Spotting Difficult Weakly Correlated Binary Knapsack Problems

Diptesh Ghosh and Tathagata Bandopadhyay
P&QM Area, IIM Ahmedabad, Vastrapur, Gujarat 380015, India.
(diptesh,tathagata)@iimahd.ernet.in

W.P. No. 2006-01-04
January 2006

INDIAN INSTITUTE OF MANAGEMENT
AHMEDABAD – 380015
INDIA

Spotting Difficult Weakly Correlated Binary Knapsack Problems

Diptesh Ghosh and Tathagata Bandopadhyay
P&QM Area, IIM Ahmedabad, Vastrapur, Gujarat 380015, India.
(diptesh,tathagata)@iimahd.ernet.in

Abstract

In this paper, we examine the possibility of quickly deciding whether or not an instance of a binary knapsack problem is difficult for branch and bound algorithms. We first observe that the distribution of the objective function values is smooth and unimodal. We define a measure of difficulty of solving knapsack problems through branch and bound algorithms, and examine the relationship between the degree of correlation between profit and cost values, the skewness of the distribution of objective function values and the difficulty in solving weakly correlated binary knapsack problems. We see that the even though it is unlikely that an exact relationship exists for individual problem instances, some aggregate relationships may be observed.
Key words: Binary Knapsack Problems; Skewness; Computational Experiments.

1 Introduction

The binary knapsack problem (BKP) is a well-known problem in combinatorial optimization (see, e.g., Martello and Toth [4]), and has been proved to be $\mathscr{NP}$-Hard in Karp [2]. In a BKP instance, we are given a set $E = \{e_1, e_2, \ldots, e_n\}$ of n (≥ 2) elements, and a number B, which is called the *budget*. Each element e_i is associated with a vector (p_i, c_i). p_i is called the *profit value* of e_i, and c_i the *cost value*. A *solution* S to the instance is a subset of E, and is called *feasible* if $\sum_{e_i \in S} c_i \leq B$. The *profit* of a solution S, also called the *objective function value* of the solution, is the quantity $\sum_{e_i \in S} p_i$. A feasible solution is called *optimal* if its profit is the largest among the profits of all feasible solutions. The objective is to identify an optimal solution to a given problem instance. Without loss of generality, $\sum_{e_i \in E} w_i > B$. For computational convenience, we choose integer values for the profit values, the cost values, and the budget.

BKP being $\mathscr{NP}$-Hard implies that unless $\mathscr{P} = \mathscr{NP}$, some (but not all) BKP instances cannot be solved to optimality within reasonable times. Exact algorithms for the BKP therefore either depend on branch and bound, or on dynamic programming. Early algorithms for the BKP were based on branch and bound; see, for example, the classic Horowitz and Sahni algorithm (Horowitz and Sahni [1]), the MT1 algorithm (Martello and Toth [6]), the MT2 algorithm (Martello and Toth [5]), and the Expknap algorithm (Pisinger [9]). Other algorithms based on dynamic programming were proposed, like the Minknap algorithm (Pisinger [8]). Combinations of branch and bound algorithms lead to the Combo algorithm (Martello et al. [3]).

BKP instances can be classified as belonging to different classes. Pisinger [7] lists several such classes, like the classes of uncorrelated instances, weakly correlated instances, strongly correlated instances, inverse strongly correlated instances, almost strongly correlated instances, subset sum problems, spanner instances, multiple strongly correlated instances, profit ceiling instances, and circle instances, and comments on their relative difficulty. The last four classes were first reported in Pisinger [7].

In this paper, we concentrate on weakly correlated BKP instances. These instances are parameterized using three parameters (say, a, b, and ρ), and are generated as follows. The cost values for all elements are first generated from a discrete uniform distribution supported on $[a, b]$. The profit values are generated randomly, with the profit value p_i of the ith element being chosen from the interval $[\lfloor(1-\rho)c_i\rfloor, \lceil(1+\rho)c_i\rceil]$, where c_i is the cost value of the element. According to Pisinger [7], these types of problems are most frequently encountered in practical applications.

For any BKP instance, the ρ value only provides an upper bound to the variability of a profit value p_i, given the corresponding cost value c_i, and hence is not a good measure of the variability between profit and cost values. We therefore define $\rho_i = (p_i/c_i) - 1$ if $p_i \geq c_i$, and $1 - (p_i/c_i)$ otherwise, and use the average of ρ_i values as a measure of the correlation between the profit and cost values in a given instance. Obviously, the average of ρ_i values is bounded above by ρ.

We define the difficulty of a BKP instance $\mathscr{I}$ for a branch and bound algorithm $\mathscr{A}$ as follows.

Definition 1 *Let a binary knapsack problem instance $\mathscr{I}$ have f feasible solutions, and a branch and bound algorithm $\mathscr{A}$ examines r of these solutions during its execution. Then the difficulty of the instance $\mathscr{I}$ for the algorithm $\mathscr{A}$ is given by*

$$D(\mathscr{A}, \mathscr{I}) = \frac{r}{f}.$$

Note that $0 < D(\cdot, \cdot) \leq 1$. Also note that the difficulty of an instance may vary from algorithm to algorithm. For an algorithm $\mathscr{A}$, an instance $\mathscr{I}_1$ is said to be more difficult (equally difficult) to solve than (respectively, as) another instance $\mathscr{I}_2$ if $D(\mathscr{A}, \mathscr{I}_1) >$ (respectively, $=$) $D(\mathscr{A}, \mathscr{I}_2)$.

In a BKP instance, it is clear that the profits for all feasible solutions would lie in $[0, z^\star]$ where $z^\star$ is the profit of an optimal solution. (The empty set is feasible, and hence there will be at least one feasible solution with 0 profit.) We plotted the frequency distribution of solution profits of several uncorrelated BKP problem instances, i.e., in instances where the profit and the cost values of each element are independently determined, with the profits of the feasible solutions on the x-axis, and the frequencies on the y-axis. For *all* instances that we generated, this empirical distribution can be approximated quite closely by a unimodal distribution. A representative plot for an instance in which there were 30 elements, the profit values and cost values were chosen independently from the interval [1,100], and budget was set to 0.6 times the sum of the costs of the elements is shown in Figure 1.

Similar experiments were performed with weakly correlated BKP instances. A representative plot of the distribution of solution profits for a weakly correlated instance with 30 elements, where the costs were generated randomly from the interval [1,100] and $\rho = 0.2$ is shown in Figure 2. Observe that the distribution of solution profits for weakly correlated instances have more negative skew than that of the distribution for uncorrelated instances.

In the next section, we examine the relationships between such skewness values, ρ values, and average of ρ_i values, and difficulty values for weakly correlated BKP instances. The paper concludes with some discussion on the usefulness of some of the observations made in Section 2.

2 Empirical Observations

In this section we examine the relationship between four "properties" of weakly correlated BKP instances, namely, the ρ value for the instance, the average of the ρ_i values for all elements in the instance, the difficulty value of the instance for the MT1 algorithm (see Definition 1), and the skewness of the distribution of profits for feasible solutions to the instance. This last measure requires

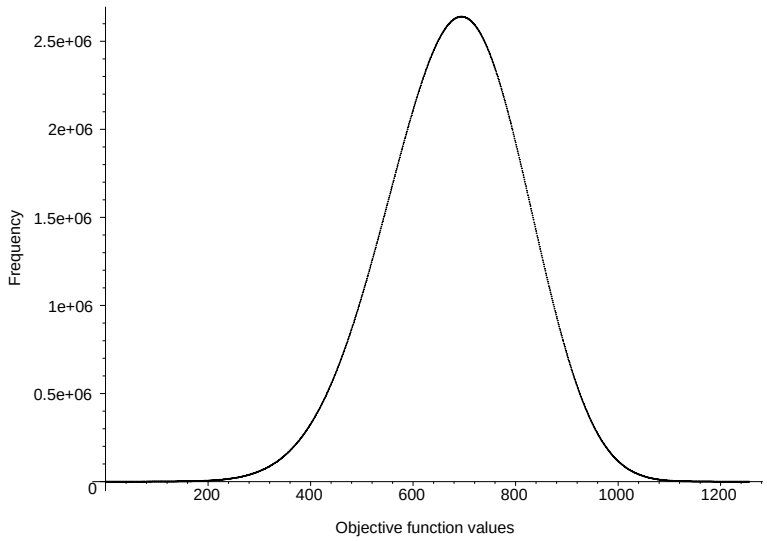

Figure 1: Distribution of solution profits for an uncorrelated BKP instance with 30 elements.

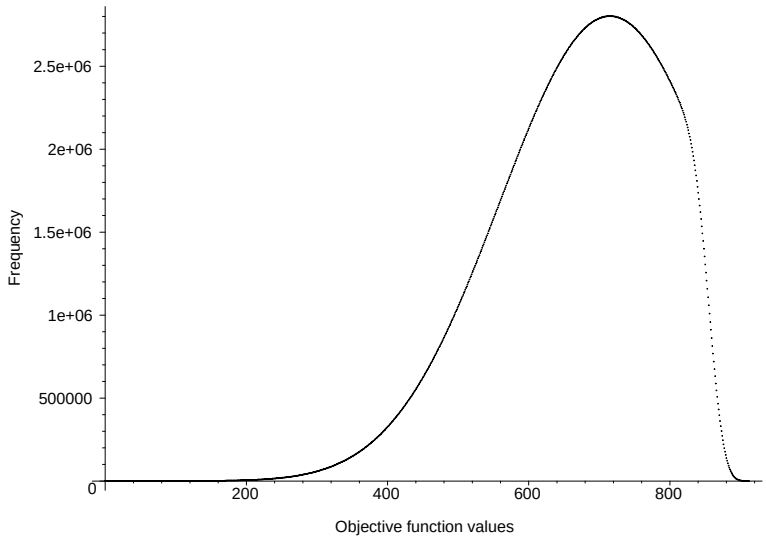

Figure 2: Distribution of solution profits for a weakly correlated BKP instance with 30 elements.

us to generate *all* feasible solutions to an instance, which restricts our study to relatively small BKP instances.

The observations in this section are based on data from 6650 randomly generated weakly correlated BKP instances. These instances are divided into 133 sets, each containing 50 instances. The sets were parameterized by two parameters, the size n of instances in the set, and the ρ value for the set. The n values we used were 16, 18, 20, 22, 24, and 26; and the ρ values we used were 0.01, 0.03, 0.05, 0.07, 0.09, 0.1, 0.2, ..., 0.8, 0.9, 0.91, 0.93, 0.95, 0.97, and 0.99. Note that the profit and cost values in instances in the sets with low ρ values were almost perfectly correlated, while those in sets with higher ρ values were allowed larger variations between the profit and cost values. The budgets for individual instances were randomly generated in the interval $[0.5 \sum_{i=1}^{n} c_i, 0.9 \sum_{i=1}^{n} c_i]$.

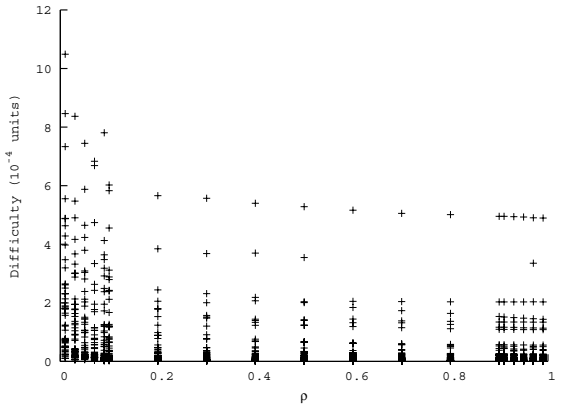

Figure 3: Distribution of difficulty values with ρ for instances with 28 elements

We first examine the distribution of difficulty values with ρ values, average of ρ_i values, and skewness values. Representative plots using 950 instances with 28 elements each are shown in Figures 3 through 5. The data for each instance is denoted by a cross mark in each figure. Although there seems to be a bounding relationship between the difficulty value and the ρ value for the instances, and to some extent, between the difficulty value and the average of ρ_i values, it is extremely unlikely that there exists a functional relationship between the measures for individual instances. In general, from the plots we see that problem instances with low ρ and average of ρ_i values (i.e., problems in which the profit and cost values are more strongly correlated) are harder for the MT1 algorithm. This fact can of course be infered from Tables 2.3 through 2.5 in Martello and Toth [4].

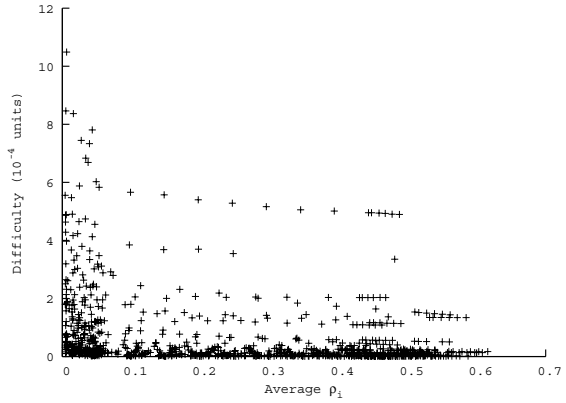

Figure 4: Distribution of difficulty values with the average of ρ_i values for instances with 28 elements

The behavior of the means and standard deviations of difficulty values with variations the other three properties is more encouraging. While the ρ values for the instances generated were controllable, the average of the ρ_i values and the skewness values were not. In our instances, the average of ρ_i values varied continuously between 0 and 0.7, while the skewness values varied between -0.1 and -0.9. In case of the average of ρ_i values (and skewness values), therefore, we divided the range of their variations into intervals of width 0.05, and partitioned the instances into subsets based on the average of ρ_i values (respectively, skewness values). We then obtained the mean and standard deviation of the difficulty values for instances in each subset of the partition and assigned these values to the midpoint of the corresponding interval. This procedure yielded the plots in Figures 6 through 8.

We see that for all problem sizes, the average of the difficulty values generally reduce at a reducing rate as ρ, average of ρ_i, or skewness decreases. The standard deviations of the difficulty values also

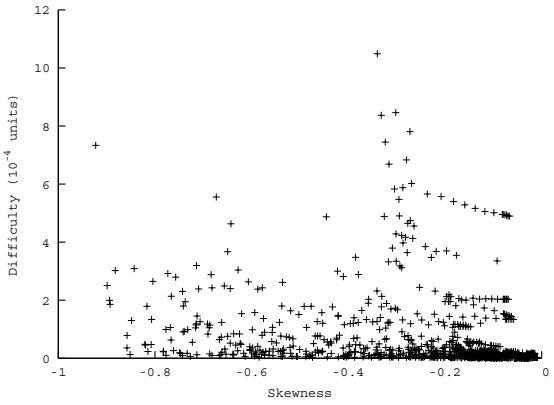

Figure 5: Distribution of difficulty values with skewness for instances with 28 elements

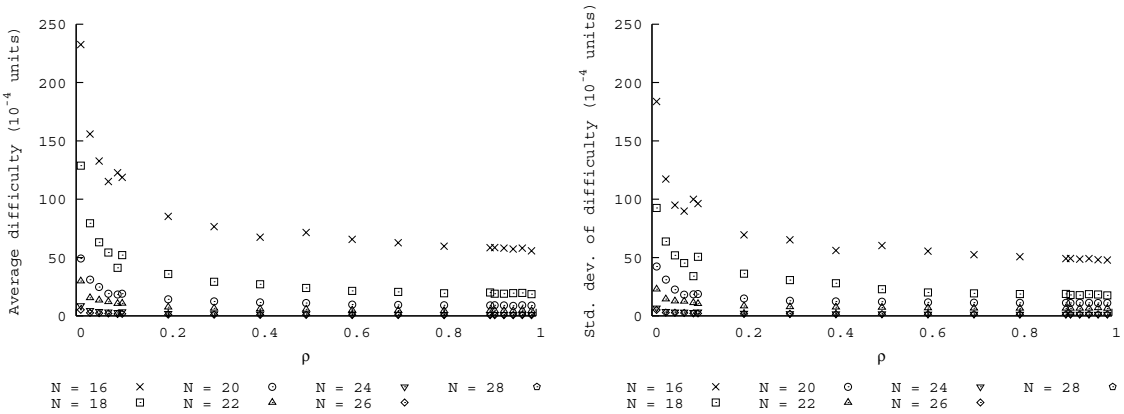

Figure 6: Variation of average of difficulty values and their std. dev. with ρ value

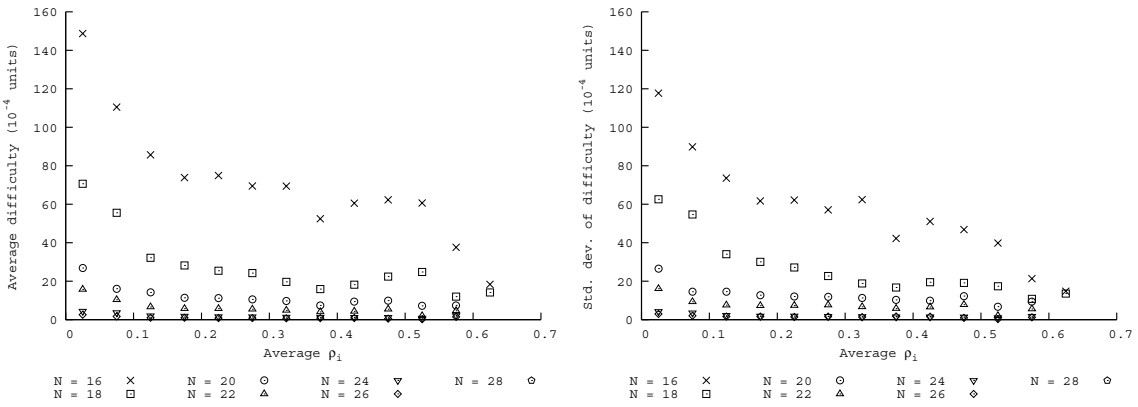

Figure 7: Variation of average of difficulty values and their std. dev. with the average of ρ_i values

show a similar trend. Notice however, that the standard deviations are of the same magnitude as the means in Figures 6 and 7, while they are smaller in Figure 8.

The relative frequency distributions for all problem sizes and for almost all average of ρ_i and skewness intervals are similar. We show representative relative frequency distributions for instances with 22 elements in Figure 9, and for instances with 28 elements in Figure 10. Both these figures show that the distributions of difficulty values for instances in each of the subsets are positively

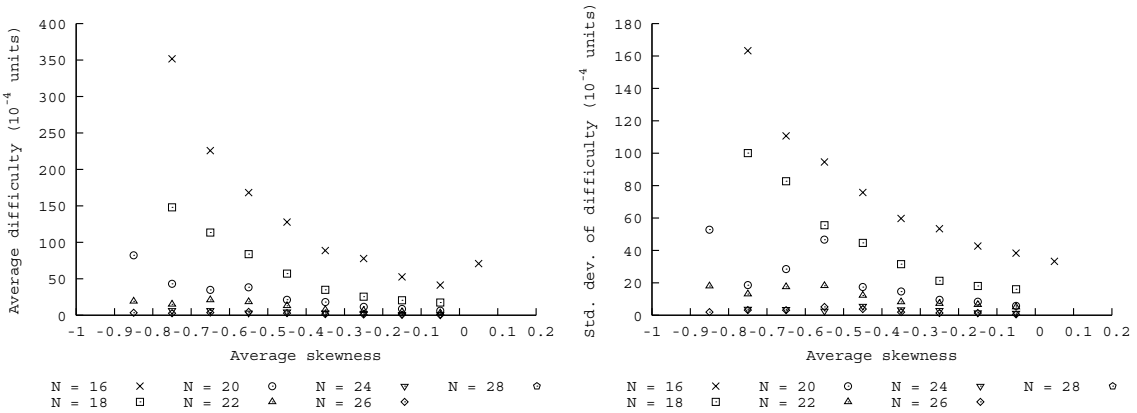

Figure 8: Variation of average of difficulty values and their std. dev. with average skewness value

skewed, which implies that the mean values shown in Figures 7 and 8 actually overestimate the most probable values. (The only cases where such right-skewed distributions are not observed are for very high average of ρ_i values, and very high skewness values. This is probably due to the fact that there are very few observations in these intervals.)

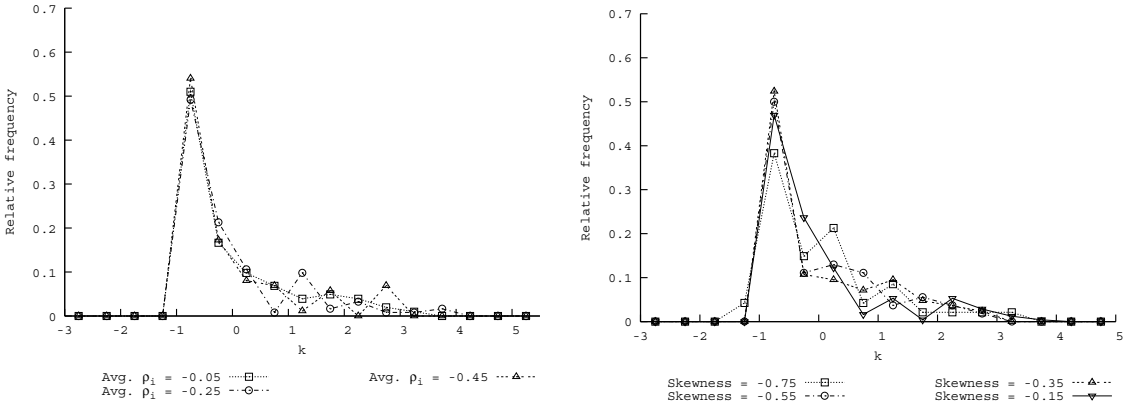

Figure 9: Relative frequency distribution of difficulty values that are k standard deviations more than the mean for instances with 22 elements

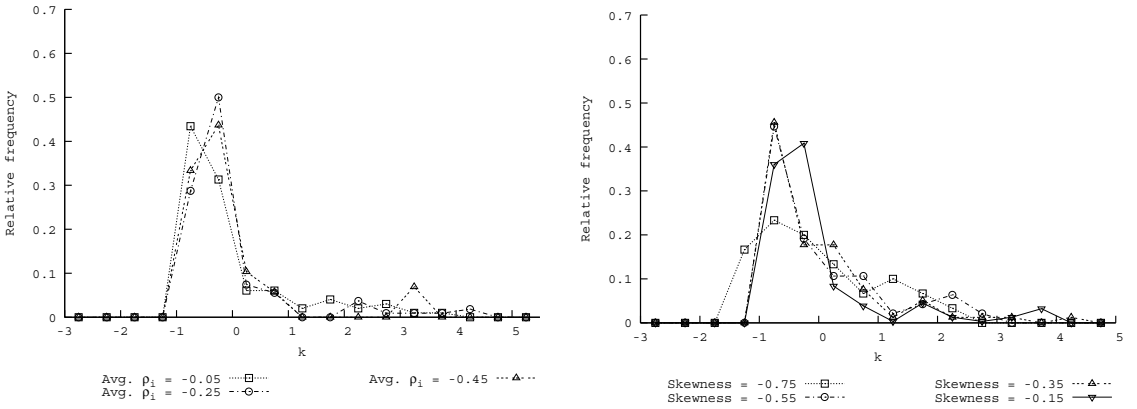

Figure 10: Relative frequency distribution of difficulty values that are k standard deviations more than the mean for instances with 28 elements

We finally examine the variation of the skewness values with ρ values and average of ρ_i values for all the problem instances experimented with. Since the averages of ρ_i values are distributed continuously, we employ the method of aggregating them as in earlier cases. The results of this experiment are depicted graphically in Figures 11 and 12. We observe that the average skewness values are directly related to both the ρ values and the average of ρ_i values. We also observe that the standard deviations of the skewness values reduce when the ρ values and the average of ρ_i values increase. Thus the variability of skewness values is more in more strongly correlated problems. In the light of these observations, the results with respect to ρ being very similar to the results with respect to the average of ρ_i values is not surprising.

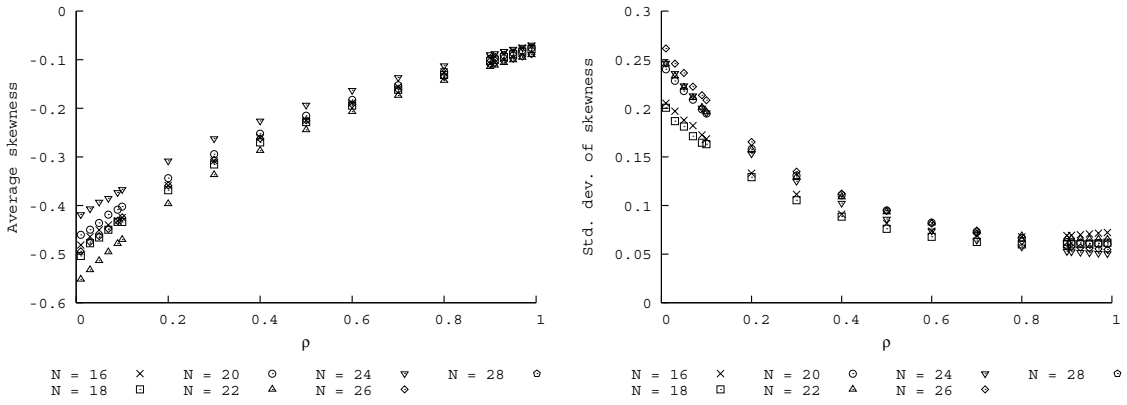

Figure 11: Variation of average skewness and std. dev. of skewness with ρ value

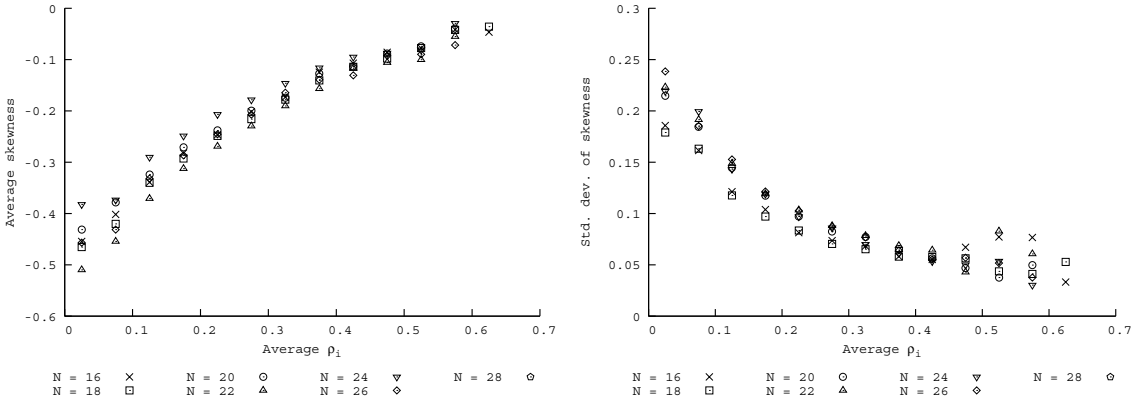

Figure 12: Variation of average skewness and std. dev. of skewness with the average of ρ_i values

3 Summary and Discussions

In this paper, we have observed relationships between the difficulty (as defined in Definition 1) in solving weakly correlated binary knapsack problem instances using the MT1 algorithm, the average of the ratios between profit and cost values of individual elements, and the skewness of the distribution of the profits of all feasible solutions to a particular instance. Our key observations are the following.

- The distribution of profits of all feasible solutions is a smooth unimodal distribution supported on $[0, z^\star]$ where $z^\star$ is the optimal solution profit. (See Figures 1 and 2.)

- The skewness of the distribution of profits of the feasible solutions decreases on average with an increase in the degree of correlation between the profit and cost values in the instance (i.e., by either the ρ value or the average of the ρ_i values).
- The difficulty in solving a particular weakly correlated binary knapsack instance is neither predictable by the degree of correlation for the instance, nor by the skewness of the distribution of profits of the feasible solutions.
- On average, the difficulty in solving knapsack problem instances reduce with increasing degree of correlation for the instance, and also with increasing skewness of the distribution of profits of the feasible solutions.
- For a fixed problem size and degree of correlation or skewness, the distribution of difficulty values is positively skewed. Therefore exceptionally hard weakly correlated binary knapsack problem instances are rare.

The third observation shows that predicting the difficulty of individual weakly correlated binary knapsack problem instances based only on problem characteristics or on the distribution of solution profits is unlikely. However, the fourth observation shows that inferences based on the average difficulty of classes of weakly correlated problems can be made.

While this paper is exploratory in nature, and does not have any predictive role, it opens up several interesting directions of research, like the one mentioned below.

Note that in all the instances that we experimented with, the distribution of solution profits was unimodal, and on a finite support. Such distributions are usually modeled as generalized Beta (GB1) distributions. These distributions have four parameters, one of which, the right end of the support is also the optimal solution profit for knapsack problem instances. An interesting line of research would be to develop sampling schemes that allow us to come up with unbiased minimum variance estimates of the optimal solution profit, and examine how sensitive the estimate is to the size of samples and sampling strategy used. Such a method would be quite different from other enumeration based techniques used to solve combinatorial optimization problems. In our limited experiments with the traveling salesperson problem, we have observed that the distribution of all tour lengths is similar, though the endpoints of the supports are both positive. This shows that the method described here can also be used for such problems.

References

[1] E. Horowitz and S. Sahni, *Computing partitions with applications to the knapsack problem* Journal of the ACM **21**, 277–292, 1974.

[2] R. M. Karp, *Reducibility among combinatorial problems*, in *Complexity of Computer Computations, Proceedings of Symposium, IBM Thomas J. Watson Research Center, Yorktown Heights, N.Y., 1972* (Eds. R. E. Miller and J. W. Thatcher), Plenum Press, New York, 85–103, 1972.

[3] S. Martello, D. Pisinger, and P. Toth, *Dynamic programming and strong bounds for the 0-1 knapsack problem*, Management Science **45**, 414–424, 1999.

[4] S. Martello and P. Toth, *Knapsack Problems: Algorithms and Computer Implementations*, Wiley, Chichester, 1990.

[5] S. Martello and P. Toth, *A new algorithm for the 0-1 knapsack problem*, Management Science **34**, 633–644, 1988.

[6] S. Martello and P. Toth, *An upper bound for the zero-one knapsack problem and a branch and bound algorithm*, European Journal of Operational Research **1**, 169–175, 1977.

[7] D. Pisinger, *Where are the hard knapsack problems?* Computers & Operations Research **32**, 2271–2284, 2005.

[8] D. Pisinger, *A minimal algorithm for the 0-1 knapsack problem*, Operations Research **45**, 758–767, 1997.

[9] D. Pisinger, *An expanding-core algorithm for the exact 0-1 knapsack problem*, European Journal of Operational Research **87**, 175–187, 1995.