Proceedings of the 1998 Winter Simulation Conference
D.J. Medeiros, E.F. Watson, J.S. Carson and M.S. Manivannan, eds.

PROPERTIES OF SYNTHETIC OPTIMIZATION PROBLEMS

Charles H. Reilly

Department of Industrial Engineering and Management Systems
University of Central Florida
P.O. Box 162450
Orlando, Florida 32816-2450, U.S.A.

ABSTRACT

In this paper, we present an approach for measuring certain properties of synthetic optimization problems based on the assumed distribution of coefficient values. We show how to estimate the proportion of all possible solutions that are feasible for the 0-1 Knapsack Problem. We calculate the population variance of the possible solution values and assess the impact of objective-constraint correlation on the variability of feasible solution values. We also show how inter-constraint correlation affects the proportion of feasible solutions in the 2-dimensional Knapsack Problem. Finally, we discuss the significance of our findings for designers of computational experiments.

1 INTRODUCTION

Many research papers introduce a new method for solving some optimization problem. Such papers typically include an evaluation of the new method's performance on synthetic instances of the problem of interest. In order to generate synthetic problem instances, a problem-generation approach is devised and parameter values are specified for certain features of the synthetic instances, for example, the tightness of the constraints or the density of the constraint coefficient matrix. Many synthetic instances are then generated, and the observed results are summarized and presented as indicative of the method's performance on all instances, including practical instances.

There is often too little understanding about how the parameters of the problem-generation methods used in computational experiments affect the characteristics of the resulting synthetic instances. When problem-generation parameters are chosen, the outcome of an experiment to be run is at least partially predetermined.

Methods for generating synthetic optimization problems ought to be subjected to the same sort of evaluations that random number generators are subjected to, such as those described in Banks, Carson, and Nelson (1996). Problem-generation methods often seem to be designed for convenience or chosen because someone else used the same method. There appears to be little regard for the consequences of selecting a particular problem-generation method.

In this paper, we present an approach for measuring certain properties of synthetic optimization problems based on the assumed distribution of coefficient values. The measurements of these properties may be made before any test problems are generated or solved. By making experimenters aware of how their parameter and distribution selections may affect the properties of the test problems they generate, we hope that better computational experiments will be designed and better understanding of the performances of solution methods will be realized.

Our paper is organized as follows. A brief literature review is presented in §2. The key insight to the approach that we use to measure properties of synthetic optimization problems is outlined in §3. We show how to estimate the proportion of possible solutions to the 0-1 Knapsack Problem (KP01) that are feasible in §4. In §5, we calculate the population variance of solution values for KP01. We consider the impact of objective-constraint correlation on the variability of feasible KP01 solution values and the impact of inter-constraint correlation on the proportion of feasible solutions for the 2-dimensional Knapsack Problem in §6. We conclude with a discussion of the significance of our findings for experimenters.

2 BACKGROUND

There have been many, many papers written about solution methods for optimization problems. Rather than cite any of those papers here, we highlight a few papers that address issues related to the properties of synthetic optimization problems or the relationship of those properties to our understanding of solution procedure performance.

Hooker (1994) advocates the development on an empirical science of algorithms. We think the present effort supports that development. By recognizing the characteristics of synthetic optimization problems, one can better interpret the results of computational experiments and better assess the true capabilities and limitations of solution methods.

Loulou and Michaelides (1979) demonstrate that the distribution of values of one coefficient type can affect the performance of heuristics, even when the expected value of the coefficients is unchanged.

Martello and Toth (1979, 1981, 1988, 1997), Balas and Martin (1980), and Balas and Zemel (1980), Reilly (1991), Rushmeier and Nemhauser (1993), Amini and Racer (1994), Cario *et al.* (1995), and Pisinger (1997) are some of the papers that include an investigation of the effect of coefficient correlation on solution procedure performance.

Reilly (1997) shows how the parameters chosen for common implicit correlation induction problem-generation methods affect the implied population correlation between the objective function and constraint coefficients for some classical optimization problems. Reilly (1998) shows how various correlation induction methods affect the distribution of objective-constraint coefficient ratios in KP01 instances.

Pilcher and Rardin (1992) describe a procedure that uses random cuts to generate symmetric traveling salesman problem instances with known optimal solution values based on a partial description of the polytope of solutions.

3 BASIC APPROACH

For the most part, we confine our attention to the 0-1 Knapsack Problem (KP01):

$$\begin{aligned} \text{Maximize } & \sum_{j=1}^{n} c_j x_j \\ \text{Subject to } & \sum_{j=1}^{n} a_j x_j \le b \\ & x_j \in \{0,1\}, \forall j, \end{aligned}$$

where all $c_j > 0$, all $a_j > 0$, $\sum_j c_j > b$, and $\max_j \{a_j\} \le b$. We assume that the c_js are i.i.d. realizations of some random variable C and that the a_js are i.i.d. realizations of some random variable A. We denote the expected values of C and A as μ_C and μ_A, respectively. The variances of C and A are denoted σ_C^2 and σ_A^2.

Suppose that a 0-1 n-vector is drawn at random from the set of all possible 0-1 solution vectors to KP01. The probability that any of the components of the vector drawn at random has value 0 is 0.5. Similarly, the probability that any of the components of the vector drawn at random has value 1 is 0.5. Therefore, the value of each vector component may be viewed as the outcome of an independent Bernoulli trial where the probability of a success and the probability of a failure are identical.

We use the notion of drawing a 0-1 n-vector at random to determine what proportion of the possible solutions to KP01 are feasible and how much variability there is among possible solution values. We consider the impact that inducing correlation between the objective and constraint coefficients has on the variability of feasible solution values. Finally, we consider how inter-constraint correlation in the 2-dimensional Knapsack Problem affects the proportion of feasible solutions.

4 PROPORTION OF FEASIBLE SOLUTIONS

When instances of KP01 are generated, the right-hand-side value in the constraint is usually set using some rule like:

$$b = t \sum_j a_j ,$$

where $0 < t < 1$. A typical value for t is 0.5.

Consider the following random variable:

$$F = \sum_j a_j x_j - t \sum_j a_j .$$

F represents the difference between the left-hand side and the right-hand side of the KP01 constraint. A non-positive value of F indicates that the KP01 solution selected at random is feasible. F is asymptotically normal with

$$\mu_F = n(0.5 - t)\mu_A$$

and

$$\sigma_F^2 = n((0.5 + t(t-1))\sigma_A^2 + \mu_A^2 / 4) .$$

The probability that a randomly selected binary n- vector is feasible, or the proportion of feasible binary n-vectors, is $\Pr(F \le 0) = \Phi(-\mu_F / \sigma_F)$, where Φ is the cumulative distribution function (c.d.f.) for the standard normal random variable.

Let $A \sim U\{1,2,\ldots,25\}$ be the random variable that represents the constraint coefficient values. (Our notation indicates that A is uniformly distributed over the integers from 1 to 25.) Table 1 shows values of $\Pr(F \le 0)$ for various values of t when $n = 100$ and $n = 500$.

Table 1: Values of $\Pr(F \le 0)$

t	$n = 100$	$n = 500$
0.25	<0.0001	<0.0001
0.30	0.0003	<0.0001
0.40	0.0409	0.0001
0.45	0.1912	0.0254
0.48	0.3633	0.2171
0.50	0.5000	0.5000
0.52	0.6367	0.7829
0.55	0.8088	0.9746
0.60	0.9591	0.9999
0.70	0.9997	>0.9999
0.75	>0.9999	>0.9999

Note that $\Pr(F \le 0)$ changes dramatically as $|0.5 - t|$ increases, especially for large values of n. This suggests that KP01 test problems generated with the same value of t and different values of n may not be comparable.

Optimization problems with very many or very few feasible solutions tend to be relatively easy to solve with an enumerative procedure (e.g., branch and bound). If nearly all solutions are feasible, a greedy procedure for selecting variables to be set to 1 will likely be quite effective. If very few solutions are feasible, then it is unlikely that there will be many deep searches along any branches in the solution tree. We conclude that values of $\Pr(F \le 0)$ that are neither small nor large would likely be associated with relatively challenging KP01 instances. However, we do not think that this measure of test problem difficulty alone will fully explain the observed performances of enumerative methods on KP01 instances. Other measures of problem characteristics or problem difficulty should be considered as well.

5 VARIABILITY OF SOLUTION VALUES

We use the same approach that we used in the last section to quantify the variability of KP01 solution values. Consider the following random variable:

$$V = \sum_j c_j x_j \ ,$$

which is asymptotically normal with

$$\mu_V = n\mu_C / 2$$

and

$$\sigma_V^2 = n(2\sigma_C^2 + \mu_C^2)/4 \ .$$

We see that the variability of the objective function coefficients and the expected value of the those coefficients determines the variability of the solution values of all possible KP01 solutions. The relative variability in solution values may determine how effective certain bounding and dominance criteria may be.

Reilly (1998) studies the relative variability in solution values of KP01 instances under various problem-generation schemes by examining the distribution of the random variable C/A.

6 EFFECT OF CORRELATION

In many KP01 papers, the effect of correlation on solution procedure performance is addressed. Examples of such papers include Martello and Toth (1979, 1988, 1997), Balas and Zemel (1980), Pisinger (1997), and Reilly (1991, 1993, 1998). Generally, KP01 instances are more difficult to solve when there is strong positive correlation between the objective and constraint coefficients. We determine how objective-constraint correlation may effect the variability of feasible solution values. In addition, we consider the proportion of feasible solutions for the 2-dimensional Knapsack Problem and how it is affected by the inter-constraint correlation.

6.1 On variability of solution values

To assess the impact that objective-constraint correlation has on the proportion of feasible solutions to KP01, we calculate the joint probability that a randomly selected 0-1 n-vector is feasible and that its solution value falls within a specified range.

Let $A \sim U\{1,2,\ldots,50\}$ and $C \sim U\{1,2,\ldots,100\}$ be the random variables representing the distributions of coefficients values. Also let $n = 100$ and $t = 0.50$. (Note that $\Pr(F \le 0) = 0.50$.) Table 2 shows values of the probabilities $\Pr(F \le 0, 0.8\mu_V \le V \le 1.1\mu_V)$ for different values of $\mathrm{Corr}(A,C)$. Table 3 shows similar values of the probability $\Pr(F \le 0, V \ge 1.1\mu_V)$.

Table 2: Values of $\Pr(F \le 0, 0.8\mu_V \le V \le 1.1\mu_V)$

Corr(A,C)	Probability
-0.9998	0.3949
-0.7499	0.3996
-0.4999	0.4044
-0.2499	0.4096
0	0.4149
0.2499	0.4206
0.4999	0.4263
0.7499	0.4320
0.9998	0.4370

Table3: Values of $\Pr(F \le 0, V \ge 1.1\mu_V)$

Corr(A,C)	Probability
-0.9998	0.0546
-0.7499	0.0479
-0.4999	0.0412
-0.2499	0.0344
0	0.0276
0.2499	0.0209
0.4999	0.0143
0.7499	0.0082
0.9998	0.0031

We observe that as the objective-constraint correlation increases, the solution values tend to become more tightly bunched around μ_V. This suggests that an enumerative algorithm would have more difficulty identifying an optimal solution. It also suggests that heuristic methods will work well on KP01 instances with strong, positive objective-constraint correlation because there are so many solutions with similar, attractive solution values.

Even though the probabilities shown in Tables 2 and 3 may not appear to change significantly, we must keep in mind that these probabilities represent the proportions of 2^n possible solutions that are feasible and whose solution values fall in a specified range. A small change in these proportions may mean a huge increase in the number of feasible solutions with competitive solution values.

6.2 Between constraints

We determine how inter-constraint correlation affects the proportion of feasible solutions to 2-dimensional Knapsack Problems by calculating the joint probability that both constraints are satisfied.

Let $A_1 \sim U\{1,2,\ldots,40\}$ and $A_2 \sim U\{1,2,\ldots,15\}$ be the random variables representing the coefficient values in the first and second constraints, respectively. Suppose that $n = 100$. Table 4 shows values of $\Pr(F_1 \le 0, F_2 \le 0)$ for different correlation levels and different combinations of t_1 and t_2 values.

We see in Table 4 that the proportion of feasible solutions increases as the inter-constraint correlation increases. This makes sense because, as the correlation increases, the constraints become increasingly similar. We see that

$$\Pr(F_1 \le 0, F_2 \le 0) \ne \Pr(F_1 \le 0)\Pr(\mathrm{F}_2 \le 0)$$

because F_1 and F_2 are correlated even if A_1 and A_2 are uncorrelated.

Table 4: Values of $\Pr(F_1 \le 0, F_2 \le 0)$

Corr(A_1, A_2)	t_1=0.48 t_2=0.53	t_1=0.50 t_2=0.50	t_1=0.52 t_2=0.47
-0.9975	0.3227	0.3397	0.2578
-0.7481	0.3299	0.3509	0.2650
-0.4988	0.3369	0.3627	0.2720
-0.2494	0.3438	0.3753	0.2789
0	0.3503	0.3890	0.2854
0.2494	0.3562	0.4044	0.2913
0.4988	0.3610	0.4222	0.2961
0.7481	0.3635	0.4451	0.2986
0.9975	0.3637	0.4939	0.2988

7 DISCUSSION

We recommend an approach that allows experimenters to determine what properties the synthetic optimization problems they intend to generate will have before any test problems are generated or solved. With this capability, experimenters can compare the characteristics of their test problems to those of practical problem instances. In addition, knowledge of these properties may facilitate better designed computational experiments and consequently better understanding of the capabilities and limitations of solution methods.

We think that our approach can facilitate better comparisons of solution procedure performance for different problem sizes (i.e., number of decision variables) and across problem classes. An experimenter can, for example, generate 100-variable and 500-variable KP01 instances with the same proportion of feasible solutions by appropriately varying t. An experimenter could also generate KP01 instances and 2-dimensional Knapsack Problem instances with the same proportion of feasible solutions to determine how the number of constraints affects the performance of a solution procedure.

KP01 instances tend to be quite challenging for a branch-and-bound routine when the objective-constraint correlation is strongly positive. KP01 solution values tend to be more similar under the same condition. This suggests that different types of test problems should be used to evaluate algorithms and heuristics. The test problems that present the greatest challenge for a branch-and-bound routine may be the same problems on which a greedy heuristic performs best.

Changing the distributions of coefficient values will change the measurements of properties that we consider here. An experimenter should not assume that the performance of a solution method is insensitive to the distribution of coefficient values.

One of the possible benefits to be realized from measuring test problem characteristics is new insights for designing solution procedures that can exploit the features of instances that are usually the most troubling for existing solution methods.

REFERENCES

Amini, M. and M. Racer. 1994. Comparison of alternative solution methods for the generalized assignment problem. *Management Science* 40: 868-890.

Balas, E. and C.H. Martin. 1980. Pivot and complement- a heuristic for 0-1 programming. *Management Science* 26: 86-96.

Balas, E. and E. Zemel. 1980. An algorithm for large zero-one knapsack problems. *Operations Research* 28: 1130-1154.

Banks, J., J.S. Carson, II, and B.L. Nelson. 1996. *Discrete-event system simulation*. 2d ed. Upper Saddle River, NJ: Prentice Hall.

Cario, M., J.J Clifford, R.R. Hill, J. Yang, K. Yang, and C.H. Reilly. 1995. Alternative methods for generating synthetic generalized assignment problems. In *Proceedings of the 4th Industrial Engineering Research Conference*, ed. R. Uszoy and B.W. Schmeiser, 1080-1089. Institute of Industrial Engineers, Atlanta, Georgia.

Hooker, J.N. 1994. Needed: an empirical science of algorithms. *Operations Research* 42 (2): 201-212.

Loulou, R. and E. Michaelides. 1979. New greedy heuristics for the multidimensional 0-1 knapsack problem. *Operations Research* 27:1101-1114.

Martello, S. and P. Toth. 1979. The 0-1 knapsack problem. In *Combinatorial optimization*, ed. A. Mingozzi and C. Sandi, 237-279. New York: John Wiley and Sons.

Martello, S. and P. Toth. 1981. An algorithm for the generalized assignment problem. In *Proceedings of the 9th IFORS Conference*, ed. J. Brans, 589-603. North-Holland.

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

Martello, S. and P. Toth. 1997. Upper bounds and algorithms for 0-1 knapsack problems. *Operations Research* 45 (5): 768-778.

Pilcher, M.G. and R.L. Rardin. 1992. Partial polyhedral description and generation of discrete optimization problems with known optima. *Naval Research Logistics* 39 (6): 839-858.

Pisinger, D. 1997. A minimal algorithm for the 0-1 knapsack problem. *Operations Research* 45 (5): 758-767.

Reilly, C.H. 1991. Optimization problems with uniformly distributed coefficients. In *Proceedings of the 1991 Winter Simulation Conference*, ed. B.L. Nelson, W.D. Kelton, and G.M. Clark, 866-874. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Reilly, C.H. 1993. A comparison of alternative input models for synthetic optimization problems. In *Proceedings of the 1993 Winter Simulation Conference*, ed. G.W. Evans, M. Mollaghasemi, E.C. Russell, and W.E. Biles, 356-364. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Reilly, C.H. 1997. Generating coefficients for optimization problems with implicit correlation induction. In *Proceedings of the 1997 IEEE International Conference on Systems, Man, and Cybernetics*, 2438-2443. Institute of Electrical and Electronics Engineers, Piscataway, New Jersey.

Reilly, C.H. 1998. On the random generation of synthetic 0-1 knapsack problems. Manuscript dated March 1, 1998, Department of Industrial Engineering and Management Systems, University of Central Florida, Orlando, Florida.

Rushmeier, R. and G.L. Nemhauser. 1993. Experiments with parallel branch-and-bound algorithms for the set covering problem. *Operations Research Letters* 13 (5): 277-285.

AUTHOR BIOGRAPHY

CHARLES H. REILLY is Chair and Professor in the Department of Industrial Engineering and Management Systems at the University of Central Florida. He earned a B.A. in mathematics and business administration at St. Norbert College in 1979, and M.S. and Ph.D. degrees in industrial engineering at Purdue University in 1980 and 1983, respectively. His current research interests include the generation of synthetic optimization problems, the evaluation of solution methods for discrete optimization problems, and simulation modeling. Dr. Reilly is a member of IIE, INFORMS, and ASEE.