Proceedings of the 1991 Winter Simulation Conference
Barry L. Nelson, W. David Kelton, Gordon M. Clark (eds.)

OPTIMIZATION TEST PROBLEMS WITH UNIFORMLY DISTRIBUTED COEFFICIENTS

Charles H. Reilly

Department of Industrial and Systems Engineering
The Ohio State University
1971 Neil Avenue
Columbus, Ohio 43210-1271

ABSTRACT

When an empirical evaluation of a solution method for an optimization problem is conducted, a standard approach is to generate test problems in which all of the coefficients are assumed to be independently and uniformly distributed. However, the performance of algorithms and heuristics can degrade as the correlation among the coefficients in integer programming test problems is strengthened. We show how to characterize the joint distribution of two discrete uniform random variables with any feasible correlation and any feasible value for the smallest joint probability when the number of possible values of one random variable is a multiple of the number of possible values of the other. We use this characterization in an experiment with randomly-generated 0-1 knapsack problems, and we summarize the results of the experiment.

1 INTRODUCTION

Empirical evaluations of solution methods for optimization problems are commonly conducted. In many cases, test problems are randomly generated under the assumption that the coefficients in the problems are independently and uniformly distributed. This approach to evaluating a solution procedure is inadequate for at least two reasons. First, the coefficients in actual instances of optimization problems may not be probabilistically independent. For example, consider a product-mix problem. We would expect the selling prices of products that require more resources for production to be greater than the selling prices of products requiring fewer resources. Second, there is usually no basis for the assumption that the coefficients in optimization problems are uniformly distributed.

Both of these concerns can be addressed if different approaches to generating test problems are taken. For example, randomly-generated test problems could include problems in which correlation among the coefficients is controlled and varied, at least when an argument for dependence among the coefficients can be made. Also, problems whose coefficients are drawn from a variety of different distributions, including the uniform distribution, could be included when there is uncertainty about the distributions of parameters in real instances of the problem of interest. In this paper, we deal only with the first concern and the first remedy.

Moore (1989) studies the effect of correlation between the objective function and constraint coefficients in 0-1 knapsack problems. She finds that the number of iterations required to achieve optimality with implicit enumeration increases exponentially as the expected correlation between the two sets of coefficients increases. Martello and Toth (1979) find that implicit enumeration performs better than dynamic programming on 0-1 knapsack problems with weakly correlated coefficients, but dynamic programming performs better when the coefficients are strongly correlated. Balas and Zemel (1980) present an algorithm for 0-1 knapsack problems in which enumerative steps play a small role. They experiment with the same types of problems that Martello and Toth (1979) use and find that the performance of their method is better on problems with perfectly correlated coefficients than on problems with uncorrelated or weakly correlated coefficients.

Moore (1990) investigates the effects of correlation between objective function coefficients and constraint-matrix column sums, distribution of the column sums, and constraint-matrix density on the performance of implicit enumeration and two heuristics for the weighted set covering problem. In this study, she finds that high, positive (expected) correlation between the two sets of problem parameters leads to poorer solution procedure performance. Also, solution procedure performance on sparse problems with column sums that are uniformly distributed is slightly worse than on problems whose column sums are binomially distributed.

Balas and Martin (1980) generate capital budgeting (multidimensional knapsack) problems in which each decision variable's constraint coefficients are related to its objective function coefficient. They find that these problems are more difficult than problems with independent coefficients.

Peterson (1990) shows how the probability mass function (pmf) for a discrete multivariate random variable with the maximum value for the smallest joint probability and a specified feasible correlation can be found by solving a linear program. In the bivariate case, this linear program is a bottleneck transportation problem with a side constraint (BTPSC). He presents a solution procedure for BTPSC that is based on (at most) four applications of the Northwest Corner Rule (NWCR) and sensitivity analysis, and he constructs a parametric curve that shows the maximum value of the smallest joint probability as a function of the expectation of the product of the random variables or, equivalently, as a function of their correlation. Peterson also presents special-case results for bivariate random variables with symmetric marginals and uniform marginals.

Whitt (1976) characterizes bivariate distributions with extreme correlation. He also suggests a procedure for constructing distributions with extreme correlation based on a rearrangement theorem. A similar notion is used by Evans (1984) to show that NWCR finds the optimal solution to the factored transportation problem. Peterson (1990) bases his solution procedure for BTPSC on the theorem by Evans.

There are many articles dealing with characterizations of discrete bivariate random variables. Examples include Kemp and Papageorgiou (1982), Kocherlakota (1989), Marshall and Olkin (1985), and Nagaraja (1983). Some of these papers deal with characterizations that preserve properties, e.g., memorylessness in the case of the bivariate geometric distribution, of the associated univariate marginal distributions.

Moore, Peterson, and Reilly (1990) summarize the work of Moore (1989, 1990) and Peterson (1990).

In this paper, we study the pmfs for random variables (Y_1, Y_2) where the marginal distributions are uniform and the number of possible values of one random variable is a multiple of the number of possible values of the other. We present closed-form expressions for the pmfs that maximize and minimize $E(Y_1Y_2)$ for a specified smallest joint probability and a procedure for finding other pmfs with the same smallest joint probability and a specified value for $E(Y_1Y_2)$. Given specified values for the smallest joint probability and $E(Y_1Y_2)$, we show that, when Y_1 and Y_2 both have at least three possible values, a pmf with at most four values can be found by mixing the extreme-correlation pmfs and the pmf for Y_1 and Y_2 independent. Finally, we summarize an experiment on 0-1 knapsack problems with independent coefficients and with correlated coefficients generated using our pmf characterization.

2 ASSUMPTIONS

To characterize pmfs of a bivariate random variable (Y_1, Y_2) with uniform marginal distributions, we make the following assumptions:

1. Y_1 is distributed over the support $I_1 = \{j_1 + 1, j_1+2, ..., j_1+n_1\}$, where j_1 is integer, according to pmf

$$f_1(y_1) = \begin{cases} \frac{1}{n_1}, & \text{for } y_1 \in I_1; \\ 0, & \text{otherwise.} \end{cases}$$

2. Y_2 is distributed over the support $I_2 = \{j_2 + 1, j_2+2, ..., j_2+n_2\}$, where j_2 is integer, according to pmf

$$f_2(y_2) = \begin{cases} \frac{1}{n_2}, & \text{for } y_2 \in I_2; \\ 0, & \text{otherwise.} \end{cases}$$

3. $n_1 | n_2$; that is, $m = n_2/n_1$ is integer.

These assumptions are typical of those made when an experiment to evaluate a solution procedure for an optimization problem is designed. The first (second) assumption implies that Y_1 (Y_2) is a discrete uniform random variable; that is, $Y_i \sim U[j_i + 1, j_i + n_i]$, $i = 1, 2$. The third assumption allows us to find closed-form solutions for the transportation problems that Peterson (1990) recommends solving to find pmfs for (Y_1, Y_2) with a specified value for the smallest joint probability and maximum and minimum correlation.

We note that

$$E(Y_i) = j_i + \frac{n_i + 1}{2}, \quad i = 1, 2;$$

and

$$\mathrm{Var}(Y_i) = \frac{n_i^2 - 1}{12}, \quad i = 1, 2.$$

The support of (Y_1, Y_2) is $I = I_1 \times I_2$.

3 ALL POSSIBLE JOINT PMFs

Let θ be the smallest joint probability for any $(y_1, y_2) \in I$. The largest possible value for θ is $\theta^* = 1/n_1 n_2$. In other words, the smallest joint probability over $(y_1, y_2) \in I$ is maximized when Y_1 and Y_2 are independent. We denote the pmf of (Y_1, Y_2) as $g^*(y_1, y_2)$ when Y_1 and Y_2 are independent.

Peterson (1990) showed that the parametric curve that plots the largest possible value for θ against $\mathrm{E}(Y_1 Y_2)$ or the correlation, ρ, is an isosceles triangle symmetric about

$$K^* = \mathrm{E}(Y_1)\mathrm{E}(Y_2)$$

or $\rho = 0$.

Let

$$\Delta = j_1 j_2 + j_1\left(\frac{n_2+1}{2}\right) + j_2\left(\frac{n_1+1}{2}\right).$$

The largest possible value for $\mathrm{E}(Y_1 Y_2)$ is

$$\begin{aligned} K_{max} &= \max\{\mathrm{E}(Y_1 Y_2 \mid \theta = 0)\} \\ &= \frac{n_1+1}{12}[4n_2 - m + 3] + \Delta, \end{aligned}$$

and the smallest possible value for $\mathrm{E}(Y_1 Y_2)$ is

$$\begin{aligned} K_{min} &= \min\{\mathrm{E}(Y_1 Y_2 \mid \theta = 0)\} \\ &= \frac{n_1+1}{12}[2n_2 + m + 3] + \Delta. \end{aligned}$$

It follows that the maximum and minimum feasible correlations for (Y_1, Y_2) are

$$\rho_{max} = (n_1 n_2 - m)\left[\frac{1}{(n_1^2-1)(n_2^2-1)}\right]^{\frac{1}{2}}$$

and

$$\rho_{min} = -\rho_{max},$$

respectively.

Suppose, for example, $j_1 = 0$, $n_1 = 2$, $j_2 = 0$, and $n_2 = 4$. Then $\mathrm{E}(Y_1) = 1.5$, $\mathrm{E}(Y_2) = 2.5$, $\mathrm{Var}(Y_1) = 0.25$, $\mathrm{Var}(Y_2) = 1.25$, $\theta^* = 0.125$, $K^* = 3.75$, $\Delta = 0$, $K_{max} = 4.25$, $K_{min} = 3.25$, $\rho_{max} = 0.894$, and $\rho_{min} = -0.894$.

The parametric curve suggested by Peterson (1990) is competely defined by the points $(K_{min}, 0)$, (K^*, θ^*), and $(K_{max}, 0)$. A typical parametric curve is shown in Figure 1.

Let K be the desired value of $\mathrm{E}(Y_1 Y_2)$. The parametric curve can be expressed as follows:

$$\theta = \begin{cases} \frac{K - K_{min}}{K^* - K_{min}}\theta^*, & \text{if } K_{min} \le K \le K^*; \\ \frac{K - K_{max}}{K^* - K_{max}}\theta^*, & \text{if } K^* \le K \le K_{max}. \end{cases}$$

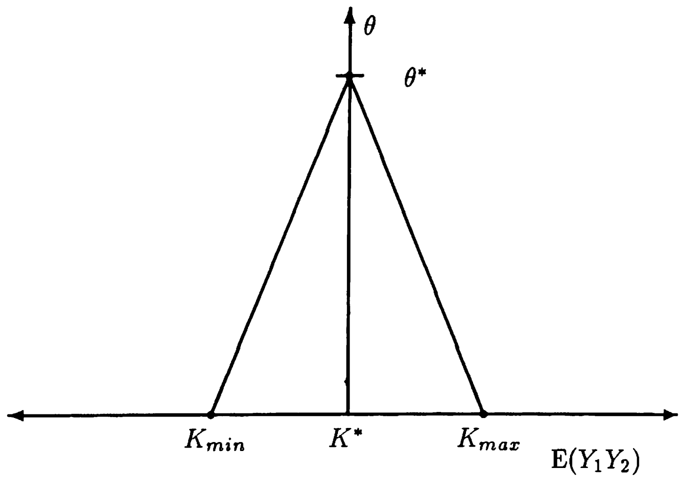


Figure 1: Typical Parametric Curve

It follows that

$$\begin{aligned} K^+(\theta) &= \max\{\mathrm{E}(Y_1 Y_2 \mid \theta)\} \\ &= K_{max} + (K^* - K_{max})\frac{\theta}{\theta^*} \end{aligned}$$

and

$$\begin{aligned} K^-(\theta) &= \min\{\mathrm{E}(Y_1 Y_2 \mid \theta)\} \\ &= K_{min} + (K^* - K_{min})\frac{\theta}{\theta^*}. \end{aligned}$$

Let $y_i' = y_i - j_i$; that is, $y_i' \in \{1, 2, ..., n_i\}$. A common way to characterize the pmf for (Y_1, Y_2) with correlation $\rho_0 \ge 0$ is to "mix" the pmfs $g^*(y_1, y_2)$ and $g_{max}(y_1, y_2)$ as follows:

$$\left(1 - \frac{\rho_0}{\rho_{max}}\right) g^*(y_1, y_2) + \left(\frac{\rho_0}{\rho_{max}}\right) g_{max}(y_1, y_2), \quad (1)$$

where

$$g_{max}(y_1, y_2) = \begin{cases} \frac{1}{n_2}, & \text{if } m(y_1' - 1) < y_2' \le m y_1'; \\ 0, & \text{otherwise;} \end{cases}$$

is the pmf with Y_1 and Y_2 maximally correlated. In this case, $\theta = (1 - \rho_0/\rho_{max})\theta^*$. A similar mixture of $g^*(y_1, y_2)$ and $g_{min}(y_1, y_2)$,

$$\left(1 - \frac{\rho_0}{\rho_{min}}\right) g^*(y_1, y_2) + \left(\frac{\rho_0}{\rho_{min}}\right) g_{min}(y_1, y_2), \quad (2)$$

where

$$g_{min}(y_1, y_2) = \begin{cases} \frac{1}{n_2}, & \text{if } m(n_1 - y_1') < y_2' \le m(n_1 - y_1' + 1); \\ 0, & \text{otherwise;} \end{cases}$$

is the minimum-correlation pmf for (Y_1, Y_2), can be used if $\rho_0 < 0$. The points on the parametric curve correspond to unique pmfs that are mixtures of the form (1) or (2). Hence, there is no need to solve BTPSC to find the pmf of (Y_1, Y_2) with the largest possible value for θ and a specified correlation. The points on or above the horizontal axis and below the parametric curve correspond to classes of other possible pmfs for (Y_1, Y_2). We may view the parametric curve as defining the envelope of all possible joint distributions of (Y_1, Y_2).

4 EXTREME-CORRELATION PMFs

Suppose $\theta = \theta_0$. The joint distribution for (Y_1, Y_2) with the maximum or minimum possible value for $E(Y_1Y_2|\theta_0)$ can be found by solving a transportation problem with NWCR (Peterson, 1990). When Y_1 and Y_2 are uniform, a closed-form expression for the pmf with $\max\{E(Y_1Y_2|\theta_0)\}$ is

$$g^+(y_1, y_2|\theta_0) = \begin{cases} \frac{1}{n_2} - (n_1 - 1)\theta_0, & \text{if } m(y_1' - 1) < y_2' \le my_1'; \\ \theta_0, & \text{if } 1 \le y_2' \le m(y_1' - 1) \\ & \quad \text{or } my_1' < y_2' \le n_2; \\ 0, & \text{otherwise.} \end{cases}$$

A similar closed-form expression for the pmf with $\min\{E(Y_1Y_2|\theta_0)\}$ is:

$$g^-(y_1, y_2|\theta_0) = \begin{cases} \frac{1}{n_2} - (n_1 - 1)\theta_0, & \text{if } m(n_1 - y_1') < y_2' \le \\ & \quad m(n_1 - y_1' + 1); \\ \theta_0, & \text{if } 1 \le y_2' \le m(n_1 - y_1') \text{ or} \\ & \quad m(n_1 - y_1' + 1) < y_2' \le n_2; \\ 0, & \text{otherwise.} \end{cases}$$

Note that

$$g^+(y_1, y_2|\theta_0) = \left(\frac{\theta_0}{\theta^*}\right) g^*(y_1, y_2) + \left(1 - \frac{\theta_0}{\theta^*}\right) g_{max}(y_1, y_2) \quad (3)$$

and

$$g^-(y_1, y_2|\theta_0) = \left(\frac{\theta_0}{\theta^*}\right) g^*(y_1, y_2) + \left(1 - \frac{\theta_0}{\theta^*}\right) g_{min}(y_1, y_2). \quad (4)$$

Y_1 \ Y_2	$j_2 + 1$	$j_2 + 2$	$j_2 + 3$	$j_2 + 4$	
$j_1 + 1$	$\frac{1}{4} - \theta_0$	$\frac{1}{4} - \theta_0$	θ_0	θ_0	$\frac{1}{2}$
$j_1 + 2$	θ_0	θ_0	$\frac{1}{4} - \theta_0$	$\frac{1}{4} - \theta_0$	$\frac{1}{2}$
	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	

Figure 2: Joint pmf for $\max\{E(Y_1Y_2|\theta_0)\}$

Y_1 \ Y_2	$j_2 + 1$	$j_2 + 2$	$j_2 + 3$	$j_2 + 4$	
$j_1 + 1$	θ_0	θ_0	$\frac{1}{4} - \theta_0$	$\frac{1}{4} - \theta_0$	$\frac{1}{2}$
$j_1 + 2$	$\frac{1}{4} - \theta_0$	$\frac{1}{4} - \theta_0$	θ_0	θ_0	$\frac{1}{2}$
	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	

Figure 3: Joint pmf for $\min\{E(Y_1Y_2|\theta_0)\}$

Mixtures of the form (3) [(4)] are actually mixtures of the form (1) [(2)], where $\theta_0/\theta^* = 1 - \rho_0/\rho_{max}$ $[1 - \rho_0/\rho_{min}]$.

Suppose $n_1 = 2$ and $n_2 = 4$. The pmf for Y_1 and Y_2 that maximizes $E(Y_1Y_2|\theta_0)$ is shown in Figure 2. See Figure 3 for the pmf that minimizes $E(Y_1Y_2|\theta_0)$. Note the staircase structure in both figures.

5 NONEXTREME-CORRELATION PMFs

Consider the pmf $g^+(y_1, y_2|\theta_0)$. Other pmfs, $g(y_1, y_2|\theta_0)$, with different values for $E(Y_1Y_2)$ can be found by redistributing probability from $(y_1, y_2) \in I$ with $g^+(y_1, y_2|\theta_0) > \theta_0$ to $(y_1, y_2) \in I$ with $g^+(y_1, y_2|\theta_0) = \theta_0$. For example, up to $\frac{1}{n_2} - n_1\theta_0$ could be redistributed from $(Y_1, Y_2) = (j_1 + 1, j_2 + 1)$ to $(j_1 + n_1, j_2 + 1)$ and from $(j_1 + n_1, j_2 + n_2)$ to $(j_1 + 1, j_2 + n_2)$ without affecting the smallest joint probability, θ_0, or violating the marginal probabilities. If β, $0 < \beta \le \frac{1}{n_2} - n_1\theta_0$, is the probability to be redistributed, the value of $E(Y_1Y_2)$ would change by $(1 - n_2)(n_1 - 1)\beta$.

Peterson (1990) uses a probability reallocation scheme in his algorithm to find a pmf when the desired value for $E(Y_1Y_2|\theta^*)$ is between $\min\{E(Y_1Y_2|\theta^*)\}$ and $\max\{E(Y_1Y_2|\theta^*)\}$. Given a target value for $E(Y_1Y_2)$, K_0, and a feasible choice for θ_0, the following procedure which is based on Peterson's method is an example of a method for finding an appropriate pmf $g(y_1, y_2|\theta_0)$:

1. Let $i = 1$, $K' = \max\{E(Y_1Y_2)|\theta_0)\}$, and $h(y_1, y_2) = g^+(y_1, y_2|\theta_0)$, $\forall y_1, y_2$.
2. $\ell \leftarrow \lfloor \frac{i}{n_1} \rfloor$. $\phi = (n_2 - 2i + 1)(n_1 - 2\ell - 1)$.
3. $\beta \leftarrow \min\left\{\frac{K'-K_0}{\phi}, \frac{1}{n_2} - n_1\theta_0\right\}$.
   $h(j_1 + \ell + 1, j_2 + i) \leftarrow h(j_1 + \ell + 1, j_2 + i) - \beta$,
   $h(j_1 + \ell + 1, j_2 + n_2 - i + 1) \leftarrow$
   $h(j_1 + \ell + 1, j_2 + n_2 - i + 1) + \beta$,
   $h(j_1 + n_1 - \ell, j_2 + i) \leftarrow$
   $h(j_1 + n_1 - \ell, j_2 + i) + \beta$, and
   $h(j_1 + n_1 - \ell, j_2 + n_2 - i + 1) \leftarrow$
   $h(j_1 + n_1 - \ell, j_2 + n_2 - i + 1) - \beta$.
4. $K' \leftarrow K' - \beta\phi$.
   If $\beta = (K' - K_0)/\phi$, $g(y_1, y_2|\theta_0) = h(y_1, y_2)$ and stop. Otherwise, $i \leftarrow i + 1$ and go to step 2.

If $n_1 > 2$, there will be some values $(y_1, y_2) \in I$ that will have joint probabilities of θ_0 in the pmfs $g^+(y_1, y_2|\theta_0)$ and $g^-(y_1, y_2|\theta_0)$. Consequently, an appropriate pmf can be found by simply mixing $g^+(y_1, y_2|\theta_0)$ and $g^-(y_1, y_2|\theta_0)$. The resulting pmf would be different than the one that would be constructed with the probabilty redistribution procedure above.

Assume $n_1 > 2$. Let $K^-(\theta_0) \le K_0 \le K^+(\theta_0)$. A pmf with $E(Y_1Y_2|\theta_0) = K_0$ can be found as follows:

$$\left(\frac{K^+(\theta_0) - K_0}{K^+(\theta_0) - K^-(\theta_0)}\right) g^-(y_1, y_2|\theta_0) + \left(\frac{K_0 - K^-(\theta_0)}{K^+(\theta_0) - K^-(\theta_0)}\right) g^+(y_1, y_2|\theta_0). \quad (5)$$

After substituting for $g^+(y_1, y_2|\theta_0)$ and $g^-(y_1, y_2|\theta_0)$ using (3) and (4), we can express this mixture as

$$\left(\frac{\theta_0}{\theta^*}\right) g^*(y_1, y_2) + \left(\frac{K^+(\theta_0) - K_0}{K_{max} - K_{min}}\right) g_{min}(y_1, y_2) + \left(\frac{K_0 - K^-(\theta_0)}{K_{max} - K_{min}}\right) g_{max}(y_1, y_2). \quad (6)$$

This indicates that a pmf for any feasible combination (K_0, θ_0) can be expressed as a mixture of the unique pmfs corresponding to the points (K^*, θ^*), $(K_{min}, 0)$, and $(K_{max}, 0)$ on the parametric curve when $n_1 > 2$.

Y_1 \ Y_2	1	2	3	4	5	6	
1	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{3}$
2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{3}$
3	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{3}$
	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	

Figure 4: Joint pmf $g^+(y_1, y_2|\frac{1}{36})$ ($n_1 = 3$, $n_2 = 6$)

The mixture (6) has three interesting properties. First, the number of joint probabilities that equal θ_0 is $n_2(n_1-2)+m$, if $n_1 \ge 3$ and odd, and $n_2(n_1-2)$, if $n_1 \ge 4$ and even. For example, if $n_1 = 50$, $n_2 = 100$, and $\theta_0 < \theta^*$, the number of joint probabilities equal to θ_0 is 4800 or 96% of all of the joint probabilities. Second, the joint probabilities that are equal to θ_0 form an "X" along the main "diagonals" in tabular representations of the pmfs like those in Figures 2 and 3. This occurs because of the staircase structures in $g^+(y_1, y_2|\theta_0)$ and $g^-(y_1, y_2|\theta_0)$ noted earlier in Figures 2 and 3 and the fact that both of these pmfs are included in this mixture. Finally, the maximum number of different probability values associated with the $(y_1, y_2) \in I$ in this mixture is four. These properties may be exploited in a random generation procedure.

Suppose $j_1 = 0$, $n_1 = 3$, $j_2 = 0$, and $n_2 = 6$. Let $K_0 = \frac{22}{3}$ and $\theta_0 = \frac{1}{36}$. Then $K^+(\frac{1}{36}) = \frac{23}{3}$. The pmf $g^+(y_1, y_2|\frac{1}{36})$ is shown in Figure 4. Our probability redistribution procedure would yield the pmf shown in Figure 5. Figure 6 shows the pmf that results from the mixture (6).

When $n_1 = 2$, the smallest joint probability for any $(y_1, y_2) \in I$ in a mixture of the form (6) would be greater than θ_0 for all but one case and $E(Y_1Y_2)$ would always be K_0. Consider the pmfs $g^+(y_1, y_2|\theta_0)$ and $g^-(y_1, y_2|\theta_0)$. The joint probability associated with $(y_1, y_2) \in I$ is at least θ_0, by definition, in both of these pmfs. Note that $\frac{1}{n_2} - (n_1 - 1)\theta_0 > \theta_0$ unless $\theta_0 = \theta^*$. If we assume $\theta_0 \ne \theta^*$, then every joint probability associated with the pmf $\lambda g^+(y_1, y_2|\theta_0) + (1-\lambda)g^-(y_1, y_2|\theta_0)$, where $0 < \lambda < 1$, is greater than θ_0 because it is a strict convex combination of θ_0 and $\frac{1}{n_2} - (n_1 - 1)\theta_0$.

		Y_2						
		1	2	3	4	5	6	
Y_1	1	$\frac{7}{90}$	$\frac{1}{9}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{11}{180}$	$\frac{1}{3}$
	2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{3}$
	3	$\frac{11}{180}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{9}$	$\frac{7}{90}$	$\frac{1}{3}$
		$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	

Figure 5: Joint pmf from redistribution

		Y_2						
		1	2	3	4	5	6	
Y_1	1	$\frac{13}{144}$	$\frac{13}{144}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{7}{144}$	$\frac{7}{144}$	$\frac{1}{3}$
	2	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{3}$
	3	$\frac{7}{144}$	$\frac{7}{144}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{13}{144}$	$\frac{13}{144}$	$\frac{1}{3}$
		$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	

Figure 6: Joint pmf from mixing

6 THE 0-1 KNAPSACK PROBLEM

The 0-1 knapsack problem has the following form:
Maximize

$$\sum_{j=1}^{n} c_j x_j$$

Subject to

$$\sum_{j=1}^{n} a_j x_j \leq b$$

$$x_j = 0 \text{ or } 1, \quad j = 1, 2, \ldots, n,$$

where

$$x_j = \begin{cases} 1 & \text{if item } j \text{ is included in the knapsack;} \\ 0 & \text{otherwise;} \end{cases}$$

$c_j > 0$ is the value of item j; $a_j > 0$ is the weight of item j; b is the capacity of the knapsack; and n is the number of items to be considered for inclusion in the knapsack.

This problem is known to be NP-complete (Garey and Johnson, 1979).

7 RANDOM GENERATION OF 0-1 KNAPSACK PROBLEMS

Suppose we are interested in evaluating the performance of a solution procedure for the 0-1 knapsack problem with randomly-generated test problems. A typical approach would be to sample n constraint coefficients a_j from the uniform distribution over the integers from 1 to 50 and n objective function coefficients c_j from the uniform distribution over the integers from 1 to 100. The constant b would be some fraction of the sum of the constraint coefficients. For example,

$$b = \left\lceil \frac{1}{2} \sum_{j=1}^{n} a_j \right\rceil . \tag{7}$$

With this approach, the values of $\mathrm{E}(Y_1 Y_2)$ and θ that correspond to the pmf for every problem generated are $\mathrm{E}(Y_1)\mathrm{E}(Y_2)$ and $1/n_1 n_2$, respectively. Only one of the infinite number of pmfs associated with the bivariate random variable (A, C) would ever be used. An alternative approach is to generate problems using pmfs of the form (6) corresponding to various feasible combinations of $\mathrm{E}(Y_1 Y_2)$ and θ.

With either approach, the problems that are randomly generated belong to the same class of problems: those with objective function and constraint coefficients uniformly distributed over specified sets of integer values. In the first case, the plan is to generate problems with objective function and constraint coefficients for each variable that are independent. Because the number of decision variables is finite, this is unlikely to ever occur. However, it would be rare to generate a problem in which there was very strong positive or negative correlation among the problem parameters. With the second approach, the expected strength of the relationship between A and C is varied from problem to problem. The correlation in some problems will be strong, sometimes positive and sometimes negative. In other problems, there will be weak correlation, and sometimes negligible correlation. In either case, 0-1 knapsack problems are "pulled out of a hat". One difference between the two approaches is that the problems are "arranged" in the hat in the first approach, with the problems with weakly correlated coefficients near the top and those with strongly correlated parameters near the bottom. Another difference is the inference space; the second approach allows experimenters to draw more general conclusions.

8 COMPUTATIONAL RESULTS

We solved 200 25-variable randomly-generated 0-1 knapsack problems with a simple implicit enumeration routine. The test problem parameters are assumed to be distributed as follows: $A \sim U[1,50]$ and $C \sim U[1,100]$. The right-hand side constant in the constraint is calculated using (7).

For the distributions we used, $\mathrm{E}(Y_1) = 25.5$, $\mathrm{E}(Y_2) = 50.5$, $\mathrm{Var}(Y_1) = 208.25$, $\mathrm{Var}(Y_2) = 833.25$, $\theta^* = 0.0002$, $K^* = 1287.75$, $\Delta = 0$, $K_{max} = 1704.25$, $K_{min} = 871.25$, $\rho_{max} = 0.99985$, and $\rho_{min} = -0.99985$.

The following procedure was used 100 times with 100 random number seeds to generate the test problems. First, desired values for $\mathrm{E}(Y_1Y_2)$ and θ were randomly generated. Next, two problems were generated using synchronized random numbers: an "independent" problem was generated according to the pmf $g^*(y_1, y_2)$ and a "dependent" problem was generated according to the pmf (6). Finally, both problems were solved to optimality. All computer runs were made on an IBM 3081-D at The Ohio State University.

For the independent problems, the average sample correlation was -0.008. The sample correlations ranged from -0.550 to 0.405. The target correlations for the dependent problems ranged from -0.930 to 0.933, with a mean of 0.042. The average of the sample correlations for the dependent problems was 0.035, and the range was -0.998 to 0.998. Clearly, the coefficients in the independent problems are correlated. However, no independent problem has coefficients that have extreme correlation.

Table 1: Iteration Count Statistics

Statistic	Independent Problems	Dependent Problems
Mean	30,637	253,816
Std. Error	3,931	50,983
Maximum	228,925	2,659,561
Median	17,875	48,521
Minimum	1639	725

Some statistics on the number of iterations to optimality are shown in Table 1. This table indicates that the independent problems are collectively easier to solve than the dependent problems. Additionally, there were 23 dependent problems for which the number of iterations exceeded the maximum number of iterations for any of the independent problems. The independent problem required more iterations than the dependent problem in only 35 pairs of problems. For 32 of these cases, the sample correlation for the dependent problem was negative.

Let $\hat{\rho}$ be a sample correlation for the coefficients in a knapsack problem. Table 2 shows a breakdown of the number of iterations for independent and dependent problems with negative and positive sample correlations. This table illustrates the effect of correlation among knapsack coefficients on the number of implicit enumeration iterations. For the dependent problems, there is exponential growth in the number of iterations as the correlation increases. The same might be said for the independent problems, but the trend is more difficult to discern. Still, independent problems with positively correlated coefficients tend to require more iterations than independent problems with negatively correlated coefficients.

The results of this experiment are not surprising, since they are consistent with the results reported by Martello and Toth (1979), Moore (1989, 1990), and Moore, Peterson, and Reilly (1990). However, our experiment is interesting because we use a different pmf, selected at random, for generating the coefficients in each dependent problem.

Table 2: Sample Correlations and Iterations

	Independent		Dependent	
Iterations	$\hat{\rho} < 0$	$\hat{\rho} > 0$	$\hat{\rho} < 0$	$\hat{\rho} > 0$
100s			3	
1000s	24	7	23	2
10000s	27	37	18	17
100000s	1	4	1	27
1000000s				9

9 DISCUSSION

Empirical evaluations of solution methods for optimization problems are conducted too often using only test problems that are randomly generated under the assumption that all coefficients are probabilistically independent. We view this practice as analogous to the *faux pas* of equating the expected value of a function of a random variable with the value of the function at the expectation of the random variable.

Test problems with controlled dependence among the problem parameters are more likely to be included in computational studies if there are easy-to-use characterizations of the joint pmf of the coefficients. We have presented a simple and general probability redistribution procedure as a means of finding a pmf for the random variable (Y_1, Y_2), given specified values for $E(Y_1 Y_2)$ and the smallest joint probability. We have shown that a pmf for a random variable (Y_1, Y_2) with uniform marginals is a mixture of three simple pmfs when the number of possible values of one random variable is a multiple of the number of possible values of the other random variable and both Y_1 and Y_2 have at least three possible values. Furthermore, this pmf has at most four values over the support of (Y_1, Y_2). Because of its simple form, this pmf may be exploited when values of (Y_1, Y_2) are simulated, making the random generation of optimization test problems with controlled dependencies relatively easy.

This mixture was used in a computational experiment in which 200 0-1 knapsack problems were randomly generated and solved to optimality. One hundred of these problems were generated with the constraint and objective function coefficients assumed to be independent. An additional 100 problems were generated after a target correlation for the two sets of coefficients and a value for the smallest joint probability were selected at random.

The results of the experiment reaffirm that implicit enumeration generally requires more iterations to solve knapsack problems with positive correlation among the coefficients than knapsack problems with negatively correlated or uncorrelated coefficients. Furthermore, the number of iterations to optimality tends to be an exponential function of the correlation induced among the coefficients.

Note that we do not advocate the exclusive use of the uniform distribution in empirical evaluations of solution methods. Rather, we suggest that optimization test problems with a variety of reasonable dependence structures and distributional assumptions be used in such evaluations and we demonstrate that this is straightforward to do when the there are two sets of coefficients, each with a uniform marginal distribution.

Possibilities for related research abound. Two examples are the characterization of pmfs for a multivariate random variable with uniform marginals and experimentation with other types of optimization problems that might be expected to have dependencies among their coefficients.

ACKNOWLEDGEMENTS

We thank the Proceedings Editor, Dr. Barry L. Nelson, for many interesting conversations about this research and Mr. Jorge E. Regalado-Vergara for his programming assistance.

REFERENCES

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

Balas, E. and E. Zemel. 1980. An algorithm for large zero-one knapsack problems. *Operations Research* **28**(5): 1130-1154.

Evans, J. R. 1984. The factored transportation problem. *Management Science* **30**(8): 1021-1024.

Garey, M. R. and D. S. Johnson. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness.* New York:W. H. Freeman and Company.

Kemp, C. D. and H. Papageorgiou. 1982. Bivariate Hermite distributions. *Sankhyā* **44**(Series A, Part 2): 269-280.

Kocherlakota, S. 1989. A note on the bivariate binomial distribution. *Statistics and Probability Letters* **8**: 21-24.

Marshall, A. W. and I. Olkin. 1985. A family of bivariate distributions generated by the bivariate

Bernoulli distribution. *Journal of the American Statistical Association* **80**: 332-338.

Martello, S. and P. Toth. 1979. The 0-1 knapsack problem. *Combinatorial Optimization*, eds. N. Christofides, A. Mingozzi, C. Sandi, 237-279. John Wiley and Sons, New York, New York.

Moore, B. A. 1989. Correlated 0-1 knapsack problems. IND ENG 854 Course Project, Department of Industrial and Systems Engineering, The Ohio State University, Columbus, Ohio.

Moore, B. A. 1990. The effect of correlation on the performance of exact and heuristic procedures for the weighted set covering problem. Master's Thesis, Department of Industrial and Systems Engineering, The Ohio State University, Columbus, Ohio.

Moore, B. A., J. A. Peterson, and C. H. Reilly. 1990. Characterizing distributions of discrete bivariate random variables for simulation and evaluation of solution methods. *Proceedings of the 1990 Winter Simulation Conference*, eds. O. Balci, R. P. Sadowski, R. E. Nance, 294-302. Institute of Electrical and Electronics Engineers, New Orleans, Louisiana.

Nagaraja, H. N. 1983. Two characterizations of a bivariate geometric distribution. *Journal of the Indian Statistical Association* **21**: 27-30.

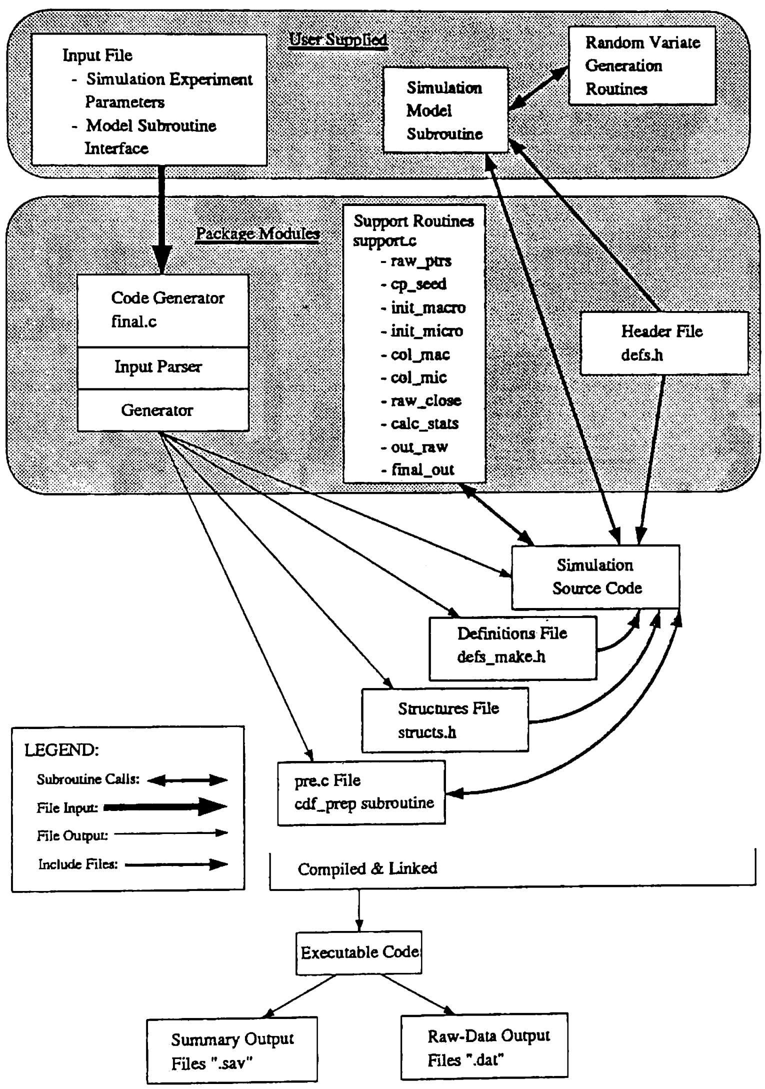
Peterson, J. A. 1990. A parametric analysis of a bottleneck transportation problem applied to the characterization of correlated discrete bivariate random variables. Master's Thesis, Department of Industrial and Systems Engineering, The Ohio State University, Columbus, Ohio

Whitt, W. 1976. Bivariate distributions with given marginals. *The Annals of Statistics* **4**(6): 1280-1289.

AUTHOR BIOGRAPHY

CHARLES H. REILLY is an Associate Professor and Vice Chair in the Department of Industrial and Systems Engineering at The Ohio State University. He earned a B.A. in mathematics and business administration at St. Norbert College in 1979 and M.S. and Ph.D. degrees at Purdue University in 1980 and 1983. His current research interests are in the areas of empirical evaluations of solution methods for optimization problems and the application of operations research techniques to problems in telecommunications. Dr. Reilly is a member of ORSA, TIMS, IIE, Alpha Pi Mu, and Omega Rho and the track coordinator for General Applications at WSC '91.

Analysis Methodology



SERVO system-integration diagram, from "SERVO: Simulation Experiments with Random-Vector Output," by Bruce W. Schmeiser and Mark D. Scott.