Discrete Optimization

A family of composite discrete bivariate distributions with uniform marginals for simulating realistic and challenging optimization-problem instances

Charles H. Reilly [a,*], Nabin Sapkota [b]

[a] Department of Industrial Engineering and Management Systems, College of Engineering and Computer Science, University of Central Florida, Engineering II, Room 312, 12800 Pegasus Drive, Orlando, FL 32816, United States
[b] Department of Engineering Technology, College of Science, Technology and Business, Northwestern State University, Natchitoches, LA 71497, United States

ARTICLE INFO

ABSTRACT

We consider a family of composite bivariate distributions, or probability mass functions (pmfs), with uniform marginals for simulating optimization-problem instances. For every possible population correlation, except the extreme values, there are an infinite number of valid joint distributions in this family. We quantify the entropy for all member distributions, including the special cases under independence and both extreme correlations. Greater variety is expected across optimization-problem instances simulated based on a high-entropy pmf. We present a closed-form solution to the problem of finding the joint pmf that maximizes entropy for a specified population correlation, and we show that this entropy-maximizing pmf belongs to our family of pmfs. We introduce the entropy range as a secondary indicator of the variety of instances that may be generated for a particular correlation. Finally, we discuss how to systematically control entropy and correlation to simulate a set of synthetic problem instances that includes challenging examples and examples with realistic characteristics.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

The simulation, or random generation, of optimization-problem instances with a target population correlation between the values of two types of coefficients, for example, the objective-function and constraint coefficients in the 0-1 Knapsack Problem (KP01), is a primary motivation for this research. To thoroughly evaluate the performance of any algorithm or heuristic, researchers must attempt to solve a convincing number of diverse problem instances that, hopefully, includes challenging and realistic examples. Our work leads to suggestions for simulating collections of such synthetic problem instances.

Bartz-Beielstein, Chiarandini, Paquete, and Preuss (2010) includes excellent papers that underscore the importance of statistics in conducting and interpreting experiments on optimization procedures. Our research contributes to both statistics and operations research by complementing previous empirical experimentation with fundamental results that facilitate better analysis and understanding of the performances of optimization solution methods, another primary motivation for our research.

Our principal contributions here are quantifying entropy for a complete family of composite probability mass functions (pmfs) with uniform marginal distributions, identifying the pmf that maximizes entropy for a specified correlation, and explaining how to systematically control entropy and coefficient correlation when simulating optimization-problem instances with explicit correlation induction (ECI). ECI allows experimenters to simulate instances based on any possible population correlation (Reilly, 1991, 2009). We explain how ECI also allows experimenters to simulate instances based on the full range of possible entropy values for any correlation level.

Suppose that Y_1 and Y_2 are discrete random variables representing the values of two types of coefficients in an optimization problem. The entropy associated with a joint distribution, or pmf $f(y_1, y_2)$, for the random variable (Y_1, Y_2),

$$H_f = -\mathrm{E}(\ln(f(Y_1, Y_2))) = -\sum_{y_1}\sum_{y_2} f(y_1, y_2)\ln(f(y_1, y_2)),$$

measures the uncertainty associated with (Y_1, Y_2) under $f(y_1, y_2)$. With any specified marginal distributions for Y_1 and Y_2, entropy is maximized when Y_1 and Y_2 are independent (Shannon, 1949). Coefficients in synthetic problem instances simulated based on a low-entropy pmf may exhibit patterns among values (y_1, y_2) of (Y_1, Y_2) and, as a result, less than desired variety may be apparent across those instances. So generally speaking, we anticipate that a wider

variety of synthetic problem instances may be simulated based on a high-entropy distribution for a target population correlation $\rho = \text{Corr}(Y_1, Y_2)$.

We expect that coefficients in many optimization problems encountered in practice are correlated. For example, it would make sense that activities associated with more resource units would be the activities that bring more profit. Computational studies on simulated instances of classical optimization problems, including KP01 (see, for example, Martello, Pisinger, & Toth, 1999, 2000; Martello & Toth, 1979, 1987,1988,1997; Pisinger, 1997), the Generalized Assignment Problem (GAP) (Amini & Racer, 1994; Cario et al., 2002; Martello & Toth, 1981), the Capital Budgeting (or Multidimensional Knapsack) Problem (Fréville & Plateau, 1994, 1996; Hill & Reilly, 2000b), and the Set Covering Problem (Rushmeier & Nemhauser, 1993; Sapkota & Reilly, 2011), show that correlation between key types of coefficients affects the performances of algorithms and heuristics. Based on results in Cario et al. (2002), there appears to be a relationship between the relative entropy of the joint distribution of GAP objective-function and capacity-constraint coefficient values and the performances of algorithms and heuristics. If entropy plays a significant role in the performances of algorithms and heuristics, experimenters may learn more about the capabilities and limitations of solution methods when they systematically control the entropy associated with the joint distribution(s) of coefficient values *and* the correlation between important coefficient types. However, heretofore entropy appears to have received little attention as a controllable factor in computational experiments with optimization methods.

Many experimenters use *ad hoc* implicit correlation-induction (ICI) procedures when simulating correlated coefficients in synthetic optimization-problem instances. The user-specified values of an ICI procedure's parameters imply the population-correlation level and determine the joint pmf of coefficient values. Under ECI, an alternative to ICI, users specify either marginal distributions of coefficient values and a target population correlation or a joint distribution of coefficient values. ECI overcomes the principal shortcomings of ICI (Reilly, 2009).

In this paper we provide closed-form expressions for the entropy associated with a family of composite pmfs for (Y_1, Y_2), where Y_1 and Y_2 are uniform random variables and ρ is specified. We consider how entropy is affected by the target correlation and the smallest joint probability value over the support of (Y_1, Y_2), or the relative entropy. The performances of optimization methods may be affected by the correlation-induction approach, as well as by the selection of joint pmfs (Cario et al., 2002). The distribution-selection criterion that we consider here for ECI is maximum entropy, principally because of the greater variety that is expected across the resulting synthetic problem instances. We formulate and solve a convex program to find the maximum-entropy pmf for (Y_1, Y_2) when ρ is specified. For all possible values of ρ, the maximum-entropy pmf belongs to our family of composite distributions.

Our paper is organized as follows. Background information about ICI and ECI problem-generation approaches, especially for KP01, is presented in the next section, along with material on the Factored Transportation Problem. We introduce the first entropy expressions for our family of composite pmfs under ECI in Section 3; some special cases are considered in Section 4. In Section 5, we present a convex programming model to find the maximum-entropy distribution for a discrete bivariate random variable with uniform marginals and a target population correlation. Closed-form solutions that satisfy the Karush–Kuhn–Tucker necessary and sufficient optimality conditions are presented for all possible target correlation values. In Section 6, we introduce the entropy range as a secondary indicator of the variety of optimization-problem instances that may be simulated for a particular correlation. The last section contains some discussion and recommendations for further research.

2. Background

In this section, we review ICI and ECI approaches for simulating coefficients in synthetic instances of optimization problems, particularly KP01. The Factored Transportation Problem has important implications for simulating coefficients for synthetic optimization-problem instances with ECI.

2.1. Simulating coefficients in synthetic optimization-problem instances

KP01 may be represented as follows:

$$\text{Maximize } \sum_{j=1}^{n} p_j x_j$$

$$\text{Subject to } \sum_{j=1}^{n} w_j x_j \le b \qquad \text{(KP01)}$$

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

where all $p_j > 0$, all $0 < w_j < b$, and $\sum_{j=1}^{n} w_j > b$. Kelleher, Pferschy, and Pisinger (2004) presents an extensive survey of exact and heuristic solution methods for KP01 and several knapsack variants.

Let P be the random variable representing the values of the objective-function coefficients and W be the random variable representing the values of the constraint coefficients. Also let ν be a positive integer and δ and γ be non-negative integers. A flexible KP01 problem simulator modeled after those outlined in the literature (e.g., see Balas & Zemel, 1980; Martello et al., 1999, 2000; Martello & Toth, 1979, 1987,1988,1997; and Pisinger, 1997) is:

$$w_j \leftarrow W \sim U\{1, 2, \ldots, \eta\}$$
$$t_j \leftarrow T \sim U\{-\delta, -\delta + 1, \ldots, \delta\}$$
$$p_j = w_j + t_j + \gamma$$

Reilly (1997, 2009) shows that:

$$\rho = \text{Corr}(W, P) = \sqrt{\frac{\eta^2 - 1}{\eta^2 + 4\delta(\delta + 1) - 1}}.$$

However, quantification of the correlation implied under ICI is rare. The implied correlation $\rho > 0$ is usually loosely classified as "weak" if $\delta > 0$ and $\gamma = 0$; as "strong" if $\delta = 0$ and $\gamma \ge 0$; and as "almost strong" if $\delta > 0$ and $\gamma > 0$. Additionally, the correlation level is confounded with the marginal distributions of coefficient values, making it impossible to systematically vary correlation alone when simulating KP01 instances. A similar problem simulator may be used if KP01 instances with negatively correlated coefficients are desired (e.g., see Cario et al., 2002 or Martello and Toth, 1981).

ICI problem simulators have also been developed for other classical optimization problems, including GAP, the Capital Budgeting (or Multi-Dimensional Knapsack) Problem, and the Set Covering Problem. Reilly (2009) examines example ICI simulators for these problems. Additionally, ICI has been used to simulate instances of single-machine scheduling problems (John, 1989; Potts & Van Wassenhove, 1988, 1992).

Alternatively, a joint pmf for KP01 coefficient values may be defined based on specified marginal distributions for the objective-function and constraint coefficient values and a target population correlation. Reilly (1991, 1993) proposes a family of joint pmfs for (W, P) indexed by two parameters: ρ and $\theta = \min_{ij}\{\text{Pr}(W = w_i, P = p_j)\}$, the smallest joint probability over the support of (W, P). These pmfs may be referred to as parametric mixtures (e.g., see Cario et al., 2002). Under such an ECI approach, correlation may be varied and systematically controlled without affecting or being confounded with the marginal distributions of coefficient values (Reilly, 2009). Hall and Posner (2010) recommend principles and properties for designing problem simulators to avoid unintended biases in the instances' coefficients and in the conclusions drawn from experiments with those