

Estimating and Validating the Cumulative Distribution of a Function of Random Variables: Toward the Development of Distribution Arithmetic

Weldon A. Lodwick¹ and K. David Jamison²

1. Department of Mathematics, Campus Box 170, University of Colorado,
P.O. Box 173364, Denver, CO 80217-3364, U.S.A.

2. Watson Wyatt & Company, 950 17th Street, Suite 1400, Denver, CO
80202, U.S.A.

e-mail: Weldon.Lodwick@cudenver.edu , Ken.Jamison@WatsonWyatt.com

Abstract: A method for estimating and validating the cumulative distribution of a function of random variables (independent or dependent) is presented and examined. The method creates a sequence of bounds that will converge to the distribution function in the limit for functions of independent random variables or of random variables of known dependencies. Moreover, an approximation is constructed from and contained in these bounds. Preliminary numerical experiments indicate that this approximation is close to the actual distribution after a few iterations. Several examples are given to illustrate the method.

Keywords: Probability Theory, Interval Analysis, Validation

1 Introduction

A method for estimating the cumulative distribution function (c.d.f.) for a real-valued function of a finite set of random variables is described. Several illustrations of the method are presented. When the number of random variables is moderate, the method gives good results quickly. The estimate is a closed form approximation that might prove useful in stochastic and fuzzy/possibilistic programming problems (see [5] and [6] for early applications to optimization) and perhaps as an alternative to simulations when closed forms are needed. We note the similarity between our method with that of R.E. Moore ([14]) and several others before and after Moore (see [3], [5], [6], [9], [12], [13], [15], [19] and [20], a discussion as to the similarities and differences between these methods and the one we develop here is given in the next section). The method is an extension and application of the ideas discussed in [10] and [11]. In essence, the c.d.f. is bounded by con-

sistent possibility and necessity distributions (consistent with the underlying probability distribution), as will be explained below.

The paper is organized in the following way. First, we discuss the problem we are solving and its relationship between what is presented and interval analysis, validation and what others have done with respect to the problem in which we are interested. A review of what has been done and how our approach is similar and different is given. Second, the algorithm to compute upper/lower bounds and an intermediate estimation to the c.d.f. of a function of random variables is presented. Third, the convergence of the algorithm is proved. Fourth, a brief explanation of extensions to non-monotonic functions is given. Fifth, numerical examples are given and the last section contains the conclusions.

2 Relationships to Interval Analysis, Validation and Various Other Approaches

In this section we look at the relationship to interval analysis, validation and the approaches developed by other researchers.

2.1 The Problem

The problem we are solving is, given $f(\vec{X})$ continuous, where each X_i is a random variable of known distribution, compute the c.d.f. of $f(\vec{X})$. We develop in detail the case where $f(\vec{X})$ is monotone in each variable and indicate how to handle broader classes of functions. An example of how our method works on non-monotone functions is given in the penultimate section. Our approach is to construct bounds $n_k(\vec{X})$ and $p_k(\vec{X})$ for the c.d.f. of $f(\vec{X})$ and an approximation, $\hat{f}(\vec{X})$, to it that possess the following properties:

$$\begin{aligned}
 n_k(\vec{X}) &\leq f(\vec{X}) \leq p_k(\vec{X}) && \text{- VALIDATION} \\
 \lim_k n_k(\vec{X}) &= f(\vec{X}) = \lim_k p_k(\vec{X}) && \text{- CONVERGENCE/RELIABILITY} \\
 &n_k(\vec{X}) \text{ and } p_k(\vec{X}) && \text{- EASILY COMPUTABLE} \\
 n_k(\vec{X}) &\leq \hat{f}(\vec{X}) \approx f(\vec{X}) \leq p_k(\vec{X}) && \text{- CONSISTENT APPROXIMATION}
 \end{aligned}$$

There are at least two areas where problems requiring the computation of the c.d.f. of a function arise, (i) risk analysis; for example, the analysis of

alternative investments, and (ii) stochastic programming and other types of optimization under uncertainty.

2.2 Relationships

What is developed herein is in the spirit of interval analysis in that we use a min/max type of computational strategy which is the approach that is used in interval arithmetic. We show that refined partitions (boxes whose diameters go to zero in the limit) lead to convergence. Second, we are concerned with validation and efficiency in computing tight bounds on functions of distributions. This is the central theme of validation-based techniques implemented on a digital computer.

There are four types of approaches in the literature for the arithmetic of distributions:

1. Compute using the definition which involves the evaluation of the convolutions and other direct methods (see equations (1)-(4) of [19], equations (2) and (3) of [12], or standard texts such as [16]). We do not discuss this method except to say that the prevailing view is that computations via direct methods are complex enough to seek alternative methods, especially when there are dependencies among the variables.

2. Compute using Monte Carlo simulations. We assume the reader is familiar with this approach and do not speak to it. We do, however, compare our approximations to the results obtained from 10,000 Monte Carlo simulations. It is mentioned that a drawback to Monte Carlo simulations is that a closed form is not obtained. Our method does. Closed forms are especially important in optimization where derivative information is used.

3. Approximate distribution arithmetic using quantile/histograms (see [2],[3], [5], [6], [12], and [15]) and compute directly with the resulting histograms. Upper/lower bounds on the c.d.f. are obtained by integrating across the left side&top/bottom&right side of the histograms. These are then pieced together to form an upper and lower c.d.f. that guarantees that the actual c.d.f is enclosed (validates). When the random variables are dependent, then a mathematical programming problem ensues for the histogram calculation. Berleant's approach, [2], [3], is the most general of these methods in that dependencies (both known and unknown) are explicitly handled where [5], [6] and [12] do not handle dependencies. Like Williamson and Downs [19] and [20], Berleant does not assume knowledge of dependencies.

4. Approximate distribution arithmetic with dependencies bounds using

Fréchet bounds (see for example [19] or [20]). In [19] and [20], it is shown how to translate the Fréchet bounds into copulas. The copulas are in turn related to dependency bounds on the associated marginal distributions. Inverse and quasi-inverses of the marginals need to be computed to calculate the dependency bounds. The dependency bounds that are worked out in [19] and [20] require monotonicity in both variables of the binary operations. The work of [7], [8] and [9] extend [19] and [20]. Ferson and his co-researchers develop methods to obtain upper/lower bounds on distributions. Once one has bounds, they compute arithmetically using the methods of [19] and [20]. Moreover, [19] and [20] obtain upper and lower estimations given no information about the dependencies. Ferson and his co-workers show how to incorporate additional information (e.g. the type of distribution along with the mean belonging to an interval, the value of the median, and so on) to tighten the bounds.

2.2.1 Similarities

Our method has parts that are similar to the histogram/quantile-based methods of [13], [14] [2], and [3] and to the Fréchet bound methods of [19], [20], [7], [8] and [9]. Our methods are similar to Williamson&Downs method in the way we calculate upper/lower bounds (see Figure 1 below and Figure 1, page 100 of [20]). Williamson&Downs have an optimization step (maximize the upper approximation and minimize the lower). We do not have this step. Moreover, we are similar to Ferson in that we use information about the distribution to obtain tighter bounds. In our case, we incorporate knowledge of the dependencies (see equation (2)). However, like Williamson&Downs and Ferson, our methods would also work without information about the dependencies.

The method developed herein is similar to [13] and [14] in that c.d.f.s are computed using min/max interval-type methods. Moore assumes uniformity on each subdivision. Berleant extends Moore to obtain lower c.d.f.s (by integrating on the left side and top of subdivisions/histograms) and upper c.d.f.s (by integrating on the bottom and right side of the subdivisions/histograms). Our method is also similar to the histogram-based methods in that we obtain upper/lower bounds over subdivisions that are similar Berleant. Since assume the distribution is known, we obtain tighter and "smoother" bounding functions (compare Figures 2-5 below to Figure 3, page 153 of [3]). This is because Berleant assumes no knowledge of the distribution within the sub-

divisions.

2.2.2 Differences

Our method is also different. Firstly, we are interested in the distribution of a function of random variables as opposed to the arithmetic of distributions. Since we are interested in obtaining the distribution of a function of n -variables, one focus is on efficiency of computations and hence, for example, we avoid the optimization that Williamson&Downs and Berleant do. Secondly, we compute an approximation to the actual distribution unlike any of the methods other than Moore. This approximation is more than picking the "midpoint" and is (from our numerical examples) very close to the Monte Carlo solution. Thirdly, we partition and obtain conditionals over boxes that are subsets of the set

$$\{[F_{X_1}^{-1}(0), F_{X_1}^{-1}(1)] \times \dots \times [F_{X_n}^{-1}(0), F_{X_n}^{-1}(1)]\}. \quad (1)$$

Here and in what follows, we mean by $F_X^{-1}(0)$ (and $F_X^{-1}(1)$) the last point coming from the left (respectively the first point going to the right). Our approach computes inequalities that bound the c.d.f. of $f(\vec{X})$ based on min/max calculations over boxes of the set (1). This is quite different from [2], [3], [13], and [14] who work on partitions of the domain of the random variables. In the sense that [19] and [20] compute inner and outer bounds on the distribution using inequalities based on inverses, our methods are similar. However, we base our inequalities on a very straightforward result (see (2) below) and not on Fréchet bounds and associated copula theory.

We obtain convergence using a partitioning approach. More importantly, since we are working with any given distribution on the domain of variables, we use the inverse of the given c.d.f.s to parameterize our approximation which leads to very good approximations as measured by comparing our approximation to that resulting from Monte Carlo simulations. In fact, the approximations that we obtain converge, when we take finer and finer partitions, much more rapidly to the underlying distribution than do the upper and lower bounds.

3 Computing Closed-Form Approximations of the c.d.f. of a Monotone Function

Let $Y = f(\vec{X})$ where the $\vec{X} = (X_1, \dots, X_n)$ is a vector of continuous random variables with joint distribution function $F_X(x)$; i.e.,

$$F_X(x) = \text{prob}(X_1 \leq x_1, \dots, X_n \leq x_n)$$

with marginals F_{X_i} and

$$G_X(x) = \text{prob}(x_1 \leq X_1, \dots, x_n \leq X_n)$$

(see; for example, [4]). We assume f is continuous and nondecreasing in each x_i (it is a simple adjustment to consider functions that increase in some variables and decrease in others) and that the support of each X_i is bounded with $\text{supp}(X_i) = [F_{X_i}^{-1}(0), F_{X_i}^{-1}(1)]$.

We construct bounds (upper/lower) and an approximation between the bounds for the c.d.f. of Y . There are three steps to the method. The first step **partitions** the domain space into smaller boxes. The second step **constructs** upper/lower bounds and an intermediate estimate for the conditional c.d.f. of Y for each box of the partition given X is in that box. The final step **combines** the conditional c.d.f.s into the final estimate.

3.1 Step 1: Partition

The first step is to construct a partition on the domain,

$$[F_{X_1}^{-1}(0), F_{X_1}^{-1}(1)] \times \dots \times [F_{X_n}^{-1}(0), F_{X_n}^{-1}(1)].$$

We do this by dividing each interval into subintervals equally spaced in probability relative to the marginal distributions (see note below). For example, to divide $[F_{X_i}^{-1}(0), F_{X_i}^{-1}(1)]$ into three pieces of equal probability we use $[F_{X_i}^{-1}(0), F_{X_i}^{-1}(\frac{1}{3})]$, $[F_{X_i}^{-1}(\frac{1}{3}), F_{X_i}^{-1}(\frac{2}{3})]$ and $[F_{X_i}^{-1}(\frac{2}{3}), F_{X_i}^{-1}(1)]$. We are not concerned with overlap since the distribution of X is assumed to be continuous. The primary consideration in this process is how it affects the size of the problem. If there are n random variables and each variable X_i is divided into k_i subintervals then we will have $\prod_{i=1}^n k_i$ conditional c.d.f.s to compute. For this study, we do not explore partitioning strategies. Therefore, in our

approach, it is desirable to minimize the number of subdivisions and only subdivide the variables that influence the results the most.

Note: Approximating of the actual c.d.f. using our approach is most difficult in $[F_Y^{-1}(0), F_Y^{-1}(0 + \delta)]$ and in $[F_Y^{-1}(1 - \delta), F_Y^{-1}(1)]$ since there are no overlaps. The best results for the intermediate approximation are obtained from overlaps where there is a cancellation of over and under approximations as will be seen in the sequel. For simplicity of exposition, we use boxes derived from partitions equally spaced in probability.

3.2 Step 2: Construct Upper/Lower Bounds and an Approximation

The second step is to construct the bounds and the estimated conditional c.d.f. for each box of the partition. Let $[b_1, c_1] \times \dots \times [b_n, c_n]$ be one such box and let A denote the event X falls in this box. Consider the family of n -dimensional boxes $\left\{ [b_1, F_{X_1|A}^{-1}(\beta)] \times \dots \times [b_n, F_{X_n|A}^{-1}(\beta)] \mid \beta \in [0, 1] \right\}$. From our assumption that f is continuous and increasing in each X_i we know that

$$f\left([b_1, F_{X_1|A}^{-1}(\beta)] \times \dots \times [b_n, F_{X_n|A}^{-1}(\beta)]\right) = \left[f(b_1, \dots, b_n), f\left(F_{X_1|A}^{-1}(\beta), \dots, F_{X_n|A}^{-1}(\beta)\right)\right]$$

Thus,

$$F_{Y|A}\left(f\left(F_{X_1|A}^{-1}(\beta), \dots, F_{X_n|A}^{-1}(\beta)\right)\right) \geq F_{X|A}\left(F_{X_1|A}^{-1}(\beta), \dots, F_{X_n|A}^{-1}(\beta)\right).$$

Now consider the n -dimensional box $[F_{X_1|A}^{-1}(\beta), c_1] \times \dots \times [F_{X_n|A}^{-1}(\beta), c_n]$. As before, we know that

$$f\left([F_{X_1|A}^{-1}(\beta), c_1] \times \dots \times [F_{X_n|A}^{-1}(\beta), c_n]\right) = \left[f\left(F_{X_1|A}^{-1}(\beta), \dots, F_{X_n|A}^{-1}(\beta)\right), f(c_1, \dots, c_n)\right].$$

This gives the inequality

$$1 - F_{Y|A}\left(f\left(F_{X_1|A}^{-1}(\beta), \dots, F_{X_n|A}^{-1}(\beta)\right)\right) \geq G_{X|A}\left(F_{X_1|A}^{-1}(\beta), \dots, F_{X_n|A}^{-1}(\beta)\right).$$

Put together we have

$$\begin{aligned} F_{X|A}\left(F_{X_1|A}^{-1}(\beta), \dots, F_{X_n|A}^{-1}(\beta)\right) &\leq F_{Y|A}\left(f\left(F_{X_1|A}^{-1}(\beta), \dots, F_{X_n|A}^{-1}(\beta)\right)\right) \\ &\leq 1 - G_{X|A}\left(F_{X_1|A}^{-1}(\beta), \dots, F_{X_n|A}^{-1}(\beta)\right). \quad (2) \end{aligned}$$

When the random variables, \vec{X}_i , are independent this becomes

$$\beta^n \leq F_{Y|A} \left(f \left(F_{X_1|A}^{-1}(\beta), \dots, F_{X_n|A}^{-1}(\beta) \right) \right) \leq 1 - (1 - \beta)^n.$$

This is a wide envelope, particularly for a large number of variables (large n). We wish to produce a reasonable estimate of the c.d.f. without having to perform the calculations needed reduce the envelope to a reasonable width. To do this we select an intermediate value $\hat{F}_{Y|A}$ for

$$F_{Y|A} \left(f \left(F_{X_1|A}^{-1}(\beta), \dots, F_{X_n|A}^{-1}(\beta) \right) \right)$$

that falls between the upper and lower estimate above. One estimate would be to average these probabilities; i.e., set

$$\begin{aligned} & \hat{F}_{Y|A} \left(f \left(F_{X_1|A}^{-1}(\beta), \dots, F_{X_n|A}^{-1}(\beta) \right) \right) \\ &= \frac{1}{2} \left(F_{X|A} \left(F_{X_1|A}^{-1}(\beta), \dots, F_{X_n|A}^{-1}(\beta) \right) + 1 - G_{X|A} \left(F_{X_1|A}^{-1}(\beta), \dots, F_{X_n|A}^{-1}(\beta) \right) \right). \end{aligned}$$

When the random variables, \vec{X}_i , are independent a reasonable choice for the intermediate estimate function, \hat{F} , is to use β . This works since $\beta^n \leq \beta \leq 1 - (1 - \beta)^n$ and has several desirable properties. First is it's simplicity. Second is that this estimate does not increase the maximum possible error in making a choice of intermediate value. This is so because the maximum of the difference $1 - (1 - \beta)^n - \beta^n$ occurs when $\beta = .5$ and at this value the midpoint estimate is $\frac{1}{2}(.5^n + 1 - .5^n) = .5$. A third property is that it is symmetric about the value $\beta = .5$. Any tendency to over/under estimate the true value when $\beta < .5$ should be offset by a tendency to under/over estimate for values of $\beta > .5$. So for independent \vec{X} we use

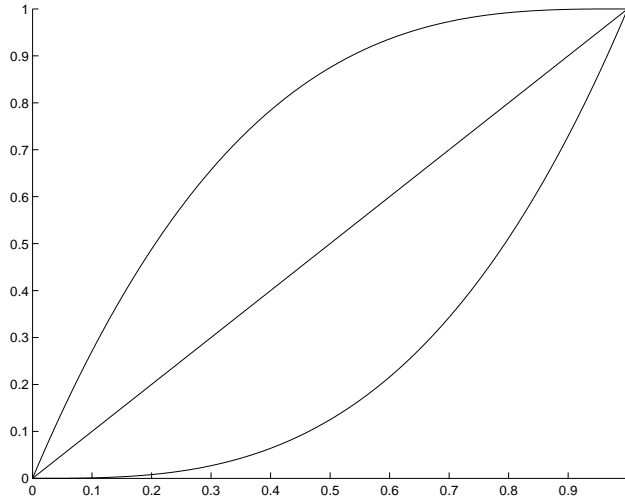
$$F_{Y|A}^- \left(f \left(F_{X_1|A}^{-1}(\beta), \dots, F_{X_n|A}^{-1}(\beta) \right) \right) = \beta^n \quad (3)$$

$$\hat{F}_{Y|A} \left(f \left(F_{X_1|A}^{-1}(\beta), \dots, F_{X_n|A}^{-1}(\beta) \right) \right) = \beta \quad (4)$$

and

$$F_{Y|A}^+ \left(f \left(F_{X_1|A}^{-1}(\beta), \dots, F_{X_n|A}^{-1}(\beta) \right) \right) = 1 - (1 - \beta)^n \quad (5)$$

to obtain a lower bound, intermediate estimate and upper bound on the actual c.d.f., $F_{Y|A}(y)$. Note that these (equations (3), (4) and (5)) define three functions of β , $H_k(\beta) : [0, 1] \rightarrow [0, 1]$, $k = 1, 2, 3$. When $n = 3$ the graphs of H_k for each of the three estimates are as follows:



Plots of $H_1(\beta) = 1 - (1 - \beta)^3$, $H_2(\beta) = \beta$, $H_3(\beta) = \beta^3$

The actual distribution lies between the upper and lower bounds while the estimate $H_2(\beta) = \beta$ gives a centrally located estimate. This is the source of the averaging of errors as the number of subdivisions increase. As the number of variables becomes large, the upper and lower bounds may become wide and the intermediate estimate becomes more important in order to keep the number of subdivisions low (otherwise combinatorial explosion makes the problem intractable).

Figure 1 illustrates the estimate and bounds for $F_Y(0.1)$ when $Y = X_1X_2$ and X_i are i.i.d. uniform $[0, 1]$. The range has not been subdivided. This example shows the inner and outer measures of the area bounded to the left of the level set $Y = X_1X_2 = 0.1$, and the area computed for the intermediate estimate which overlaps the level set, illustrating the inequalities (2), (3), (4) and (5).

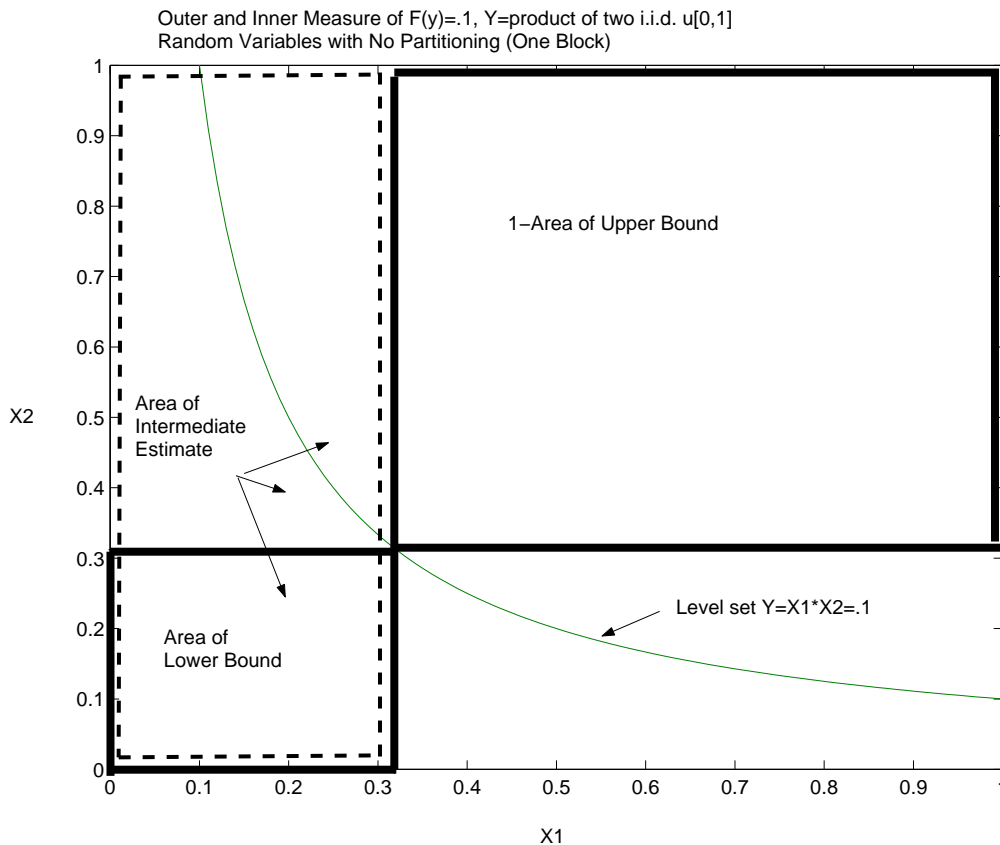


FIGURE 1

3.3 Step 3: Combine

The last step of the method is to combine the estimated conditional c.d.f.s into an approximation of the c.d.f. for Y . Assume we have divided the support of each X_i into subintervals creating a partition of the support of \vec{X} . If $\{A_j \mid j = 1, m\}$ is the partition (so each A_j is also an n -dimensional box) and if $F_{Y|A_i}^-(y)$, $\hat{F}_{Y|A_i}(y)$ and $F_{Y|A_i}^+(y)$ are the values calculated as above for the lower bound, intermediate approximation, and the upper bound (respectively) of the c.d.f. for the variable $Y \mid X \in A_i$, then we can combine these c.d.f.s to produce the bounds and estimate for the c.d.f. of interest. For example, set $\hat{F}_Y(y) = \sum_{j=1}^m \hat{F}_{Y|A_i}(y) P(X \in A_j)$ (where $P(E)$ equals the probability of the event E). The upper and lower bounds ($F_Y^-(y)$ and

$F_Y^+(y)$) are similarly calculated.

4 Convergence

This process will converge to the actual c.d.f. if the supports for each X_i are subdivided into finer and finer subintervals since $F_Y^-(y)$ and $F_Y^+(y)$ are simply inner and outer measures (relative to the measure defined by F_X) for the region $\{x \mid f(x) \leq y\}$.

Theorem 1 *Given $f(\vec{X})$ continuous and monotonically increasing in each variable, then the lower and upper bound converge to the c.d.f. F_X of $f(\vec{X})$; that is, $F_Y^-(y) \rightarrow F_Y$ from below and $F_Y^+(y) \rightarrow F_Y$ from above*

Proof (Sketch): Since f is monotone and thus measurable it can be approximated with arbitrary precision by a simple step function φ that takes on a finite number of values (steps). Since f is monotone we can chose $\varphi \leq f$ at each point. Each measurable set $A_i = \{x \mid \varphi(x) = a_i, i = 1, \dots, I\}$ can be approximated with arbitrary precision by a finite number of n -dimensional rectangles, $\cup_{j=1}^J R_{ij} \subseteq A_i$. The calculation $F_Y^-(y)$ from the partition that includes each end point of each n -dimensional box R_{ij} ($i = 1, \dots, I, j = 1, \dots, J$) gives an arbitrarily close approximation to F_Y .(see [1] or [17]).□

5 Extension to Non-Monotonic Functions

Non-monotonic functions are more challenging. The initial approach is to discover the regions over which the function is monotonic. Our example below (see Figure 6) uses this approach. This is similar to what [20] suggests (see pages 139-142). Without knowing where the regions of monotonicity are, upper and lower estimates over the boxes using a global optimization technique would be required. Clearly, this would make the problem more complex.

6 Examples and Numerical Results

We consider four examples. The first is the sum of twenty independent identically distributed (i.i.d.) uniform on $[0, 1]$ random variables. Here the

resulting c.d.f. is a bell shaped curve, with mean 10 and about 1.3 for a standard deviation. The second example is the function $Y = (\max\{X_1^3, X_2\})^2 + X_1X_2X_3$ where X_i are i.i.d. uniform on $[0, 1]$ random variables. The third is the computation of the c.d.f. associated with an investment strategy containing various percentages of thirteen asset classes. The fourth is a non-monotonic function.

6.1 Bounds on the c.d.f. of the sum of 20 independent identically distributed uniform on $[0, 1]$ random variables

This first example computes the bounds and approximation on the c.d.f. of

$$Y = \sum_{n=1}^{N=20} X_n.$$

The strong law of large numbers states that the distribution for Y should be quite concentrated about its mean and indeed, the resulting probability density function approaches the standard bell-shaped curve with mean at 10 and standard deviation of about 1.3. The associated c.d.f. is a symmetric "S" shaped curve that is 0 on $(-\infty, 0)$ and 1 on $(20, \infty)$.

The strategy for computing the bound and estimate for this distribution is to take advantage of the separability of the random variables by independent sums. We do this by calculating the distribution two variables at a time as follows:

Step 1: Calculate the bound and estimate for the c.d.f. for $Y_1 = X_1 + X_2$.

Step 2: Calculate the bound and estimate for the c.d.f. for $Y_2 = Y_1 + X_3$

...

Step 19: Calculate the bound and estimate for the c.d.f. for $Y = Y_{19} + X_{20}$.

The advantage of this approach is the lower number of subdivisions needed to obtain a good approximation. For example if we were to subdivide each variable by 10, the number of subdivisions needed to calculate Y in one step would be 10^{20} compared to the number in the approach above which for ten subdivisions is $19(10^2) = 1900$.

The result of the calculations are illustrated in the following graphs.

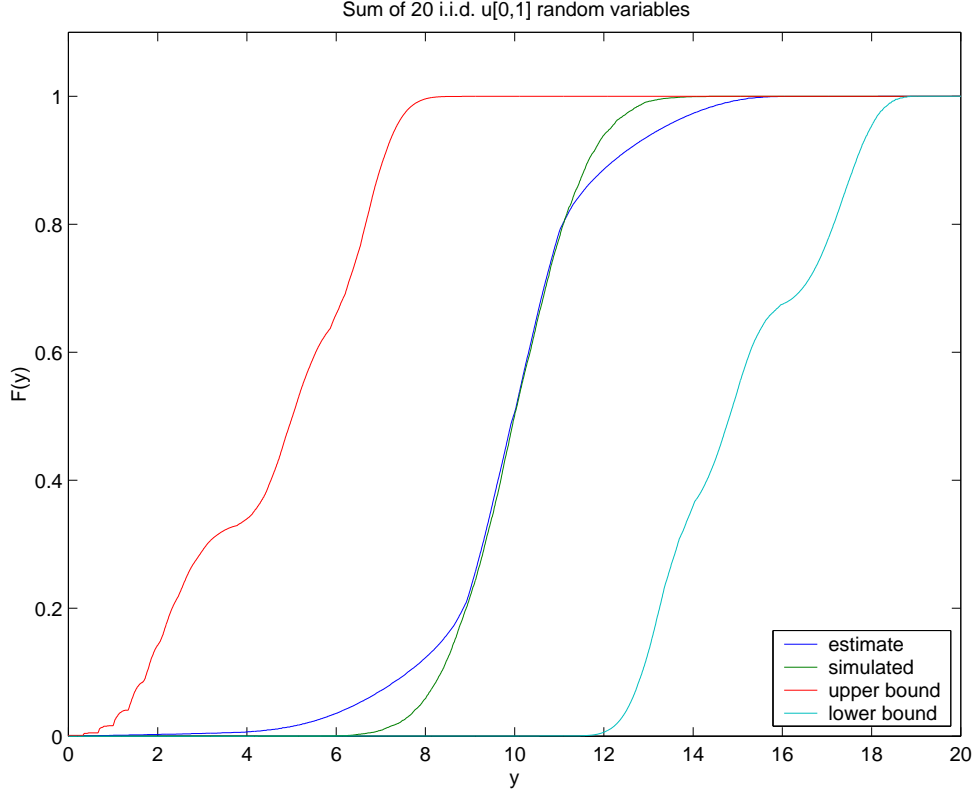


FIGURE 2

6.2 Bounds on the c.d.f. with dependencies

Consider $Y = (\max\{X_1^3, X_2\})^2 + X_1 X_2 X_3$ where X_i are i.i.d. uniform on $[0, 1]$ random variables. Then on $A = [a, b] \subseteq [0, 1]$ $F_{X|A}(x) = \frac{x-a}{b-a}$ and the inverse of this is $F_{X|A}^{-1}(\beta) = \beta(b-a) + a$. Assume we have subdivided $[0, 1]^3$ into 512 boxes by dividing each interval $[0, 1]$ into eight intervals $[0, \frac{1}{8}]$, $[\frac{1}{8}, \frac{2}{8}]$, ..., $[\frac{7}{8}, 1]$ (step 1). For example, on $A = [0, \frac{1}{8}] \times [\frac{7}{8}, 1] \times [\frac{2}{8}, \frac{3}{8}]$ we have $F_{X_1|A}^{-1}(\beta) = \frac{1}{8}\beta$, $F_{X_2|A}^{-1}(\beta) = \frac{1}{8}\beta + \frac{7}{8}$ and $F_{X_3|A}^{-1}(\beta) = \frac{1}{8}\beta + \frac{2}{8}$. Next we bound and estimate the conditional c.d.f. at $y = (\max\{(\frac{1}{8}\beta)^3, \frac{1}{8}\beta + \frac{7}{8}\})^2 + \frac{1}{8}\beta (\frac{1}{8}\beta + \frac{7}{8}) (\frac{1}{8}\beta + \frac{2}{8})$ by $F_{Y|A}^-(y) = \beta^3$, $\hat{F}_{Y|A}(y) = \beta$ and $F_{Y|A}^+(y) = 1 - (1 - \beta)^3$ (step 2). Then the c.d.f. for Y is estimated by summing over the 512 conditional c.d.f.s multiplied by $(\frac{1}{8})^3$, the probability that all three random variables fall into any one of the 512 boxes (step 3). The result of this calculation compared

to the empirical distribution function for 10000 Monte Carlo simulations is as follows.

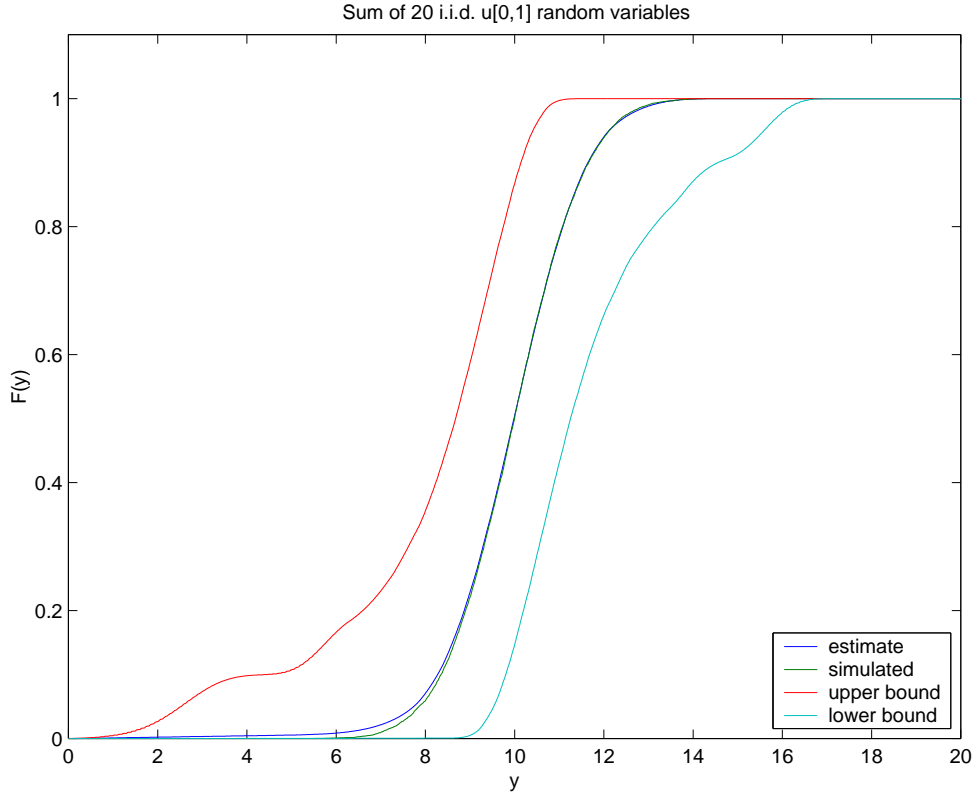


FIGURE 3

6.3 Bounds on the c.d.f. of thirteen investments

In this example we calculated the distribution for the one year return on a portfolio invested in thirteen asset classes. We assumed the return on each asset class was lognormally distributed and that the distributions for each asset class and their dependencies are specified by their means, standard deviations and correlation coefficients. Using these specifications we used the model $Y = e^{(\vec{X}A+m)}$ and the rate of return, $RofR = Y * alloc'$, where \vec{X} is a vector of thirteen i.i.d. $n(0, 1)$ random variables, A and m are calculated so that Y has the means, standard deviations and correlations desired (see [18])

and *alloc* is a vector whose elements sum to one, representing the beginning of year allocation among the asset classes. The result of the calculations is given below in Figure 5.

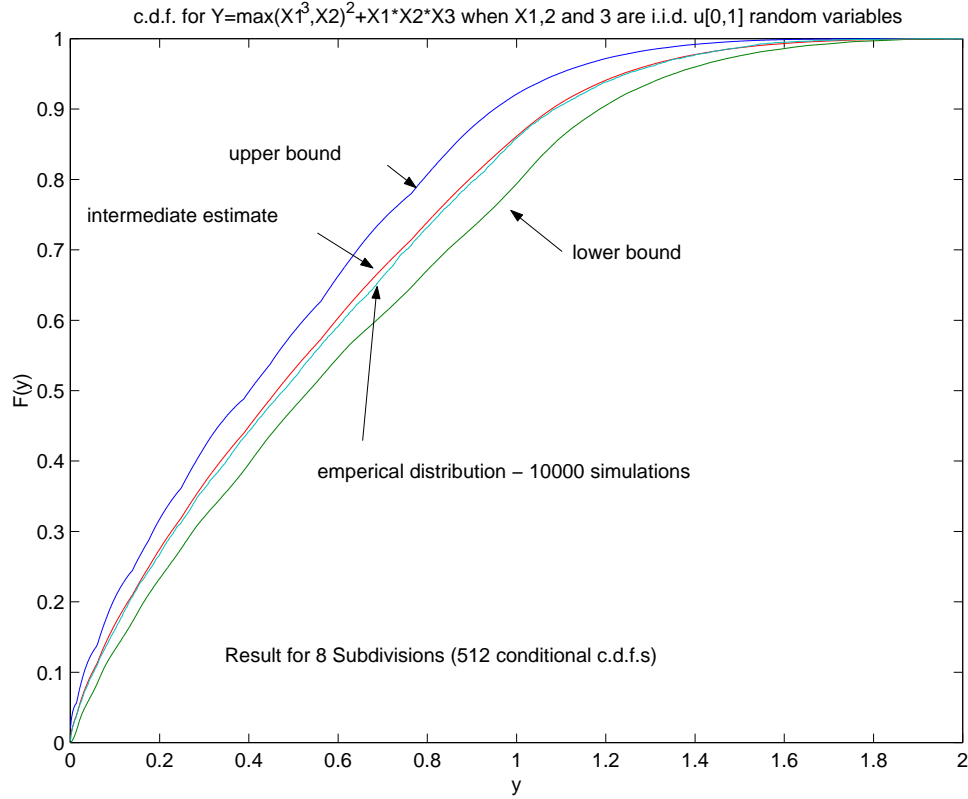


FIGURE 4

6.4 Bounds on the c.d.f. for a nonmonotone function

Let $Y = 0.5(X_1 + 3) \sin(X_2 X_3 + 3) + 3$ where $X_1 \sim u[-3, 3]$ and X_2, X_3 are i.i.d. $n(0, 1)$. The result of the calculation using sixteen subdivisions of each variable follows.

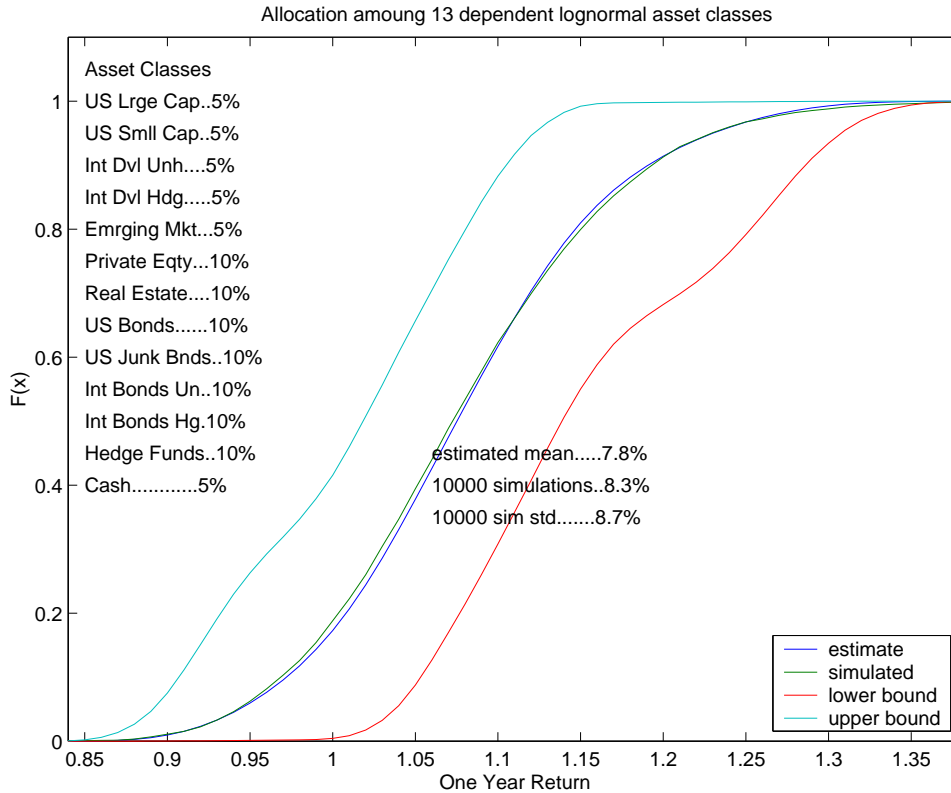


FIGURE 5

The (naive) strategy used for this problem was to first determine the points where the direction of the function $f(t)$ changed sign and divided the range of each variable at these points. We further divided each resulting region in two.

7 Conclusion

The method outlined herein gives a reasonable estimate of the c.d.f. in reasonable computational times without optimized code. The intermediate estimate for the c.d.f. for up to twenty independent random variables has been calculated to a good degree of accuracy in several minutes on a laptop even while the upper and lower bounds are still quite wide. Further research is needed to make the method of greater use when a large number of variables

are involved, to extend the results to non-monotone functions and to the selection of better intermediate estimates. Moreover, partitioning strategies need to be explored.

References

- [1] S.K. Berberian, *Measure and Integration*. Chelsea Publishing Company, Bronx, New York, 1970.
- [2] D. Berleant, "Automatically verified reasoning with both intervals and probability density functions," *Interval Computations 2* (1993), pp. 48-70.
- [3] D. Berleant and C. Goodman-Stauss, "Bounding results of arithmetic operations of random variables of unknown dependencies using interval arithmetic," *Reliable Computation 4* (1998), pp. 147-165.
- [4] L. Breiman. *Probability*, SIAM, Philadelphia, 1992.
- [5] M.A.H. Dempster, "Distributions in interval and linear programming," in *Topics in Interval Analysis* (edited by E. R. Hansen), Oxford Press, 1969, pp. 107-127.
- [6] M.A.H. Dempster, "An application of quantile arithmetic to the distribution problem in stochastic linear programming," *Bulletin of the Institute of Mathematics and its Applications*, 10 (1974), pp. 186-194.
- [7] S. Ferson, "Quality assurance for Monte Carlo risk assessment," *IEEE Proceedings of ISUMA-NAFIPS'95*.(1995), pp. 14-19.
- [8] S. Ferson, "Probability bounds analysis software," *Computing in Environmental Resource Management: Proceedings of a Speciality Conference*, Research Triangle Park, NC, December 2-4, 1996, Pittsburgh, PA, Air and Waste Management Association, pp. 669-677.
- [9] S. Ferson, L. Ginzburg, and R. Akçakaya, "Whereof one cannot speak: when input distributions are unknown," *Risk Analysis* (to appear).
- [10] K.D. Jamison and W.A. Lodwick, "The Construction of Consistent Possibility and Necessity Measures," *Fuzzy Sets and Systems* (accepted 2002).

- [11] K.D. Jamison, W.A. Lodwick and M. Kawai, "A simple closed form estimation for the cumulative distribution function of a monotone function of random variables," *UCD/CCM Report No. 187*, May 2002.
- [12] S. Kaplan, "On the method of discrete probability distributions in risk and reliability calculations - Applications to seismic risk assessment," *Journal of Risk*, 1(3), (1981), pp. 189-196.
- [13] A.S. Moore, "Interval Risk Analysis of Real Estate Investment: A Non-Monte Carlo Approach," *Freiburger Intervall-Berichte* 85/3 (1985), pp. 23-49.
- [14] R.E. Moore, "Risk Analysis without Monte Carlo methods," *Freiburger Intervall-Berichte* 84/1 (1984), pp. 1-48.
- [15] K. Nickel, "Triplex-Algol and its applications," in *Topics in Interval Analysis* (edited by E. R. Hansen), Oxford Press, 1969, pp. 10-24.
- [16] S. Ross, *A First Course in Probability*. Macmillan Publishing Co., New York, 1976.
- [17] H.L. Royden, *Real Analysis*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1988.
- [18] Shapiro, M. and Wilcox, D., "Generating non-standard multivariate distributions with an application to mismeasurement in the CPI," *Technical Working Paper 196*, National Bureau of Economic Research, 1050 Massachusetts Avenue, Cambridge, MA 02138, May 1996.
- [19] R.C. Williams, "Interval arithmetic and probability arithmetic," in C. Ullrich (editor), *Contributions to Computer Arithmetic and Self-Validating Numerical Methods*, J.C. Baltzer AG, IMACS, 1990, pp. 67-80.
- [20] R.C. Williamson and T. Downs, "Probabilistic Arithmetic I: Numerical Methods for Calculating Convolutions and Dependency Bounds," *International Journal of Approximate Reasoning* 4 (1990), pp. 89-158.