A Software Tool for Automatically Verified Reasoning with Intervals and Probability Distributions

Hang Cheng and Daniel Berleant

Abstract— **This work reports a software tool for performing automatically verified arithmetic operations on independent operands when these operands are intervals, are probability distribution functions, or one operand is an interval and the other is a distribution. Because our histogram based representation can express both intervals and distributions, the algorithms do not need to distinguish between intervals and distributions in their operation. While our previous work (Berleant 1993) discussed concepts and algorithms but did not emphasize computer implementations, here we emphasize a PC-based calculating tool that implements the algorithms, interacts with the user via a graphical user interface, and saves, retrieves, and prints the results of its calculations.**

I. Introduction to the Representation

We take an interval to be an incompletely specified description of some value such that there is a probability of one that the actual but unknown value falls within that interval. A probability distribution function may be either a probability density function (PDF) or its integral, a cumulative distribution function (CDF). Thus (excepting nontypical cases, which we do not address) the same distribution function can be described as either a PDF or a CDF, a CDF is the integral of a corresponding PDF, and a PDF is the derivative of a corresponding CDF.

Since the interval representation of value does not assume any particular distribution of probability mass within the interval, the interval representation is consistent with any cumulative distribution whose value is zero below the interval and one above it. To represent the interval using CDF's, we determine upper and lower CDF's which together bound the space of CDF's consistent with the interval. These two most extreme CDF's are the CDF that rises faster than any other CDF consistent with the interval, and the CDF that rises slower than any other

CDF consistent with the interval. The CDF that rises fastest is obviously the one corresponding to the situation in which the variable's value is certainly no higher than the interval's low bound (and can't be lower than the interval's low bound because we assume it lies within the interval with probability one). Likewise, the CDF that rises slowest is obviously the one corresponding to the situation in which the variable's value is certainly equal to the interval's high bound. The interval is therefore representable as a family of CDF's whose two bounding CDF's each have one "step." Figure 1 shows the interval $[0, 1]$ and its representation as a bounded family of CDF's.
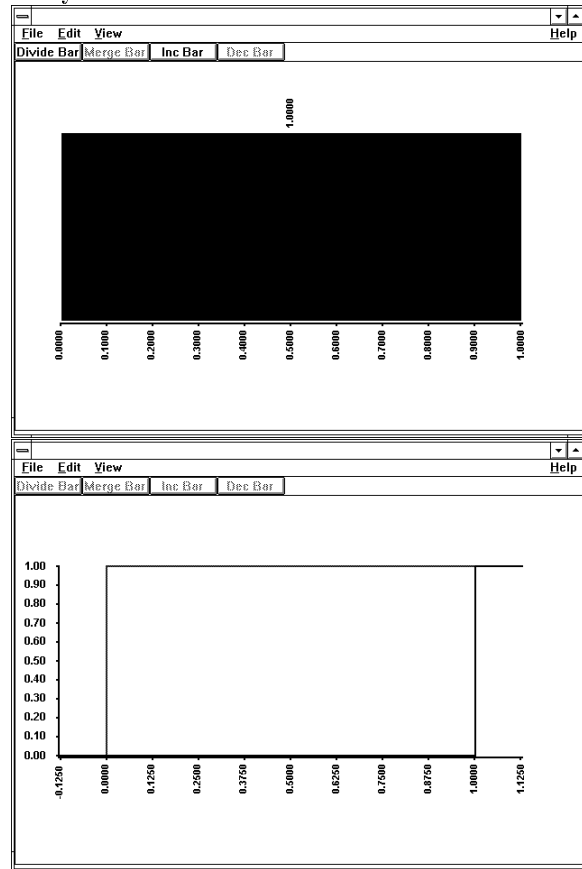

Figure 1: An interval and its representation as two CDF's bounding a family of CDF's.

The authors are with the Dept. of Computer Systems Engineering, University of Arkansas, Fayetteville, AR 72701. Contact: djb@engr.uark.edu .

Some other CDF's may be represented using multiple steps. A CDF corresponding to a set of sample values, for example, would have one step for each distinct sample value (Figure 2).
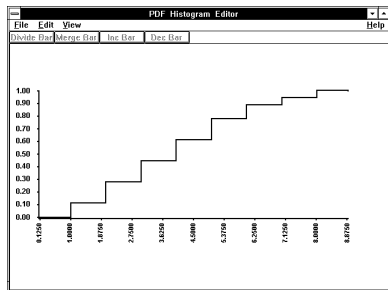

Figure 2: A CDF corresponding to a set of sample values.

Many CDF's do not have a stepwise form. For example, many are smooth curves. Our stepwise representation captures these CDF's correctly, by bounding them from above and below with stepwise CDF's that we can represent. Of course, while this method of representing non-stepwise CDF's is correct in that it does not introduce incorrect approximations, it does lose information due to its use of bounding CDF's that are consistent with other CDF's besides the one they are generated to represent. The user generates these CDF bounds interactively, using a graphical histogram editor to edit a PDF-like histogram which is automatically integrated into bounding CDF's: each bar of the histogram consists of an interval and a corresponding probability, and is treated analogously to the interval whose integration was described previously. (Figure 3 shows a user-generated histogram and its corresponding CDF bounds.)
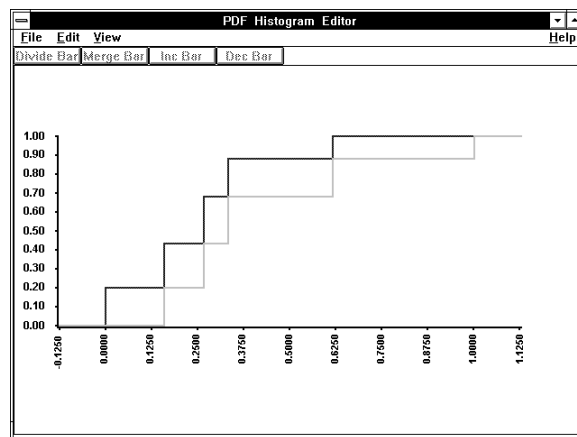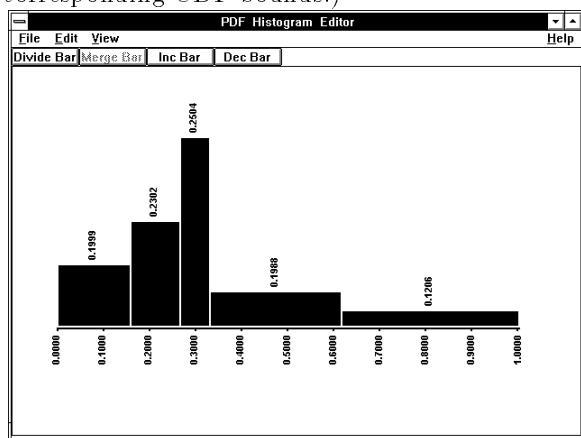

Figure 3: A histogram edited by a user and its corresponding CDF bounds.

Berleant (1993) describes in more detail how a histogram can correctly (although with information loss) represent a smooth PDF curve; the integral of the histogram forms stepwise CDF's bounding the smooth CDF corresponding to the smooth PDF.

Not all pairs of bounding stepwise CDF's can be used by our algorithms. A pair of bounding CDF's must correspond to the integral of a set of intervals, each of which is associated with its probability. While intervals in those sets can overlap or subsume one another, the integrals are still generated analogously to the way a single interval was integrated into single-step CDF bounds (Figure 1) and a histogram was integrated into a multi-step CDF bounds (Figure 2). The common approach to integrating intervals, histograms, and sets of possibly overlapping intervals is due to the fact that each is a set of intervals and associated probability masses.

> *the underlying representation of CDF bounds in our system is a set of intervals and their associated probability masses. This set of intervals we call an <u>intermediate distribution.</u>*

While a histogram generated by a user corresponds to a set of non-overlapping intervals and their probability masses, the result of an arithmetic operation often produces a set of intervals that contains overlaps, and this set may in turn be used as an operand in further arithmetic operations.

An intermediate distribution represents stepwise CDF's that bound a family of CDF's. Obviously other members of that family of CDF's need not themselves be stepwise. Therefore, other than the fact that a CDF to be provided with containing CDF bounds must have values of zero below a known finite value and one above a known finite value, arbitrary CDF's can be provided with stepwise bounds. The intermediate distribution corresponding to these bounds therefore can represent various CDF's in an

automatically verified fashion, though usually with information loss. Therefore they can also represent PDF's and their histogram based discretizations correctly. Future work is needed to extend the representation to handle distribution functions whose curves are asymptotic.

II. Introduction to Arithmetic Operations

The algorithmic framework for performing an arithmetic operation on operands that are intermediate distributions is exemplified in Figure 4. A fuller description of the algorithms is given in Berleant (1993). Related algorithms appeared beginning in 1968 (Ingram et al.), however these algorithms were not automatically verifying. An approach to using CDF bounds in computer systems administration appears in Post and Diltz (1986) under the rubric of another related field called **stochastic dominance,** which recognizes the usefulness of bounded families of CDF's. We have not been able to find work by others discussing automatically verified operations where one operand is an interval and the other is a distribution function. A fuller review of related work appears in Berleant (1993).

III. A Software Tool

A personal computer based calculating tool for doing automatically verified arithmetic operations on operands each of which may be either an interval or a distribution function has been written. The tool allows the user to specify operands as histograms. These histograms, as well as the results of operations, are represented internally as intermediate distributions. Sets of exact numbers may be read in and are also represented internally as intermediate distributions (whose intervals have equal low and high bounds). All of these may be displayed graphically as CDF families (note that sets of exact numbers correspond to stepwise CDF's and hence can be bounded above and below with the same stepwise CDF). All of these may also be displayed graphically as histograms representing discretized PDF curves which are in most cases approximate, when this is desired by the user for intuition or convenience. Tables of intervals and their associated probabilities, coding intermediate distributions directly, may be printed out on paper.

Addition, subtraction, multiplication, and division are currently implemented. Several of these operations can be done in sequence using a primitive batch command facility.

Our software is available at no charge (for non-commercial purposes) from the authors.

IV. Conclusion

Our previous work (Berleant 1993) described how automatically verified arithmetic operations may be carried out on intervals and distribution functions. The present work reports a computer tool with graphical user interface capabilities for allowing users to specify operands as intervals and distribution functions and allowing them to perform automatically verified arithmetic operations on them. An important next stage in this research is to find applications for the techniques in order to demonstrate practical use of the tool and its underlying ideas.

V. Acknowledgements

References

[1] Berleant, D., Automatically Verified Reasoning with Both Intervals and Probability Density Functions, *Interval Computations* 1993 No. 2, pp. 48–70.

[2] G. E. Ingram, E. L. Welker, and C. R. Herrmann, Designing for Reliability Based on Probabilistic Modeling Using Remote Access Computer Systems, *Proceedings 7th Reliability and Maintainability Conference,* American Society of Mechanical Engineers, 1968, pp. 492–500.

[3] G. V. Post and J. D. Diltz, A Stochastic Dominance Approach to Risk Analysis of Computer Systems, *Management Information Systems Quarterly,* Vol. 10, No. 4, Dec. 1986, pp. 363–375.
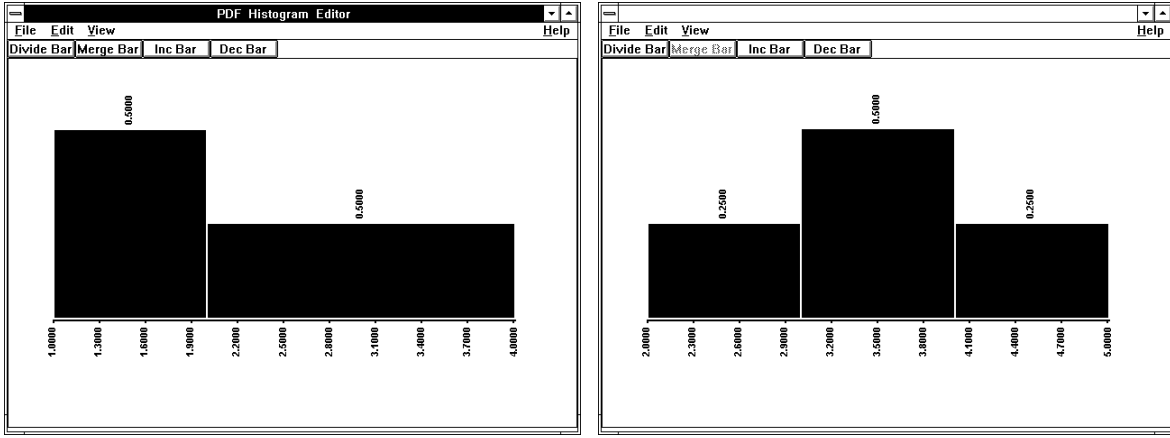
The left histogram X corresponds to the intermediate distribution $\{p([1,2]) = \frac{1}{2}, p([2,4]) = \frac{1}{2}\}$

The right histogram Y corresponds to the intermediate distribution $\{p([2,3]) = \frac{1}{4}, p([3,4]) = \frac{1}{2}, p([4,5]) = \frac{1}{4}\}$

Multiplying their intermediate distributions leads to a Cartesion product:

Cartesian product term #		Operand 1 (X_i)	Operand 2 (Y_j)	Resulting intermediate distribution elements
#1	interval	$[1,2]$	$[2,3]$	$[2,6]$
	probability	$1/2$	$1/4$	$1/8$
#2	interval	$[1,2]$	$[3,4]$	$[3,8]$
	probability	$1/2$	$1/2$	$1/4$
#3	interval	$[1,2]$	$[4,5]$	$[4,10]$
	probability	$1/2$	$1/4$	$1/8$
#4	interval	$[2,4]$	$[2,3]$	$[4,12]$
	probability	$1/2$	$1/4$	$1/8$
#5	interval	$[2,4]$	$[3,4]$	$[6,16]$
	probability	$1/2$	$1/2$	$1/4$
#6	interval	$[2,4]$	$[4,5]$	$[8,20]$
	probability	$1/2$	$1/4$	$1/8$

Integrating the resulting intermediate distribution above produces the curves below ($\downarrow$).
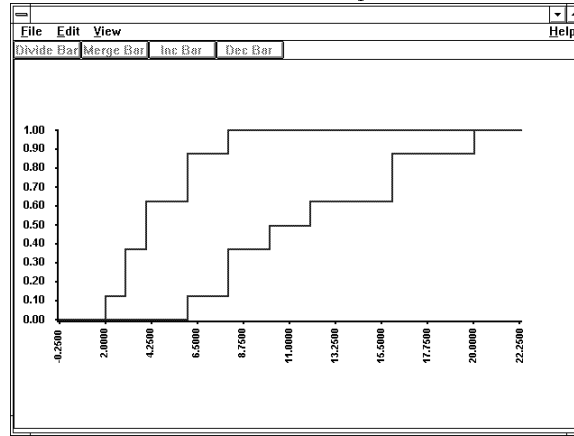

Figure 4: Multiplying two operands. At top, the user creates two histograms X and Y using a graphical editor. The Cartesian product of the intermediate distributions corresponding to these histograms is then used to calculate the members of the intermediate distribution of the answer, whose CDF bounds are shown at bottom.