Reliable Computing 2 (2) (1996), pp. 155–160

Interval approach challenges Monte Carlo simulation

JANNE PESONEN and EERO HYVÖNEN

Intervals are used to represent imprecise numerical values. Modelling uncertain values with precise bounds without considering their probability distribution is infeasible in many applications. As a solution, this paper proposes the use of probability density functions instead of intervals; we consider evaluation of an arithmetical function of random variables. Since the result density cannot in general be solved algebraically, an interval method for determining its *guaranteed* bounds is developed. This possibility challenges traditional Monte Carlo methods in which only stochastic characterizations for the result distribution, such as confidence bounds for fractiles, can be determined.

Интервальный подход составляет конкуренцию моделированию методом Монте-Карло

Я. ПЕСОНЕН, Э. ХИВОНЕН

Интервалы используются для представления неточных численных значений. Однако моделирование неточных величин интервалами с точными границами без учета распределений вероятности этих величин во многих случаях неприемлемо. В качестве одного из решений предложено использовать функции плотности вероятности вместо интервалов. Рассматривается вычисление арифметической функции случайных переменных. Поскольку плотность результата не может быть в общем случае получена алгебраическими методами, предлагается интервальный подход, дающий для этой величины *гарантированные* границы. Такой подход составляет конкуренцию традиционным методам Монте-Карло, которые позволяют определить только стохастические характеристики распределения результата, такие как доверительные границы квантилей.

1. Introduction

An interval is a range of possible values but says nothing about the probability distribution. In many applications such information is essential. For example, the manufacturing tolerance of a resistance R is not an interval but rather a truncated normal distribution. This means that R should be treated as a random variable when used in an arithmetical function $Y = h(\dots, R, \dots)$. The function value Y is then also a probability density function (PDF) $f_Y(y)$ over the feasible y-values. Unfortunately, it is not possible to solve the algebraic form of $f_Y(y)$ in the general case [3]. One usually has to use numerical or approximative techniques.

The most widely used numerical technique for the problem is Monte Carlo simulation (MCS) [4]. Here h is evaluated at different points by stochastic sampling. The resulting sample set is a stepwise approximation of the cumulative density function (CDF) for which stochastic confidence bounds can be calculated.

Also interval techniques have been applied to the problem. Berleant [1] discretizes independent argument densities into histograms representing probability mass distributions, after which stepwise guaranteed bounds for the CDF can be determined. However, since this approach transforms argument densities into probability masses, information concerning the density is lost. Williamson [6] derives stepwise bounded CDFs for binary operations with another method, but the resulting density is not considered.

There are good reasons for determining bounds for the actual continuous density function instead of calculating the approximative or bounded CDF. Firstly, bounded PDF is more fundamental than bounded or approximated CDF: bounds for the CDF can be constructed from the bounded density function, but bounded PDF cannot be formed from the CDF-approximation or from the stepwise CDF-bounds. Secondly, from the user's viewpoint, the form of the density function is more informative than the form of the CDF in many situations, like when displaying relative probability values, the most likely values (modes), the actual shape of the distribution (skewness, kurtosis, etc.) and small changes in probability density [4]. In this paper we present an interval method that calculates guaranteed bounds for the continuous PDF.

2. Calculation of density bounds

Consider a function $Y = h(X_1, X_2, \ldots, X_n)$ of independent random variables X_i, each with a known PDF $f_{X_i}(x_i)$. We want to know $f_Y(y)$, the density of Y. If at least one variable, say X_1, can be solved in terms of Y and other X's, then under certain assumptions (cf. [3])[1] $f_Y(y)$ can be represented by the integral form

$$f_Y(y) = \int \cdots \int g(y, x_2, \ldots, x_n) f_{X_2}(x_2) \ldots f_{X_n}(x_n)\, dx_2 \ldots dx_n,$$
$$g(y, x_2, \ldots, x_n) = f_{X_1}(g_{X_1}(y, x_2, \ldots, x_n), x_2, \ldots, x_n) g_J(y, x_2, \ldots, x_n) \quad (1)$$

where g_{X_1} is x_1 solved out from $Y = h(X_1, X_2, \ldots, X_n)$ and g_J is the Jacobian term

$$g_J(y, x_2, \ldots, x_n) = \left| \frac{\partial g_{X_1}(y, x_2, \ldots, x_n)}{\partial y} \right|. \quad (2)$$

If a closed form solution for (1) cannot be determined, then it is possible to consider $\underline{\overline{f}}_Y(\underline{\overline{y}})$ in subboxes by partitioning $\underline{\overline{y}}$ and the argument distribution support intervals $\underline{\overline{x}}_2, \ldots, \underline{\overline{x}}_n$ exhaustively and exclusively into sets $\{\underline{\overline{y}}^r\}, \{\underline{\overline{x}}_2^{s_2}\}, \ldots, \{\underline{\overline{x}}_n^{s_n}\}$. Under the assumptions of (1) the following interval formula can then be used for bounding $f_Y(y)$:

$$\underline{\overline{f}}_Y(\underline{\overline{y}}^r) = \sum_S \underline{\overline{g}}(\underline{\overline{y}}^r, \underline{\overline{x}}_2^{s_2}, \ldots, \underline{\overline{x}}_n^{s_n}) F_2(\underline{\overline{x}}_2^{s_2}) \ldots F_n(\underline{\overline{x}}_n^{s_n}). \quad (3)$$

Here the sum is taken over the Cartesian index set $S = \{(s_2, \ldots, s_n)\}$, $F_i(\underline{\overline{x}}_i^{s_i})$ is the probability mass of f_{X_i} within interval $\underline{\overline{x}}_i^{s_i}$, and $\underline{\overline{g}}$ is an interval extension of g, i.e., its value is $[\underline{g}, \overline{g}]$ where $\underline{g} \leq \inf g(y, x_2, \ldots, x_n)$ and $\overline{g} \geq \sup g(y, x_2, \ldots, x_n)$, $y \in \underline{\overline{y}}^r, x_2 \in \underline{\overline{x}}_2^{s_2}, \ldots, x_n \in \underline{\overline{x}}_n^{s_n}$. It is next shown that (3) bounds the density of Y.

Proposition 1. *Formula (3) provides bounds for (1), i.e.,* $f_Y(y) \in \underline{\overline{f}}_Y(\underline{\overline{y}}^r)$, *if* $y \in \underline{\overline{y}}^r$.

Proof. Formula (1) can be expressed as

[1] $\frac{\partial y}{\partial x_1}$ is continuous and $\neq 0$ at all $(x_1, \ldots, x_n)$ and $f_{X_1}(x_1) \ldots f_{X_n}(x_n)$ is continuous at all but a finite number of points $x = (x_1, \ldots, x_n)$.

$$f_Y(y) = \sum_S \int_{\underline{x}_2^s}^{\overline{x}_2^s} \cdots \int_{\underline{x}_n^s}^{\overline{x}_n^s} g(y, x_2, \ldots, x_n) f_{X_2}(x_2) \ldots f_{X_n}(x_n)\, dx_2 \ldots dx_n \tag{4}$$

where each sum term has a corresponding term in (3). Thus it is sufficient to prove that each sum term in (4) is contained in the corresponding sum term in (3), i.e.,

$$\int_{\underline{x}_2^s}^{\overline{x}_2^s} \cdots \int_{\underline{x}_n^s}^{\overline{x}_n^s} g(y, x_2, \ldots, x_n) f_{X_2}(x_2) \ldots f_{X_n}(x_n)\, dx_2 \ldots dx_n \in \overline{\underline{g}}(\overline{\underline{y}}^r, \overline{\underline{x}}_2^{s_2}, \ldots, \overline{\underline{x}}_n^{s_n}) F_2(\overline{\underline{x}}_2^{s_2}) \ldots F_n(\overline{\underline{x}}_n^{s_n}) \tag{5}$$

For the left hand side (lhs) of (5) we can write inequalities

$$\inf_s \left(g(y, x_2, \ldots, x_n)\right) \int_{\underline{x}_2^s}^{\overline{x}_2^s} \cdots \int_{\underline{x}_n^s}^{\overline{x}_n^s} f_{X_2}(x_2) \ldots f_{X_n}(x_n)\, dx_2 \ldots dx_n \leq lhs \leq \sup_s \left(g(y, x_2, \ldots, x_n)\right) \int_{\underline{x}_2^s}^{\overline{x}_2^s} \cdots \int_{\underline{x}_n^s}^{\overline{x}_n^s} f_{X_2}(x_2) \ldots f_{X_n}(x_n)\, dx_2 \ldots dx_n \tag{6}$$

from which it is evident that

$$\underline{g}(\overline{\underline{y}}^r, \overline{\underline{x}}_2^{s_2}, \ldots, \overline{\underline{x}}_n^{s_n}) F_2(\overline{\underline{x}}_2^{s_2}) \ldots F_n(\overline{\underline{x}}_n^{s_n}) \leq lhs \leq \overline{g}(\overline{\underline{y}}^r, \overline{\underline{x}}_2^{s_2}, \ldots, \overline{\underline{x}}_n^{s_n}) F_2(\overline{\underline{x}}_2^{s_2}) \ldots F_n(\overline{\underline{x}}_n^{s_n}) \qquad \square$$

Using (3) provides thus safe outer bounds for the actual density within the given subinterval $\overline{\underline{y}}^r$. The finer partitions one makes, the narrower densities one gets. Formula (3) approaches the analytic integral form (1), when individual subinterval widths in all partitions approach zero.

In order to bound the whole density curve $f_Y(y)$ one has to calculate PDF bounds for all support subintervals in the partition of y. The result is step functions for lower and upper density bounds. The area between bounds defines a family of PDFs all of which must have quadrature one. Based on these PDF-bounds piecewise linear bounds for the CDF can also be computed.

The derivative $f'_Y(y)$ can be bounded in a similar way if the derivative of the solved variable does not contain singularities. This makes it possible to further constrain the uncertain area between bounds, because the density can now be bounded by linear splines with varying slopes instead of flat steps only. Then CDF bounds are piecewise defined curves of second degree.

The complexity of (3) is exponential w.r.t. the number of variables. If we have n variables each with a partition of size s and the result is wanted in m subintervals, then calculating bounds for the density curve requires the integrand to be evaluated in $ms^{(n-1)}$ subboxes. This complexity can be reduced if the original function h can be decomposed into algebraically independent parts. Such parts can be calculated separately and the intermediate bounded PDFs can then be used in the subsequent operations. For example, if h is the sum of 8 variables each of which is partitioned into 100 subintervals and result is wanted in 100 parts, then 10^{16} subboxes are processed with (3). But if we do the 7 additions one after another with 100 subintervals in the result at each step, then only 7×10^4 boxes are needed. Our experiments seem to indicate that with pairwise evaluations narrower bounds are obtained even with smaller amount of subbox evaluations.

The sharpness of the bounds obtained depends not only on the subboxes used but also on how the integrand bounds in each subbox are evaluated. Local interval arithmetic that neglects dependences between multiple occurrences of a variable is fast but gives overestimations that in some cases can be very large. Then, global evaluation of the integrand is needed.

3. An example

Assume a resistance R and voltage U given as the triangular distributions $R = \text{Triangle}(1, 2, 3)$ and $U = \text{Triangle}(2, 3, 4)$, where in $\text{Triangle}(a, b, c)$ a is the left bound, b the most probable value and c the right bound. The task is to compute the support and PDF of the current $I = U/R$. The result support is $\underline{\overline{i}} = \underline{\overline{u}}/\underline{\overline{r}} = [2/3, 4]$. Suppose that we want the PDF bounds in 30 equally wide subintervals of I-support. U is selected to be the solved variable and R-support is partitioned into 100 equally wide subintervals. Then the integral form (1) is

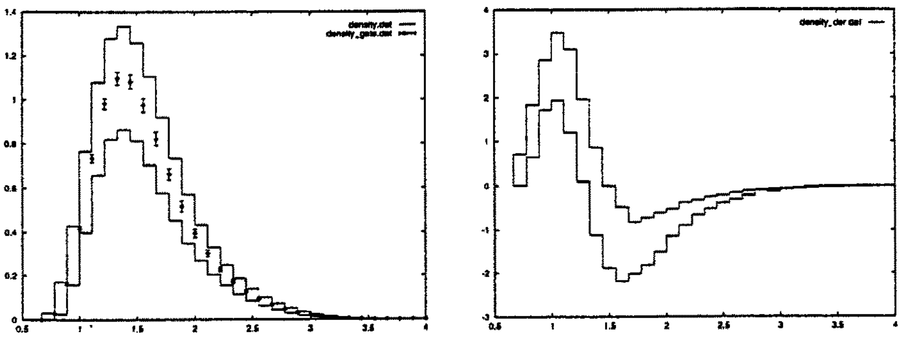
$$f_I(i) = \int_{\underline{r}}^{\overline{r}} f_U\big(g_U(i,r)\big) g'_U(i) f_R(r) dr = \int_{\underline{r}}^{\overline{r}} f_U(ir) r f_R(r) dr \tag{7}$$

Now we can use (3) to calculate bounds for both PDF and the derivative of PDF. Integrand bounds are calculated here with local interval arithmetic, which gives us unnecessary wide overestimates due to neglecting the dependence between two occurrences of variable r.

Figure 1 shows results of such an evaluation: (a) depicts stepwise density bounds for intervals and short vertical segments show density bounds at partition boundary points; (b) shows derivative bounds. By using also derivative information, the smoother piecewise linear PDF-bounds shown in Figure 2 are obtained.

The bounding error between lower and upper bounds can be reduced by making partitions of i and r finer. In the example, partitions of equally wide subintervals were used, but a better strategy is to do partitioning dynamically in order to distribute the error evenly. For instance, Figure 2 shows that finer partitioning around the peak is needed than in the right tail, if the same relative accuracy is wanted.

Another way to increase accuracy is to use interval analytic techniques such as [5] for determining stricter bounds for the interval integrand expression in the subboxes. In some cases this really is obligatory. Consider for example a situation, where R and U are lognormal distributions with the support maximums at infinity. Let's say that the rightmost r-part is $\underline{\overline{r}}^s = [a, \infty]$ and that we are after the density bound at the point $i = 1$. One integrand term in the sum is the monotonicly decreasing product $f_U(i\underline{\overline{r}}^s)\underline{\overline{r}}^s = f_U([a, \infty])[a, \infty]$. For this local IA gives $[0, f_U(a)] \times [a, \infty] = [0, \infty]$, while the true value is $[0, f_U(a)a]$. This means that without



(a) Density bounds

(b) Derivative bounds

Figure 1. Bounds for 30 subintervals of i with 100 iterations over r

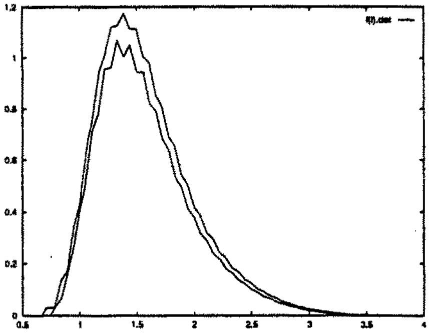


Figure 2. Bounded probability density

considering the two occurrences of r the upper density bound is always overestimated to be plus infinity.

4. Conclusion

Interval arithmetic can be used for bounding the result distribution of an arithmetical function of random variables. This approach challenges Monte Carlo simulation (MCS), where only stochastic bounds can be derived.

The fundamental qualitative advantage of the interval approach is that guaranteed bounds for both PDF and CDF can be determined. MCS uses pointwise evaluations and can only characterize. the accuracy of the resulting CDF statistically, e.g., determine stochastic confidence bounds for the mean or fractiles. From the sampled set one can try to visualize the PDF by smoothing and derivating the stepwise CDF and then by adjusting the PDF-curve to have quadrature one. This may produce a curve close to reality or something totally wrong. Another method to perceive the distribution form is to construct a histogram from the sample. Two histograms from the same sample can however look very different if the histogram bins are different. Neither of these methods produces reliable results, because knowledge of the probability density is lost in simulation.

To the user the actual form of the PDF (or its bounds) is often more informative than the CDF approximation from MCS. Consider, for example, risk analysis. In the interval approach any risk peaks—that may have very small probability—can be directly detected, while in MCS such points may be lost in the noise or are difficult to find at least to an unexperienced user. An inherent property of risks and exceptional events, for which MCS is widely used, is their low probability. These phenomena therefore by their very nature escape standard MCS analysis.

In addition, the interval approach also challenges basic MCS on its home ground in fractile confidence analysis. Consider the given circuit example again. Suppose, that we want to determine an upper limit for the current such that exceeding it is at most 10^{-5} probable, and that one wants to be 0.95 confident of the result. To achieve this degree of certainty the required number of sample points is $\approx 4 \times 10^5$ [4]. To be 0.9996 certain requires $\approx 1.2 \times 10^6$ samples. Using the bounded PDF (Figure 2) derived by our interval method one can say that the upper limit is *guaranteed* to be greater than 3.7355 and smaller than 3.8387. This result was achieved by evaluating only 3×10^3 subboxes with simple local IA. Interval approach seems promising in regions with low probability mass, although there are advanced MCS techniques for better management of tail probabilities (such as importance sampling [2]).

If only a tail fractile is of concern, then interval computations can be focused to the tail only—there is no need to construct the whole distribution. More generally, any result distribution part can be calculated independently of other parts. This enables, for example, directing computational effort towards distribution parts having largest uncertainty between the density bounds.

A nice feature of MCS is that its time complexity is linear w.r.t. the number of function arguments [4], while for the interval approach this is a major problem. This suggests that the method is feasible only for small problems. There are, however, means to reduce the complexity. For example, a large function can be computed in smaller independent parts. Another problem is that, in contrast to MCS, the interval approach is very sensitive to the algebraic form of the function.

In this paper, stochastic independence of argument variables was assumed; their joint distribution was the product of their densities (1). However, the product can be replaced by any other joint density function [3]. Density bounds can still be calculated, but the procedure is somewhat different from the one presented here.

The fundamental question when comparing MCS with the interval approach is how high precision is actually enough? In the interval approach all events, even those which have very low probability, can be captured within bounds. We believe that this possibility alone justifies further research on the interval approach.

Acknowledgements

This research was mainly funded by the Technology Development Centre of Finland (TEKES) and VTT. Cooperation with Trema Inc., Tekla Inc. and Mintax Inc. is acknowledged. Thanks to Stefano De Pascale and Ilkka Karanta for discussions.

References

[1] Berleant, D. *Automatically verified reasoning with both intervals and probability density functions.* Interval Computations 2 (1993), pp. 48–70.

[2] Clark, C. E. *Importance sampling in Monte Carlo analysis.* Operations Research 9 (1961), pp. 603–620.

[3] Dudewicz, E. J. and Mishra, S. N. *Modern mathematical statistics.* John Wiley & Sons, 1988.

[4] Morgan, M. and Henrion, M. *Uncertainty.* Cambridge University Press, 1990.

[5] Ratschek, H. and Rokne, J. *Computer methods for the range of functions.* Ellis Horwood Limited, Chichester, England, 1984.

[6] Williamson, R. C. and Downs, T. *Probabilistic arithmetic. i. numerical methods for calculating convolutions and dependency bounds.* International Journal of Approximate Reasoning 4 (2) (1990), pp. 88–158.

Received: November 1, 1995
Revised version: November 30, 1995

VTT Information Technology
P.O. Box 1201, 02044, Finland
E-mail: {Janne.Pesonen, Eero.Hyvonen}@vtt.fi