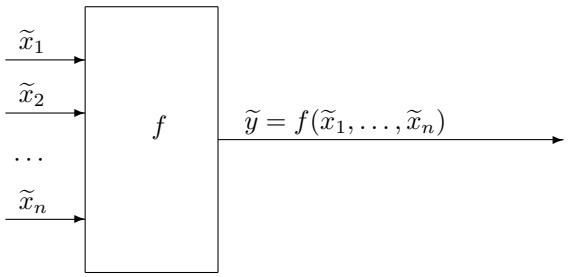
Statistical Data Processing under Interval Uncertainty: Algorithms and Computational Complexity

Vladik Kreinovich

Department of Computer Science, University of Texas at El Paso,
El Paso, TX 79968, USA vladik@utep.edu

1 Main Problem

Why indirect measurements? In many real-life situations, we are interested in the value of a physical quantity y that is difficult or impossible to measure directly. Examples of such quantities are the distance to a star and the amount of oil in a given well. Since we cannot measure y directly, a natural idea is to measure y *indirectly*. Specifically, we find some easier-to-measure quantities $x_1, \ldots, x_n$ which are related to y by a known relation $y = f(x_1, \ldots, x_n)$; this relation may be a simple functional transformation, or complex algorithm (e.g., for the amount of oil, numerical solution to an inverse problem). Then, to estimate y, we first measure the values of the quantities $x_1, \ldots, x_n$, and then we use the results $\widetilde{x}_1, \ldots, \widetilde{x}_n$ of these measurements to to compute an estimate $\widetilde{y}$ for y as $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$:



For example, to find the resistance R, we measure current I and voltage V, and then use the known relation $R = V/I$ to estimate resistance as $\widetilde{R} = \widetilde{V}/\widetilde{I}$.

Computing an estimate for y based on the results of direct measurements is called *data processing*; data processing is the main reason why computers were invented in the first place, and data processing is still one of the main uses of computers as number crunching devices.

Comment. In this paper, for simplicity, we consider the case when the relation between x_i and y is known exactly; in some practical situations, we only known an approximate relation between x_i and y.

Why interval computations? From computing to probabilities to intervals. Measurement are never 100% accurate, so in reality, the actual value x_i of i-th measured quantity can differ from the measurement result $\widetilde{x}_i$. Because of these *measurement errors* $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$, the result $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ of data processing is, in general, different from the actual value $y = f(x_1, \ldots, x_n)$ of the desired quantity y.

It is desirable to describe the error $\Delta y \stackrel{\text{def}}{=} \widetilde{y} - y$ of the result of data processing. To do that, we must have some information about the errors of direct measurements.

What do we know about the errors Δx_i of direct measurements? First, the manufacturer of the measuring instrument must supply us with an upper bound Δ_i on the measurement error. If no such upper bound is supplied, this means that no accuracy is guaranteed, and the corresponding "measuring instrument" is practically useless. In this case, once we performed a measurement and got a measurement result $\widetilde{x}_i$, we know that the actual (unknown) value x_i of the measured quantity belongs to the interval $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$, where $\underline{x}_i = \widetilde{x}_i - \Delta_i$ and $\overline{x}_i = \widetilde{x}_i + \Delta_i$.

In many practical situations, we not only know the interval $[-\Delta_i, \Delta_i]$ of possible values of the measurement error; we also know the probability of different values Δx_i within this interval. This knowledge underlies the traditional engineering approach to estimating the error of indirect measurement, in which we assume that we know the probability distributions for measurement errors Δx_i.

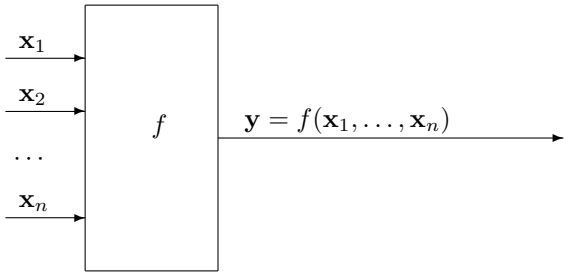
In practice, we can determine the desired probabilities of different values of Δx_i by comparing the results of measuring with this instrument with the results of measuring the same quantity by a standard (much more accurate) measuring instrument. Since the standard measuring instrument is much more accurate than the one use, the difference between these two measurement results is practically equal to the measurement error; thus, the empirical distribution of this difference is close to the desired probability distribution for measurement error. There are two cases, however, when this determination is not done:

- First is the case of cutting-edge measurements, e.g., measurements in fundamental science. When a Hubble telescope detects the light from a distant galaxy, there is no "standard" (much more accurate) telescope floating nearby that we can use to calibrate the Hubble: the Hubble telescope is the best we have.
- The second case is the case of measurements on the shop floor. In this case, in principle, every sensor can be thoroughly calibrated, but sensor calibration is so costly – usually costing ten times more than the sensor itself – that manufacturers rarely do it.

In both cases, we have no information about the probabilities of Δx_i; the only information we have is the upper bound on the measurement error.

In this case, after we performed a measurement and got a measurement result $\widetilde{x}_i$, the only information that we have about the actual value x_i of the measured quantity is that it belongs to the interval $\mathbf{x}_i = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$. In such situations, the only information that we have about the (unknown) actual value of $y = f(x_1, \ldots, x_n)$ is that y belongs to the range $\mathbf{y} = [\underline{y}, \overline{y}]$ of the function f over the box $\mathbf{x}_1 \times \ldots \times \mathbf{x}_n$:

$$\mathbf{y} = [\underline{y}, \overline{y}] = \{f(x_1, \ldots, x_n) \,|\, x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\}.$$



The process of computing this interval range based on the input intervals $\mathbf{x}_i$ is called *interval computations*; see, e.g., [19, 37].

Interval computations techniques: brief reminder. Historically the first method for computing the enclosure for the range is the method which is sometimes called "straightforward" interval computations. This method is based on the fact that inside the computer, every algorithm consists of elementary operations (arithmetic operations, min, max, etc.). For each elementary operation $f(a, b)$, if we know the intervals $\mathbf{a}$ and $\mathbf{b}$ for a and b, we can compute the exact range $f(\mathbf{a}, \mathbf{b})$. The corresponding formulas form the so-called *interval arithmetic*. For example,

$$[\underline{a}, \overline{a}] + [\underline{b}, \overline{b}] = [\underline{a} + \underline{b}, \overline{a} + \overline{b}]; \quad [\underline{a}, \overline{a}] - [\underline{b}, \overline{b}] = [\underline{a} - \overline{b}, \overline{a} - \underline{b}];$$

$$[\underline{a}, \overline{a}] \cdot [\underline{b}, \overline{b}] = [\min(\underline{a} \cdot \underline{b}, \underline{a} \cdot \overline{b}, \overline{a} \cdot \underline{b}, \overline{a} \cdot \overline{b}), \max(\underline{a} \cdot \underline{b}, \underline{a} \cdot \overline{b}, \overline{a} \cdot \underline{b}, \overline{a} \cdot \overline{b})].$$

In straightforward interval computations, we repeat the computations forming the program f step-by-step, replacing each operation with real numbers by the corresponding operation of interval arithmetic. It is known that, as a result, we get an enclosure $\mathbf{Y} \supseteq \mathbf{y}$ for the desired range.

In some cases, this enclosure is exact. In more complex cases (see examples below), the enclosure has excess width.

Example. Let us illustrate the above idea on the example of estimating the range of the function $f(x) = (x - 2) \cdot (x + 2)$ on the interval $x \in [1, 2]$.

We start with parsing the expression for the function, i.e., describing how a computer will compute this expression; it will implement the following sequence of elementary operation:

$$r_1 := x - 2; \quad r_2 := x + 2; \quad r_3 := r_1 \cdot r_2.$$

The main idea behind straightforward interval computations is to perform the same operations, but with *intervals* instead of *numbers*:

$$\mathbf{r}_1 := [1,2] - [2,2] = [-1,0]; \quad \mathbf{r}_2 := [1,2] + [2,2] = [3,4];$$

$$\mathbf{r}_3 := [-1,0] \cdot [3,4] = [-4,0].$$

For this function, the actual range is $f(\mathbf{x}) = [-3,0]$.

Comment: this is just a toy example, there are more efficient ways of computing an enclosure $\mathbf{Y} \supseteq \mathbf{y}$.

There exist more sophisticated techniques for producing a narrower enclosure, e.g., a centered form method. However, for each of these techniques, there are cases when we get an excess width. Reason: as shown in [25], the problem of computing the exact range is known to be NP-hard even for polynomial functions $f(x_1, \ldots, x_n)$ (actually, even for quadratic functions f).

Practical problem. In some practical situations, in addition to the lower and upper bounds on each random variable x_i, we have some additional information about x_i.

So, we arrive at the following problem:

- we have a data processing algorithm $f(x_1, \ldots, x_n)$, and
- we have some information about the uncertainty with which we know x_i (e.g., measurement errors).

We want to know the resulting uncertainty in the result $y = f(x_1, \ldots, x_n)$ of data processing.

In interval computations, we assume that the uncertainty in x_i can be described by the interval of possible values. In real life, in addition to the intervals, we often have some information about the probabilities of different values within this interval. What can we then do?

2 What is the Best Way to Describe Probabilistic Uncertainty?

In order to describe how uncertainty in x_i affects y, we need to know what is the best way to represent the corresponding probabilistic uncertainty in x_i.

In probability theory, there are many different ways of representing a probability distribution. For example, one can use a probability density function (pdf), or a cumulative distribution function (CDF), or a probability measure, i.e., a function which maps different sets into a probability that the corresponding random variable belongs to this set. The reason why there are many different representations is that in different problems, different representations turned out to be the most useful.

We would like to select a representation which is the most useful for problems related to risk analysis. To make this selection, we must recall where the information about probabilities provided by risk analysis is normally used.

How is the partial information about probabilities used in risk analysis? The main objective of risk analysis is to make decisions. A standard way of making a decision is to select the action a for which the expected utility (gain) is the largest possible. This is where probabilities are used: in computing, for every possible action a, the corresponding expected utility. To be more precise, we usually know, for each action a and for each actual value of the (unknown) quantity x, the corresponding value of the utility $u_a(x)$. We must use the probability distribution for x to compute the expected value $E[u_a(x)]$ of this utility.

In view of this application, the most useful characteristics of a probability distribution would be the ones which would enable us to compute the expected value $E[u_a(x)]$ of different functions $u_a(x)$.

Which representations are the most useful for this intended usage? General idea. Which characteristics of a probability distribution are the most useful for computing mathematical expectations of different functions $u_a(x)$? The answer to this question depends on the type of the function, i.e., on how the utility value u depends on the value x of the analyzed parameter.

Smooth utility functions naturally lead to moments. One natural case is when the utility function $u_a(x)$ is smooth. We have already mentioned, in Section I, that we usually know a (reasonably narrow) interval of possible values of x. So, to compute the expected value of $u_a(x)$, all we need to know is how the function $u_a(x)$ behaves on this narrow interval. Because the function is smooth, we can expand it into Taylor series. Because the interval is narrow, we can safely consider only linear and quadratic terms in this expansion and ignore higher-order terms: $u_a(x) \approx c_0 + c_1 \cdot (x - x_0) + c_2 \cdot (x - x_0)^2$, where x_0 is a point inside the interval. Thus, we can approximate the expectation of this function by the expectation of the corresponding quadratic expression: $E[u_a(x)] \approx E[c_0 + c_1 \cdot (x - x_0) + c_2 \cdot (x - x_0)^2]$, i.e., by the following expression: $E[u_a(x)] \approx c_0 + c_1 \cdot E[x - x_0] + c_2 \cdot E[(x - x_0)^2]$. So, to compute the expectations of such utility functions, it is sufficient to know the first and second moments of the probability distribution.

In particular, if we use, as the point x_0, the average $E[x]$, the second moment turns into the variance of the original probability distribution. So, instead of the first and the second moments, we can use the mean E and the variance V.

In risk analysis, non-smooth utility functions are common. In engineering applications, most functions are smooth, so usually the Taylor expansion works pretty well. In risk analysis, however, not all dependencies are smooth. There is often a threshold x_0 after which, say, a concentration of a certain chemical becomes dangerous.

This threshold sometimes comes from the detailed chemical and/or physical analysis. In this case, when we increase the value of this parameter, we see the drastic increase in effect and hence, the drastic change in utility value. Sometimes, this threshold simply comes from regulations. In this case, when

we increase the value of this parameter past the threshold, there is no drastic increase in effects, but there is a drastic decrease of utility due to the necessity to pay fines, change technology, etc. In both cases, we have a utility function which experiences an abrupt decrease at a certain threshold value x_0.

Non-smooth utility functions naturally lead to CDFs. We want to be able to compute the expected value $E[u_a(x)]$ of a function $u_a(x)$ which changes smoothly until a certain value x_0, then drops it value and continues smoothly for $x > x_0$. We usually know the (reasonably narrow) interval which contains all possible values of x. Because the interval is narrow and the dependence before and after the threshold is smooth, the resulting change in $u_a(x)$ before x_0 and after x_0 is much smaller than the change at x_0. Thus, with a reasonable accuracy, we can ignore the small changes before and after x_0, and assume that the function $u_a(x)$ is equal to a constant u^+ for $x < x_0$, and to some other constant $u^- < u^+$ for $x > x_0$.

The simplest case is when $u^+ = 1$ and $u^- = 0$. In this case, the desired expected value $E[u_a^{(0)}(x)]$ coincides with the probability that $x < x_0$, i.e., with the corresponding value $F(x_0)$ of the cumulative distribution function (CDF). A generic function $u_a(x)$ of this type, with arbitrary values u^- and u^+, can be easily reduced to this simplest case, because, as one can easily check, $u_a(x) = u^- + (u^+ - u^-) \cdot u^{(0)}(x)$ and hence, $E[u_a(x)] = u^- + (u^+ - u^-) \cdot F(x_0)$.

Thus, to be able to easily compute the expected values of all possible non-smooth utility functions, it is sufficient to know the values of the CDF $F(x_0)$ for all possible x_0.

3 How to Represent Partial Information about Probabilities

General idea. In many cases, we have a complete information about the probability distributions that describe the uncertainty of each of n inputs.

However, a practically interesting case is how to deal with situations when we only have partial information about the probability distributions. How can we represent this partial information?

Case of cdf. If we use cdf $F(x)$ to represent a distribution, then full information corresponds to the case when we know the exact value of $F(x)$ for every x. Partial information means:

- either that we only know approximate values of $F(x)$ for all x, i.e., that for every x, we only know the interval that contains $F(x)$; in this case, we get a *p-box*;
- or that we only know the values of $F(x)$ for some x, i.e, that we only know the values $F(x_1)$, ..., $F(x_n)$ for finitely many values $x = x_1, \ldots, x_n$; in this case, we have a *histogram*.

It is also possible that we know only approximate values of $F(x)$ for some x; in this case, we have an *interval-valued histogram*.

Case of moments. If we use moments to represent a distribution, then partial information means that we either know the exact values of finitely many moments, or that we know intervals of possible values of several moments.

4 Resulting Problems

This discussion leads to a natural classification of possible problems:

- If we have complete information about the distributions of x_i, then, to get validated estimates on uncertainty of y, we have to use Monte-Carlo-type techniques; see, in particular, papers by D. Lodwick et al. [33, 34]
- If we have p-boxes, we can use methods proposed by S. Ferson et al. [13, 14, 15, 23, 43, 46].
- If we have histograms, we can use methods proposed by D. Berleant et al. [6, 7, 8, 9, 10, 44, 53].
- If we have moments, then we can use methods proposed by S. Ferson, V. Kreinovich, M. Orshansky, et al. [18, 22, 30, 41, 42].

There are also additional issues, including:

- how we get these bounds for x_i?
- specific practical applications, like the appearance of histogram-type distributions in problems related to privacy in statistical databases,
- etc.

5 Case Study

Practical problem. In some practical situations, in addition to the lower and upper bounds on each random variable x_i, we know the bounds $\mathbf{E}_i = [\underline{E}_i, \overline{E}_i]$ on its mean E_i.

Indeed, in measurement practice (see, e.g., [11]), the overall measurement error Δx is usually represented as a sum of two components:

- a *systematic* error component $\Delta_s x$ which is defined as the expected value $E[\Delta x]$, and
- a *random* error component $\Delta_r x$ which is defined as the difference between the overall measurement error and the systematic error component: $\Delta_r x \stackrel{\text{def}}{=} \Delta x - \Delta_s x$.

In addition to the bound Δ on the overall measurement error, the manufacturers of the measuring instrument often provide an upper bound Δ_s on the systematic error component: $|\Delta_s x| \le \Delta_s$.

This additional information is provided because, with this additional information, we not only get a bound on the accuracy of a single measurement, but we also get an idea of what accuracy we can attain if we use

repeated measurements to increase the measurement accuracy. Indeed, the very idea that repeated measurements can improve the measurement accuracy is natural: we measure the same quantity by using the same measurement instrument several (N) times, and then take, e.g., an arithmetic average $\bar{x} = \dfrac{\widetilde{x}^{(1)} + \ldots + \widetilde{x}^{(N)}}{N}$ of the corresponding measurement results $\widetilde{x}^{(1)} = x + \Delta x^{(1)}, \ldots, \widetilde{x}^{(N)} = x + \Delta x^{(N)}$.

- If systematic error is the only error component, then all the measurements lead to exactly the same value $\widetilde{x}^{(1)} = \ldots = \widetilde{x}^{(N)}$, and averaging does not change the value – hence does not improve the accuracy.
- On the other hand, if we know that the systematic error component is 0, i.e., $E[\Delta x] = 0$ and $E[\widetilde{x}] = x$, then, as $N \to \infty$, the arithmetic average tends to the actual value x. In this case, by repeating the measurements sufficiently many times, we can determine the actual value of x with an arbitrary given accuracy.

In general, by repeating measurements sufficiently many times, we can arbitrarily decrease the random error component and thus attain accuracy as close to Δ_s as we want.

When this additional information is given, then, after we performed a measurement and got a measurement result $\widetilde{x}$, then not only we get the information that the actual value x of the measured quantity belongs to the interval $\mathbf{x} = [\widetilde{x} - \Delta, \widetilde{x} + \Delta]$, but we can also conclude that the expected value of $x = \widetilde{x} - \Delta x$ (which is equal to $E[x] = \widetilde{x} - E[\Delta x] = \widetilde{x} - \Delta_s x$) belongs to the interval $\mathbf{E} = [\widetilde{x} - \Delta_s, \widetilde{x} + \Delta_s]$.

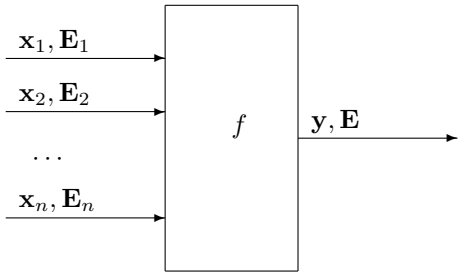
If we have this information for every x_i, then, in addition to the interval $\mathbf{y}$ of possible value of y, we would also like to know the interval of possible values of $E[y]$. This additional interval will hopefully provide us with the information on how repeated measurements can improve the accuracy of this indirect measurement. Thus, we arrive at the following problem:

Precise formulation of the problem. Given an algorithm computing a function $f(x_1, \ldots, x_n)$ from R^n to R, and values $\underline{x}_1$, $\overline{x}_1$, …, $\underline{x}_n$, $\overline{x}_n$, $\underline{E}_1$, $\overline{E}_1$, …, $\underline{E}_n$, $\overline{E}_n$, we want to find

$$\underline{E} \stackrel{\text{def}}{=} \min\{E[f(x_1, \ldots, x_n)] \mid \text{ all distributions of } (x_1, \ldots, x_n) \text{ for which}$$

$$x_1 \in [\underline{x}_1, \overline{x}_1], \ldots, x_n \in [\underline{x}_n, \overline{x}_n], E[x_1] \in [\underline{E}_1, \overline{E}_1], \ldots E[x_n] \in [\underline{E}_n, \overline{E}_n]\};$$

and $\overline{E}$ which is the maximum of $E[f(x_1, \ldots, x_n)]$ for all such distributions.



In addition to considering all possible distributions, we can also consider the case when all the variables x_i are independent.

How we solve this problem. The main idea behind straightforward interval computations can be applied here as well. Namely, first, we find out how to solve this problem for the case when $n = 2$ and $f(x_1, x_2)$ is one of the standard arithmetic operations. Then, once we have an arbitrary algorithm $f(x_1, \ldots, x_n)$, we parse it and replace each elementary operation on real numbers with the corresponding operation on quadruples $(\underline{x}, \underline{E}, \overline{E}, \overline{x})$.

To implement this idea, we must therefore know how to, solve the above problem for elementary operations.

For *addition*, the answer is simple. Since $E[x_1 + x_2] = E[x_1] + E[x_2]$, if $y = x_1 + x_2$, there is only one possible value for $E = E[y]$: the value $E = E_1 + E_2$. This value does not depend on whether we have correlation or nor, and whether we have any information about the correlation. Thus, $\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2$.

Similarly, the answer is simple for *subtraction:* if $y = x_1 - x_2$, there is only one possible value for $E = E[y]$: the value $E = E_1 - E_2$. Thus, $\mathbf{E} = \mathbf{E}_1 - \mathbf{E}_2$.

For *multiplication*, if the variables x_1 and x_2 are independent, then $E[x_1 \cdot x_2] = E[x_1] \cdot E[x_2]$. Hence, if $y = x_1 \cdot x_2$ and x_1 and x_2 are independent, there is only one possible value for $E = E[y]$: the value $E = E_1 \cdot E_2$; hence $\mathbf{E} = \mathbf{E}_1 \cdot \mathbf{E}_2$.

The first non-trivial case is the case of multiplication in the presence of possible correlation. When we know the exact values of E_1 and E_2, the solution to the above problem is as follows:

Theorem 1. *For multiplication $y = x_1 \cdot x_2$, when we have no information about the correlation,*

$$\underline{E} = \max(p_1 + p_2 - 1, 0) \cdot \overline{x}_1 \cdot \overline{x}_2 + \min(p_1, 1 - p_2) \cdot \overline{x}_1 \cdot \underline{x}_2 +$$
$$\min(1 - p_1, p_2) \cdot \underline{x}_1 \cdot \overline{x}_2 + \max(1 - p_1 - p_2, 0) \cdot \underline{x}_1 \cdot \underline{x}_2;$$

and

$$\overline{E} = \min(p_1, p_2) \cdot \overline{x}_1 \cdot \overline{x}_2 + \max(p_1 - p_2, 0) \cdot \overline{x}_1 \cdot \underline{x}_2 +$$
$$\max(p_2 - p_1, 0) \cdot \underline{x}_1 \cdot \overline{x}_2 + \min(1 - p_1, 1 - p_2) \cdot \underline{x}_1 \cdot \underline{x}_2,$$

where $p_i \stackrel{\text{def}}{=} (E_i - \underline{x}_i)/(\overline{x}_i - \underline{x}_i)$.

Theorem 2. *For multiplication under no information about dependence, to find $\underline{E}$, it is sufficient to consider the following combinations of p_1 and p_2:*

- $p_1 = \underline{p}_1$ *and* $p_2 = \underline{p}_2$; $p_1 = \underline{p}_1$ *and* $p_2 = \overline{p}_2$; $p_1 = \overline{p}_1$ *and* $p_2 = \underline{p}_2$; $p_1 = \overline{p}_1$ *and* $p_2 = \overline{p}_2$;
- $p_1 = \max(\underline{p}_1, 1 - \overline{p}_2)$ *and* $p_2 = 1 - p_1$ *(if* $1 \in \mathbf{p}_1 + \mathbf{p}_2$*); and*
- $p_1 = \min(\overline{p}_1, 1 - \underline{p}_2)$ *and* $p_2 = 1 - p_1$ *(if* $1 \in \mathbf{p}_1 + \mathbf{p}_2$*).*

The smallest value of $\underline{E}$ for all these cases is the desired lower bound $\underline{E}$.

Theorem 3. *For multiplication under no information about dependence, to find $\overline{E}$, it is sufficient to consider the following combinations of p_1 and p_2:*

- $p_1 = \underline{p}_1$ *and* $p_2 = \underline{p}_2$; $p_1 = \underline{p}_1$ *and* $p_2 = \overline{p}_2$; $p_1 = \overline{p}_1$ *and* $p_2 = \underline{p}_2$; $p_1 = \overline{p}_1$ *and* $p_2 = \overline{p}_2$;
- $p_1 = p_2 = \max(\underline{p}_1, \underline{p}_2)$ *(if* $\mathbf{p}_1 \cap \mathbf{p}_2 \neq \emptyset$*); and*
- $p_1 = p_2 = \min(\overline{p}_1, \overline{p}_2)$ *(if* $\mathbf{p}_1 \cap \mathbf{p}_2 \neq \emptyset$*).*

The largest value of $\overline{E}$ for all these cases is the desired upper bound $\overline{E}$.

For the *inverse* $y = 1/x_1$, the finite range is possible only when $0 \notin \mathbf{x}_1$. Without losing generality, we can consider the case when $0 < \underline{x}_1$. In this case, we get the following bound:

Theorem 4. *For the inverse $y = 1/x_1$, the range of possible values of E is $\mathbf{E} = [1/E_1, p_1/\overline{x}_1 + (1 - p_1)/\underline{x}_1]$.*

(Here p_1 denotes the same value as in Theorem 1).

Theorem 5. *For minimum $y = \min(x_1, x_2)$, when x_1 and x_2 are independent, we have $\overline{E} = \min(E_1, E_2)$ and*

$$\underline{E} = p_1 \cdot p_2 \cdot \min(\overline{x}_1, \overline{x}_2) + p_1 \cdot (1 - p_2) \cdot \min(\overline{x}_1, \underline{x}_2) +$$

$$(1 - p_1) \cdot p_2 \cdot \min(\underline{x}_1, \overline{x}_2) + (1 - p_1) \cdot (1 - p_2) \cdot \min(\underline{x}_1, \underline{x}_2).$$

Theorem 6. *For maximum $y = \min(x_1, x_2)$, when x_1 and x_2 are independent, we have $\underline{E} = \max(E_1, E_2)$ and*

$$\overline{E} = p_1 \cdot p_2 \cdot \max(\overline{x}_1, \overline{x}_2) + p_1 \cdot (1 - p_2) \cdot \max(\overline{x}_1, \underline{x}_2) +$$

$$(1 - p_1) \cdot p_2 \cdot \max(\underline{x}_1, \overline{x}_2) + (1 - p_1) \cdot (1 - p_2) \cdot \max(\underline{x}_1, \underline{x}_2).$$

Theorem 7. *For minimum $y = \min(x_1, x_2)$, when we have no information about the correlation between x_1 and x_2, we have $\overline{E} = \min(E_1, E_2)$,*

$$\underline{E} = \max(p_1 + p_2 - 1, 0) \cdot \min(\overline{x}_1, \overline{x}_2) + \min(p_1, 1 - p_2) \cdot \min(\overline{x}_1, \underline{x}_2) +$$

$$\min(1 - p_1, p_2) \cdot \min(\underline{x}_1, \overline{x}_2) + \max(1 - p_1 - p_2, 0) \cdot \min(\underline{x}_1, \underline{x}_2).$$

Theorem 8. *For maximum $y = \max(x_1, x_2)$, when we have no information about the correlation between x_1 and x_2, we have $\underline{E} = \max(E_1, E_2)$ and*

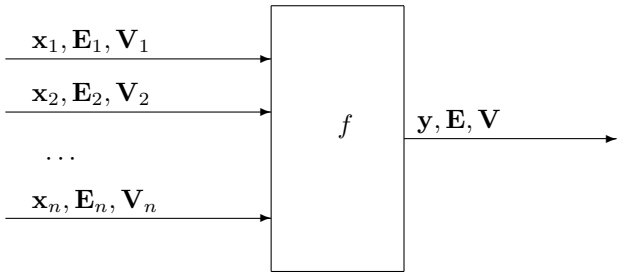
$$\overline{E} = \min(p_1, p_2) \cdot \max(\overline{x}_1, \overline{x}_2) + \max(p_1 - p_2, 0) \cdot \max(\overline{x}_1, \underline{x}_2) +$$

$$\max(p_2 - p_1, 0) \cdot \max(\underline{x}_1, \overline{x}_2) + \min(1 - p_1, 1 - p_2) \cdot \max(\underline{x}_1, \underline{x}_2).$$

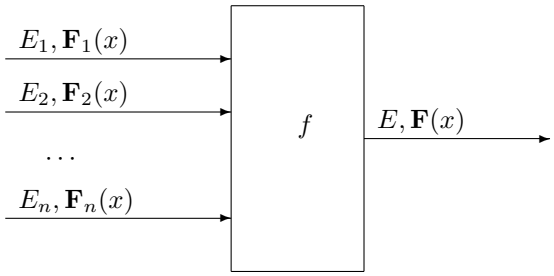
Similar formulas can be produced for the cases when there is a strong correlation between x_i: namely, when x_1 is (non-strictly) increasing or decreasing in x_2.

For products of several random variables, the corresponding problem is already NP-hard [24].

Challenges. What is, in addition to intervals and first moments, we also know second moments (this problem is important for design of computer chips):



What if, in addition to moments, we also know p-boxes?



6 Additional Results and Challenges

Estimating bounds on statistical characteristics. The above techniques assume that we already know the moments etc. How can we compute them based on the measurement results – taking into account that these results represent the actual (unknown) values with measurement uncertainty.

For example, in the case of interval uncertainty, instead of the exact sample values, we have only interval ranges $[\underline{x}_i, \overline{x}_i]$ for the sample values $x_1, \ldots, x_n$.

In this situation, we want to compute the ranges of possible values of the population mean $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$, population variance $V = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$, etc.

It turns out that most such problems are, in general, computationally difficult (to be more precise, NP-hard). Even computing the range $[\underline{V}, \overline{V}]$ of the population variance V is an NP-hard problem [16, 17]. In many practical situations, there exist feasible algorithms that compute the bounds of desirable statistical characteristics [4, 5, 11, 16, 17, 18, 22, 28, 40, 45, 49, 51]. For example, there exist efficient algorithms for computing $\underline{V}$ and efficient algorithms for computing $\overline{V}$ for several reasonable situations (e.g., when measurements are sufficiently accurate); efficient algorithms are also known for detecting outliers [12, 27].

An important issue is whether we can perform these computations on-line, updated the statistical characteristics as new measurements appear [29, 48].

When efficient algorithms are not known, we can use parallelization and quantum computing to speed up computation [26]. In many practical situations, there are still important open problems.

Computing amount of information. Another important problem is estimating amount of information, i.e., entropy. The traditional Shannon's definition described amount of information as the average number of "yes"-"no" questions that we need to ask to find the actual value (with given accuracy).

When we have finitely many alternatives, and we know the probabilities $p_1, \ldots, p_n$ of these alternatives, then this average number of "yes"-"no" questions is described by Shannon's entropy formula $S = -\sum_{i=1}^{n} p_i \cdot \log(p_i)$. In practice, we only have partial information about the probabilities, e.g., only intervals $[\underline{p}_i, \overline{p}_i]$ of possible values of p_i. Different values $p_i \in \mathbf{p}_i$ lead, in general, to different values S, so it is desirable to compute the range $[\underline{S}, \overline{S}]$ of possible values of S [20].

Since entropy S is a concave function, standard feasible algorithms for minimizing convex functions (= maximizing concave ones) enable us to compute $\overline{S}$; see, e.g., [20, 21]. Computing $\underline{S}$ is, in general, NP-hard [50], but for reasonable cases, feasible algorithms are possible [1, 2, 3, 31, 50].

Decision making. Computational aspects of decision making under interval and probabilistic uncertainty are discussed, e.g., in [52].

Acknowledgments. This work was supported in part by NASA under cooperative agreement NCC5-209, NSF grants EAR-0225670 and DMS-0532645, Star Award from the University of Texas System, and Texas Department of Transportation grant No. 0-5453.

References

1. J. Abellan and S. Moral, Range of entropy for credal sets, In: M. López-Diaz et al. (eds.), *Soft Methodology and Random Information Systems*, Springer, Berlin and Heidelberg, 2004, pp. 157–164.
2. J. Abellan and S. Moral, Difference of entropies as a nonspecificity function on credal sets, *Intern. J. of General Systems*, 2005, Vol. 34, No. 3, pp. 201–214.
3. J. Abellan and S. Moral, An algorithm that attains the maximum of entropy for order-2 capacities, *Intern. J. of Uncertainty, Fuzziness, and Knowledge-Based Systems* (to appear).
4. R. Aló, M. Beheshti, and G. Xiang, Computing Variance under Interval Uncertainty: A New Algorithm and Its Potential Application to Privacy in Statistical Databases, *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems IPMU'06*, Paris, France, July 2–7, 2006 (to appear).
5. J. B. Beck, V. Kreinovich, and B. Wu, Interval-Valued and Fuzzy-Valued Random Variables: From Computing Sample Variances to Computing Sample Covariances, In: M. Lopez, M. A. Gil, P. Grzegorzewski, O. Hrynewicz, and J. Lawry (eds.), *Soft Methodology and Random Information Systems*, Springer-Verlag, 2004, pp. 85–92.
6. D. Berleant, M.-P. Cheong, C. Chu, Y. Guan, A. Kamal, G. Sheblé, S. Ferson, and J. F. Peters, Dependable handling of uncertainty, *Reliable Computing* 9(6) (2003), pp. 407–418.
7. D. Berleant, L. Xie, and J. Zhang, Statool: a tool for Distribution Envelope Determination (DEnv), an interval-based algorithm for arithmetic on random variables, *Reliable Computing* 9(2) (2003), pp. 91–108.
8. D. Berleant and J. Zhang, Using Pearson correlation to improve envelopes around the distributions of functions, *Reliable Computing*, 10(2) (2004), pp. 139–161.
9. D. Berleant and J. Zhang, Representation and Problem Solving with the Distribution Envelope Determination (DEnv) Method, *Reliability Engineering and System Safety*, 85 (1–3) (July-Sept. 2004).
10. D. Berleant and J. Zhang, Using Pearson correlation to improve envelopes around the distributions of functions, *Reliable Computing*, 10(2) (2004), pp. 139–161.
11. E. Dantsin, V. Kreinovich, A. Wolpert, and G. Xiang, Population Variance under Interval Uncertainty: A New Algorithm, *Reliable Computing*, 2006, Vol. 12, No. 4, pp. 273–280.
12. E. Dantsin, A. Wolpert, M. Ceberio, G. Xiang, and V. Kreinovich, Detecting Outliers under Interval Uncertainty: A New Algorithm Based on Constraint Satisfaction, *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems IPMU'06*, Paris, France, July 2–7, 2006 (to appear).
13. S. Ferson, L. Ginzburg, and R. Akcakaya, *Whereof One Cannot Speak: When Input Distributions Are Unknown*, Applied Biomathematics Report, 2001.
14. S. Ferson, J. Hajagos, D. Berleant, J. Zhang, W. T. Tucker, L. Ginzburg, and W. Oberkampf, *Dependence in Dempster-Shafer Theory and Probability Bounds Analysis*, Technical Report SAND2004-3072, Sandia National Laboratory, 2004.

15. S. Ferson, V. Kreinovich, L. Ginzburg, D. S. Myers, and K. Sentz, *Constructing Probability Boxes and Dempster-Shafer Structures*, Sandia National Laboratories, Report SAND2002-4015, January 2003.
16. S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpre, and M. Aviles, Exact Bounds on Finite Populations of Interval Data, *Reliable Computing*, 2005, Vol. 11, No. 3, pp. 207–233.
17. S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, and M. Aviles, Computing Variance for Interval Data is NP-Hard, *ACM SIGACT News*, 2002, Vol. 33, No. 2, pp. 108–118.
18. L. Granvilliers, V. Kreinovich, and N. Mueller, Novel Approaches to Numerical Software with Result Verification, In: R. Alt, A. Frommer, R. B. Kearfott, and W. Luther (eds.), *Numerical Software with Result Verification*, International Dagstuhl Seminar, Dagstuhl Castle, Germany, January 19–24, 2003, Revised Papers, Springer Lectures Notes in Computer Science, 2004, Vol. 2991, pp. 274–305.
19. L. Jaulin, M. Kieffer, O. Didrit, and E. Walter, *Applied interval analysis: with examples in parameter and state estimation, robust control and robotics*, Springer Verlag, London, 2001.
20. G. J. Klir, *Uncertainty and Information: Foundations of Generalized Information Theory*, J. Wiley, Hoboken, New Jersey, 2005.
21. V. Kreinovich, Maximum entropy and interval computations, *Reliable Computing*, 1996, Vol. 2, No. 1, pp. 63–79.
22. V. Kreinovich, Probabilities, Intervals, What Next? Optimization Problems Related to Extension of Interval Computations to Situations with Partial Information about Probabilities, *Journal of Global Optimization*, 2004, Vol. 29, No. 3, pp. 265–280.
23. V. Kreinovich and S. Ferson, Computing Best-Possible Bounds for the Distribution of a Sum of Several Variables is NP-Hard, *International Journal of Approximate Reasoning*, 2006, Vol. 41, pp. 331–342.
24. V. Kreinovich, S. Ferson, and L. Ginzburg, Exact Upper Bound on the Mean of the Product of Many Random Variables With Known Expectations, *Reliable Computing* (to appear).
25. V. Kreinovich, A. Lakeyev, J. Rohn, P. Kahl, *Computational complexity and feasibility of data processing and interval computations*, Kluwer, Dordrecht, 1997.
26. V. Kreinovich and L. Longpre, Fast Quantum Algorithms for Handling Probabilistic and Interval Uncertainty, *Mathematical Logic Quarterly*, 2004, Vol. 50, No. 4/5, pp. 507–518.
27. V. Kreinovich, L. Longpre, P. Patangay, S. Ferson, and L. Ginzburg, Outlier Detection Under Interval Uncertainty: Algorithmic Solvability and Computational Complexity, *Reliable Computing*, 2005, Vol. 11, No. 1, pp. 59–75.
28. V. Kreinovich, L. Longpré, S. A. Starks, G. Xiang, J. Beck, R. Kandathi, A. Nayak, S. Ferson, and J. Hajagos, Interval Versions of Statistical Techniques, with Applications to Environmental Analysis, Bioinformatics, and Privacy in Statistical Databases", *Journal of Computational and Applied Mathematics* (to appear).
29. V. Kreinovich, H. T. Nguyen, and B. Wu, On-Line Algorithms for Computing Mean and Variance of Interval Data, and Their Use in Intelligent Systems, Information Sciences (to appear).
30. V. Kreinovich, G. N. Solopchenko, S. Ferson, L. Ginzburg, and R. Alo, Probabilities, intervals, what next? Extension of interval computations to situations with

partial information about probabilities, *Proceedings of the 10th IMEKO TC7 International Symposium on Advances of Measurement Science*, St. Petersburg, Russia, June 30–July 2, 2004, Vol. 1, pp. 137–142.
31. V. Kreinovich, G. Xiang, and S. Ferson, How the Concept of Information as Average Number of 'Yes-No' Questions (Bits) Can Be Extended to Intervals, P-Boxes, and more General Uncertainty, *Proceedings of the 24th International Conference of the North American Fuzzy Information Processing Society NAFIPS'2005*, Ann Arbor, Michigan, June 22–25, 2005, pp. 80–85.
32. V. Kuznetsov, *Interval Statistical Models* (in Russian), Radio i Svyaz, Moscow, 1991.
33. W. A. Lodwick and K. D. Jamison, Estimating and validating the cumulative distribution of a function of random variables: toward the development of distribution arithmetic, *Reliable Computing*, 2003, Vol. 9, No. 2, pp. 127–141.
34. W. A. Lodwick, A. Neumaier, and F. Newman, Optimization under uncertainity: methods and applications in radiation therapy, *Proc. 10th IEEE Int. Conf. Fuzzy Systems*, December 2–5, 2001, Melbourne, Australia.
35. C. Manski, *Partial Identification of Probability Distributions*, Springer-Verlag, New York, 2003.
36. A. S. Moore, Interval risk analysis of real estate investment: a non-Monte-Carlo approach, *Freiburger Intervall-Berichte* 85/3, Inst. F. Angew. Math., Universitaet Freiburg I. Br., 23–49 (1985).
37. R. E. Moore, *Automatic error analysis in digital computation*, Technical Report Space Div. Report LMSD84821, Lockheed Missiles and Space Co., 1959.
38. R. E. Moore, Risk analysis without Monte Carlo methods, *Freiburger Intervall-Berichte* 84/1, Inst. F. Angew. Math., Universitaet Freiburg I. Br., 1–48 (1984).
39. A. Neumaier, Fuzzy modeling in terms of surprise, *Fuzzy Sets and Systems* 135, 2003, 21–38.
40. H. T. Nguyen, V. Kreinovich, and G. Xiang, Foundations of Statistical Processing of Set-Valued Data: Towards Efficient Algorithms, *Proceedings of the Fifth International Conference on Intelligent Technologies InTech'04*, Houston, Texas, December 2–4, 2004.
41. M. Orshansky, W.-S. Wang, M. Ceberio, and G. Xiang, Interval-Based Robust Statistical Techniques for Non-Negative Convex Functions, with Application to Timing Analysis of Computer Chips, *Proceedings of the Symposium on Applied Computing SAC'06*, Dijon, France, April 23–27, 2006, pp. 1645–1649.
42. M. Orshansky, W.-S. Wang, G. Xiang, and V. Kreinovich, Interval-Based Robust Statistical Techniques for Non-Negative Convex Functions, with Application to Timing Analysis of Computer Chips, *Proceedings of the Second International Workshop on Reliable Engineering Computing*, Savannah, Georgia, February 22–24, 2006, pp. 197–212.
43. H. Regan, S. Ferson and D. Berleant, Equivalence of methods for uncertainty propagation of real-valued random variables, *International Journal of Approximate Reasoning*, in press.
44. H.-P. Schröcker and J. Wallner, Geometric constructions with discretized random variables, *Reliable Computing*, 2006, Vol. 12, No. 3, pp. 203–223.
45. S. A. Starks, V. Kreinovich, L. Longpre, M. Ceberio, G. Xiang, R. Araiza, J. Beck, R. Kandathi, A. Nayak, and R. Torres, Towards combining probabilistic and interval uncertainty in engineering calculations, *Proceedings of the Workshop on Reliable Engineering Computing*, Savannah, Georgia, September 15–17, 2004, pp. 193–213.

46. W. T. Tucker and S. Ferson, *Probability Bounds Analysis in Environmental Risk Assessments*, Applied Biomathematics Report.
47. P. Walley, *Statistical Reasoning with Imprecise Probabilities*, Chapman & Hall, N.Y., 1991.
48. B. Wu, H. T. Nguyen, and V. Kreinovich, Real-Time Algorithms for Statistical Analysis of Interval Data, *Proceedings of the International Conference on Information Technology InTech'03*, Chiang Mai, Thailand, December 17–19, 2003, pp. 483–490.
49. G. Xiang, Fast algorithm for computing the upper endpoint of sample variance for interval data: case of sufficiently accurate measurements, *Reliable Computing*, 2006, Vol. 12, No. 1, pp. 59–64.
50. G. Xiang, O. Kosheleva, and G. J. Klir, Estimating Information Amount under Interval Uncertainty: Algorithmic Solvability and Computational Complexity, *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems IPMU'06*, Paris, France, July 2–7, 2006 (to appear).
51. G. Xiang, S. A. Starks, V. Kreinovich, and L. Longpre, New Algorithms for Statistical Analysis of Interval Data, *Proceedings of the Workshop on State-of-the-Art in Scientific Computing PARA'04*, Lyngby, Denmark, June 20–23, 2004, Vol. 1, pp. 123–129.
52. R. R. Yager and V. Kreinovich, Decision Making Under Interval Probabilities, *International Journal of Approximate Reasoning*, 1999, Vol. 22, No. 3, pp. 195–215.
53. J. Zhang and D. Berleant, Envelopes around cumulative distribution functions from interval parameters of standard continuous distributions, *Proceedings, North American Fuzzy Information Processing Society (NAFIPS 2003)*, Chicago, pp. 407–412.