Computing Best-Possible Bounds for the Distribution of a Sum of Several Variables is NP-Hard ⋆

Vladik Kreinovich *

Department of Computer Science, University of Texas at El Paso, El Paso, TX 79968, USA

Scott Ferson

Applied Biomathematics, 100 North Country Road. Setauket, NY 11733, USA

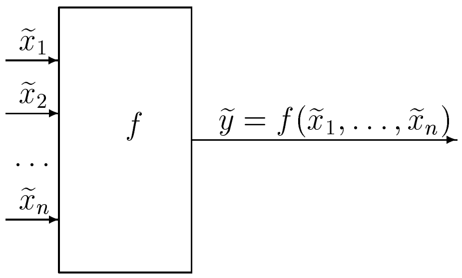
Abstract

In many real-life situations, we know the probability distribution of two random variables x_1 and x_2, but we have no information about the correlation between x_1 and x_2; what are the possible probability distributions for the sum $x_1 + x_2$? This question was originally raised by A. N. Kolmogorov. Algorithms exist that provide best-possible bounds for the distribution of $x_1 + x_2$; these algorithms have been implemented as a part of the efficient software for handling probabilistic uncertainty. A natural question is: what if we have several ($n > 2$) variables with known distribution, we have no information about their correlation, and we are interested in possible probability distribution for the sum $y = x_1 + \ldots + x_n$? Known formulas for the case $n = 2$ can be (and have been) extended to this case. However, as we prove in this paper, not only are these formulas not best-possible anymore, but in general, computing the best-possible bounds for arbitrary n is an NP-hard (computationally intractable) problem.

Key words: sum of random variables, best-possible bounds, NP-hard

1 Error Estimating for Indirect Measurements: Practical Problem; What Is Known about Its Solution

Real-life problem: error estimation for 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 $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.

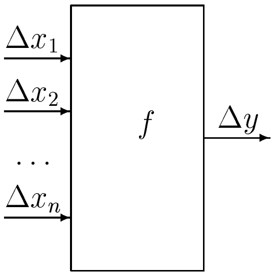


Computing an estimate for y based on the results of direct measurements is called *data processing*; data processing is the main reason why computers were invented in the first place, and data processing is still one of the main uses of computers as number crunching devices.

* This work was supported in part by NASA under cooperative agreement NCC5-209, by the Future Aerospace Science and Technology Program (FAST) Center for Structural Integrity of Aerospace Systems, effort sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant F49620-00-1-0365, by NSF grants EAR-0112968, EAR-0225670, and EIA-0321328, and by the Army Research Laboratories grant DATM-05-02-C-0046.
This work was also supported, in part, by Small Business Innovation Research grant 9R44CA81741 to Applied Biomathematics from the National Cancer Institute (NCI), a component of the National Institutes of Health (NIH), and by a research grant from Sandia National Laboratories as part of the Department of Energy Accelerated Strategic Computing Initiative (ASCI).
The authors are thankful to the anonymous referees for useful suggestions.

* Corresponding author
Email addresses: `vladik@utep.edu` (Vladik Kreinovich), `scott@ramas.com` (Scott Ferson).

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; see, e.g., [19]. 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.

Let us review the main types of this information.

Case of interval uncertainty. For measuring instruments, a manufacturer usually provides an upper bound Δ on the measurement error $\Delta x = \widetilde{x} - x$.

Indeed, a word "measurement" usually means that as a result, we get some guaranteed information about the measured quantity; so after we observe a measurement result $\widetilde{x}$, we should be able to conclude that the actual (unknown) values of the measured quantity is bounded by some bounds.

In many practical situations, this upper bound Δ is the only information that we have about the measurement's accuracy. In such situations, after we perform the measurement and record the measured value as $\widetilde{x}$, the only information that we have about the actual (unknown) value of the measured quantity x is that x must be within the interval $\mathbf{x} = [\widetilde{x} - \Delta, \widetilde{x} + \Delta]$.

As a result, we arrive at the following problem: based on n known intervals $\mathbf{x}_1, \ldots, \mathbf{x}_n$, and on a known function $y = f(x_1, \ldots, x_n)$, we must determine the range $\mathbf{y}$ of the function $y = f(x_1, \ldots, x_n)$ when $x_i \in \mathbf{x}_i$ for all i. The problem of computing such a range is known as the problem of *interval computations*; see, e.g., [7,10].

It is known that in general, the problem of computing the range exactly (or even with a given accuracy ε) is NP-hard (computationally intractable). Specifically:

- for *linear* functions $f(x_1,\ldots,x_n) = a_0 + a_1 \cdot x_1 + \ldots + a_n \cdot x_n$, there is an *efficient* algorithm for computing the exact range for y;
- however, already for general *quadratic* functions $f(x_1,\ldots,x_n) = a_0 + \sum_{i=1}^{n} a_i \cdot x_i + \sum_{i=1}^{n} \sum_{=1}^{n} a_{ij} \cdot x_i \cdot x_j$, the problem of computing the exact (or approximate) range is *NP-hard*.

Case of fuzzy uncertainty. Often, in addition to (or instead of) the guaranteed bound for Δx_i, an expert can provide bounds that contain Δx_i with a certain degree of confidence. Usually, we know several such bounding intervals corresponding to different degrees of confidence. Such a nested family of intervals is also called a *fuzzy set*, because it turns out to be equivalent to a more traditional definition of fuzzy set [1,8,13–15] (if a traditional fuzzy set is given, then different intervals from the nested family can be viewed as α-cuts corresponding to different levels of uncertainty α).

In the case of fuzzy uncertainty, for each degree of confidence α, we must solve the problem corresponding to the α-cut intervals; thus, instead of a *single* interval of possible values of $y = f(x_1,\ldots,x_n)$, we get a nested *family* of intervals corresponding to different α – i.e., a fuzzy-valued range for $y = f(x_1,\ldots,x_n)$.

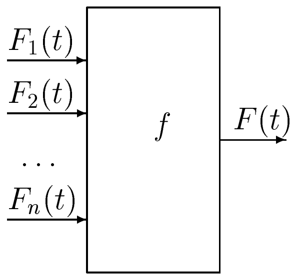
Similarly to the interval case:

- the problem of computing this fuzzy range can be effectively solved for linear functions $f(x_1,\ldots,x_n)$;
- however, for quadratic functions $y = f(x_1,\ldots,x_n)$, this problem is NP-hard.

Ideal case: Probabilistic uncertainty. In the ideal case, if each measuring instrument has been thoroughly analyzed and calibrated, we know the exact probability distribution for each random variable Δx_i. As a result, after i-th measurement, we know the probability distribution of actual values x_i. This probability distribution can be described, e.g., by the corresponding cumulative distribution function (cdf) $F_i(t) \stackrel{\rm def}{=} \text{Prob}(x_i < t)$.

It is worth mentioning that in most practical cases, the distribution functions $F_i(t)$ are not Gaussian; see, e.g., [16,17].

Based on n known cdfs $F_1(t),\ldots,F_n(t)$ and on a known function $y = f(x_1,\ldots,x_n)$, we must determine the distribution (cdf) $F(t)$ for $y = f(x_1,\ldots,x_n)$.



When measurement errors of individual x_i are possibly correlated, the corresponding probabilistic problem becomes difficult. When all the measurement errors are independent, i.e., if x_i are independent random variables, then we can, e.g., use Monte-Carlo simulations and/or analytical formulas to come up with the desired distribution for y. In many practical situations, however, we know that the measurement errors of different measuring instruments are not independent, because they contain components that come from the same outside error source (e.g., from the power grid).

Ideally, we should find out how exactly the variables x_i are correlated, i.e., we should get the joint probability distribution of the corresponding n variables. Unfortunately, this is very difficult:

- to get a distribution of a *single* variable with "k bins" accuracy, it is sufficient to divide the real line into k bins;
- however, to describe a joint distribution of n variables with the same accuracy, we need k^n bins.

For large n, the number k^n becomes larger than the number of particles in the Universe (see, e.g., [10,18]), so this is not practically possible.

As a result, for n variables, we face the following problem: we know the distributions $F_i(t)$ for n variables $x_1, \dots, x_n$, we know the function $y = f(x_1, \dots, x_n)$, but we do not have any information about the correlation between x_i. In such situation, there may be many different joint distributions for $x_1, \dots, x_n$, and for these different joint distributions, we may get different distributions $F(t)$ for y. What we would like to find, in this situation, is the range $[\underline{F}(t), \overline{F}(t)]$ of possible values of $F(t)$ for each t. In other words, we would like to find the best-possible bounds for a distribution of a function $y = f(x_1, \dots, x_n)$ of several random variables $x_1, \dots, x_n$. Let us formulate this problem in precise mathematical terms.

Formulation of the problem in mathematical terms. We know n cdfs $F_1(t), \dots, F_n(t)$, and we know a function $y = f(x_1, \dots, x_n)$ from R^n to R. Based on this information, we would like to compute the range $[\underline{F}(t), \overline{F}(t)]$, where:

- $\underline{F}(t)$ is the infimum of possible values $F(t) = \text{Prob}(f(x_1, \ldots, x_n) < t)$ over all joint distributions of $(x_1, \ldots, x_n)$ for which the marginal distributions coincide with the given cdfs $F_i(t)$, and
- $\overline{F}(t)$ is the supremum of possible values $F(t)$ over all such joint distributions.

What is known: case of $n = 2$ variables. In spite of the clear practical importance of this problem, no general solution was known until the early 1980s, when G. D. Makarov, a student of A. N. Kolmogorov, in his paper [12], provided the exact formulas for $\underline{F}(t)$ and $\overline{F}(t)$ for the simplest case when $n = 2$ and $f(x_1, x_2) = x_1 + x_2$. These formulas were later simplified, in the paper [5], into the following form:

$$\underline{F}(t) = \max_{t_1, t_2 : t_1 + t_2 = t} \max(F_1(t_1) + F_2(t_2) - 1, 0); \tag{1}$$

$$\overline{F}(t) = \min_{t_1, t_2 : t_1 + t_2 = t} \min(F_1(t_1) + F_2(t_2), 1). \tag{2}$$

The fact that these formulas do provide lower and upper bounds for $F(t)$ is reasonably easy to understand. Indeed, it is well known that for any two events A and B, the probability $P(A \vee B)$ cannot exceed $P(A) + P(B)$. Since $A \,\&\, B$ is equivalent to $\neg(\neg A \vee \neg B)$, we conclude that

$$1 - P(A \,\&\, B) = P(\neg A \vee \neg B) \le P(\neg A) + P(\neg B) = (1 - P(A)) + (1 - P(B)),$$

hence $P(A \,\&\, B) \ge P(A) + P(B) - 1$. Since the probability is always non-negative, we conclude that $P(A \,\&\, B) \ge \max(P(A) + P(B) - 1, 0)$.

For every t_1 and t_2 for which $t_1 + t_2 = t$, the inequalities $x_1 < t_1$ and $x_2 < t_2$ imply that $y \stackrel{\text{def}}{=} x_1 + x_2 < t_1 + t_2$. Thus, the probability $F(t)$ that $y < t$ cannot be smaller than the probability $\text{Prob}((x_1 < t_1) \,\&\, (x_2 < t_2))$. Due to the above inequality, this probability, in turn, cannot be smaller than

$$\max(\text{Prob}(x_1 < t_1) + \text{Prob}(x_2 < t_2) - 1, 0) = \max(F_1(t_1) + F_2(t_2) - 1, 0),$$

so $F(t) \ge \max(F_1(t_1) + F_2(t_2) - 1, 0)$. Since $F(t)$ is larger than or equal to this expression for all t_1 and t_2 for which $t_1 + t_2 = 1$, it must be also larger than or equal to the largest of these expressions – which is exactly the above lower bound $\underline{F}(t)$.

The proof that the expression (2) is the upper bound is similar. The non-trivial part of the result (1), (2) is proving that these bounds are indeed the best possible.

Further developments: brief overview. The seminal paper [22] extended the above formulas to the situations with more complex functions $f(x_1, x_2)$ and/or situations in which we have some information about the correlation between x_1 and x_2. The formulas proposed in [22] formed the basis for an efficient software system RiskCalc for handling probabilistic uncertainty (see, e.g., [2]); for a theoretical foundation of the corresponding formulas, see, e.g., [4,20].

Similar formulas have also been analyzed, clarified, implemented, and tested in [11].

Case of $n > 2$ variables: what is known. For the sum of $n > 2$ random variables $x_1, \ldots, x_n$, similar arguments lead to similar formulas. Specifically, it is known that for any two sequence of events $A_1, \ldots, A_n$, the probability $P(A_1 \vee \ldots \vee A_n)$ cannot exceed $P(A_1) + \ldots + P(A_n)$. Since $A_1 \,\&\, \ldots \,\&\, A_n$ is equivalent to $\neg((\neg A_1) \vee \ldots \vee (\neg A_n))$, we conclude that

$$1 - P(A_1 \,\&\, \ldots \,\&\, A_n) = P((\neg A_1) \vee \ldots \vee (\neg A_n)) \le$$

$$P(\neg A_1) + \ldots + P(\neg A_n) = (1 - P(A_1)) + \ldots + (1 - P(A_n)),$$

hence $P(A_1 \,\&\, \ldots \,\&\, A_n) \ge P(A_1) + \ldots + P(A_n) - (n-1)$. Since the probability is always non-negative, we conclude that

$$P(A_1 \,\&\, \ldots \,\&\, A_n) \ge \max(P(A_1) + \ldots + P(A_n) - (n-1), 0).$$

Now, for every tuple $(t_1, \ldots, t_n)$ for which $t_1 + \ldots + t_n = t$, the inequalities $x_1 < t_1$, $\ldots$, $x_2 < t_2$ imply that $y \stackrel{\text{def}}{=} x_1 + \ldots + x_n < t_1 + \ldots + t_n$. Thus, the probability $F(t)$ that $y < t$ cannot be smaller than the probability

$$\text{Prob}((x_1 < t_1) \,\&\, \ldots \,\&\, (x_n < t_n)).$$

Due to the above inequality, this probability, in turn, cannot be smaller than

$$\max(\text{Prob}(x_1 < t_1) + \ldots + \text{Prob}(x_n < t_n) - (n-1), 0) =$$

$$\max(F_1(t_1) + \ldots + F_n(t_n) - (n-1), 0),$$

so $F(t) \ge \max(F_1(t_1) + \ldots + F_n(t_n) - (n-1), 0)$. Since $F(t)$ is larger than or equal to this expression for all tuples $(t_1, \ldots, t_n)$ for which $t_1 + \ldots + t_n = 1$, it must be also larger than or equal to the largest of these expressions – hence $F(t) \ge F^-(t)$, where

$$F^-(t) \stackrel{\text{def}}{=} \max_{t_1,\ldots,t_n : t_1+\ldots+t_n=t} \max(F_1(t_1) + \ldots + F_n(t_n) - (n-1), 0). \qquad (3)$$

Similarly, we can conclude that $F(t) \le F^+(t)$, where

$$F^+(t) \stackrel{\text{def}}{=} \min_{t_1,\ldots,t_n : t_1+\ldots+t_n=t} \min(F_1(t_1) + \ldots + F_n(t_n), 1). \qquad (4)$$

2 New Result: For Probabilistic Uncertainty, the Problem of Error Estimation for Indirect Measurements Is NP-hard

Precise formulation of the problem: reminder. We know n cdfs $F_1(t), \ldots, F_n(t)$, and we know a function $y = f(x_1, \ldots, x_n)$ from R^n to R. Based on this information, we would like to compute the range $[\underline{F}(t), \overline{F}(t)]$, where:

- $\underline{F}(t)$ is the infimum of possible values $F(t) = \text{Prob}(f(x_1, \ldots, x_n) < t)$ over all joint distributions of $(x_1, \ldots, x_n)$ for which the marginal distributions coincide with the given cdfs $F_i(t)$, and
- $\overline{F}(t)$ is the supremum of possible values $F(t)$ over all such joint distributions.

Question: are the known bounds (3)–(4) best possible? In the first section, we have described formulas that bound the desired values $\underline{F}(t)$ and $\overline{F}(t)$. For $n = 2$, as we have mentioned, these bounds are the best possible. A natural question is: are the corresponding bounds (3)–(4) best possible for $n > 2$ as well?

The paper [21] implicitly formulates a hypothesis that these bounds are indeed the best possible.

Our results. In this paper:

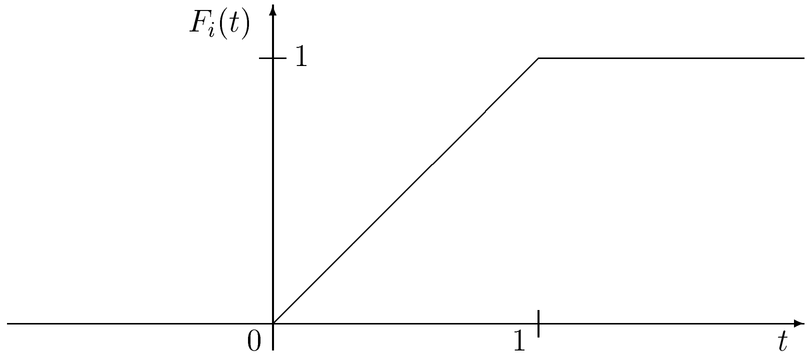
- first, we show that these bounds are not the best possible;
- second, we prove that computing the best-possible bounds for a general n is an NP-hard (computationally intractable) problem.

Comments.

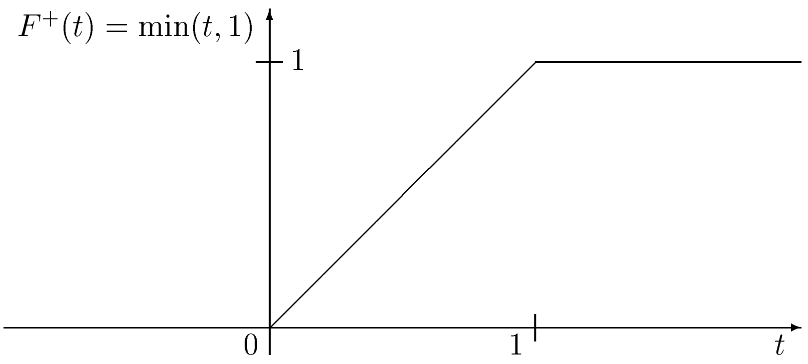
- It is worth repeating that, as we have mentioned, for linear functions $f = a_0 + \sum\limits_{i=1}^{n} a_i \cdot x_i$, under *interval* and *fuzzy* uncertainties the problem of computing the uncertainty of y can be efficiently solved – in contrast to the *probabilistic* case.
- Since in the probabilistic case, the problem is NP-hard already for *linear* functions $y = f(x_1, \ldots, x_n)$, this problem remains NP-hard for more general classes of functions, e.g., for all *quadratic* functions, for all *smooth* functions, etc.
- Sometimes, the data processing algorithm is not smooth. For example, we may select the largest value $y = \max(x_1, \ldots, x_n)$, or we may deal with a threshold $y = \theta(a_0 + a_1 \cdot x_1 + \ldots + a_n \cdot x_n)$, where $\theta(x) = 0$ for $x < 0$ and $\theta(x) = 1$ for $x \geq 0$. What can we then say about the cdf $F(t)$ of the corresponding quantity y?

- For *maximum*, there is an explicit formula for the bounds on $F(t)$. Indeed, $\max(x_1, \ldots .x)n) < t$ if and only if $(x_1 < t) \,\&\, \ldots (x_n < t)$. Since we know the values $F_i(t) = \text{Prob}(x_i < t)$, we can thus conclude that $\underline{F}(t) = \max(F_1(t) + \ldots + F_n(t) - (n-1), 0)$ and $\overline{F}(t) = \min(F_1(t), \ldots, F_n(t))$.
- In contrast, for the *threshold function* $y = \theta(a_0 + a_1 \cdot x_1 + \ldots + a_n \cdot x_n)$, the condition $y < 1$ is equivalent to $a_1 \cdot x_1 + \ldots + a + n \cdot x_n < -a_0$. Thus, even for $a_1 = \ldots = a_n = 1$, the value $F(1)$ is equal to the probability $\text{Prob}(x_1 + \ldots + x_n < -a_0)$. Since, as we prove in this paper, the problem of computing the best-possible bound for this probability is NP-hard, the problem of computing the best-possible bounds for $F(t)$ is NP-hard as well.

First result: Example when the bounds (3)–(4) are not the best possible. We will consider the simplest possible example when $n = 3$ and all 3 distributions are uniform distributions on the interval $[0, 1]$, i.e., $F_i(t) = 0$ for $t \leq 0$, $F_i(t) = t$ for $0 \leq t \leq 1$, and $F_i(t) = 1$ for $t \geq 1$.



In this case, once $t_i \in [0, 1]$, we have $F_1(t_1) + F_2(t_2) + F_3(t_3) = t_1 + t_2 + t_3$. Therefore, once $t_1 + t_2 + t_3 = t$, we have $F_1(t_1) + F_2(t_2) + F_3(t_3) = t_1 + t_2 + t_3 = t$ hence $\min(F_1(t_1) + F_2(t_2) + F_3(t_3), 1) = \min(t, 1)$. Therefore, the minimum in the formula (4) is the minimum of identical values, hence $F^+(t) = \min(t, 1)$. In particular, for $t = 1$, we have $F^+(1) = \min(1, 1) = 1$.



So, for an arbitrary joint distribution of 3 random variables x_1, x_2, x_3 for which each marginal distribution is uniform on $[0,1]$, for the cdf $F(t)$ of the sum $y = x_1 + x_2 + x_3$, we have $F(1) \le F^+(1) = 1$.

Let us now show that this bound $F(1) \le F^+(1) = 1$ cannot be the best possible, i.e., that we cannot have $F(1) = 1$. Indeed, if $F(1) = 1$, this means that with probability 1, we have $y < 1$. Thus, the expected value $E[y]$ of y cannot exceed 1: $E[y] \le 1$. On the other hand, since $y = x_1 + x_2 + x_3$, we have $E[y] = E[x_1] + E[x_2] + E[x_3] = 3 \cdot 0.5 = 1.5$ – a contradiction with the fact that $E[y] = 1$.

We can show that not only $F(1)$ cannot be equal to 1, it cannot be even close to 1: e.g., if $E(1) \ge 0.9$, this means that the probability that $y \ge 1$ is at most 0.1. So, with probability ≤ 1, we have $y \le 1$, and with probability ≤ 0.1, we have $y = x_1 + x_2 + x_3 \le 3 \cdot 1 = 3$. Thus, the expected value $E[y]$ of y cannot exceed $1 \cdot 1 + 0.1 \cdot 3 = 1.3$ – still a contradiction.

In this particular example, we can add additional inequalities on the cdf $F(t)$ caused by the fact that we know the value $E[y] = 1.5$ of the first moment (see, e.g., [4]). We will show, however, that in general, the problem of computing the best-possible bounds on $F(t)$ is NP-hard.

Second result: Computing best-possible bounds for the distribution of a sum of several variables is NP-hard – a proof. To prove NP-hardness of the problem of computing the best-possible bounds for $F(t)$, we will reduce, to this problem, a known NP-problem, namely, the following *partition* problem (see, e.g., [10,18]): given n positive integers $s_1, \ldots, s_n$, check whether it is possible to find values $\varepsilon_i \in \{-1, 1\}$ for which $\varepsilon_1 \cdot s_1 + \ldots + \varepsilon_n \cdot s_n = 0$.

We will reduce each instance of this problem to the case when we have n random variables; for every i from 1 to n, i-th variable x_i is equal to $-s_i$ with probability 1/2 and to s_i with probability 1/2. For each of these variables, we have $E[x_i] = (1/2) \cdot (-s_i) + (1/2) \cdot s_i = 0$, hence for their sum $y \stackrel{\text{def}}{=} x_1 + \ldots + x_n$, we have $E[y] = E[x_1] + \ldots + E[x_n] = 0$.

Let us show that $\underline{F}(0) = 0$ if and only if the original instance of the partition problem has a solution. Indeed, if the original instance has a solution $(\varepsilon_1, \ldots, \varepsilon_n)$, then we can take the joint distribution in which $x = (x_1, \ldots, x_n)$ is equal to $(\varepsilon_1 \cdot s_1, \ldots, \varepsilon_n \cdot s_n)$ with probability 1/2 and to $(-\varepsilon_1 \cdot s_1, \ldots, -\varepsilon_n \cdot s_n)$ with probability 1/2. In this case, all n marginal distributions are as desired; on the other hand, the sum $y = x_1 + \ldots + x_n$ is equal to 0 with probability 1, hence $F(0) = \text{Prob}(y < 0) = 0$.

Vice versa, let us assume that $\underline{F}(0) = 0$. By definition of $\underline{F}(t)$, this means that for every $\delta > 0$, there exists a joint distribution for which $F(0) \le \delta$. Let us select some small $\varepsilon > 0$ (we will later determine which value to select), and let us select a distribution F that satisfies the above inequality for this δ. We will use reduction to a contradiction to prove that in this case, the original instance of the partition problem has a solution.

According to our choice of the random variables x_i, the only possible values of x_i are $\pm s_i$, i.e., the values $\varepsilon_i \cdot s_i$ for some $\varepsilon_i \in \{-1, 1\}$. So, if the original instance does not have a solution, then all possible values of $y = \sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} \varepsilon_i \cdot s_i$ are non-zero integers. Thus, if $y \ge 0$, we have $y \ge 1$.

The smallest possible value of y is $-S$, where $S \stackrel{\text{def}}{=} s_1 + \ldots + s_n$.

The expected value $E[y] = \sum\limits_j p_j \cdot y_j$ of y can be represented as the sum $E = E^+ + E^-$ of two sub-sums E^+ and E^- corresponding to positive and negative y_j.

For the joint distribution F for which $F(0) = \text{Prob}(y < 0) \le \delta$, with probability $\le \delta$, we have values $\ge -S$, and with probability at least $1 - \delta$, we have values ≥ 1.

The overall probability of positive values is at least $1 - \delta$, and each positive value is at least 1, so $E^+ \ge (1 - \delta) \cdot 1 = 1 - \delta$. On the other hand, the probability of negative values is $\le \delta$, and each negative value is $\ge -S$, so $E^- \ge -\delta \cdot S$. Therefore, $E[y] = E^- + E^+ \ge (1 - \delta) - \delta \cdot S = 1 - (S + 1) \cdot \delta$; so, for $\delta < 1/(S + 1)$, we have $E[y] > 0$ – a contradiction with $E[y] = 0$. This contradiction shows that our assumption was false; hence, the original instance of the partition problem has a solution.

The reduction is proven, thus computing best-possible bounds for the distribution of a sum of several variables is indeed NP-hard.

Comment. The fact that problem turns out to be NP-hard is not very surprising: many other interval problems are NP-hard (see, e.g., [10]), as well as many problems related to combination of interval and probabilistic uncertainty (see, e.g., [3,6,9]).

Conclusion

In many practical situations, we know the probability distribution of several random variables $x_1, \ldots, x_n$, but we have no information about the correlation between x_i; what are the possible probability distributions for the sum $y = x_1 + \ldots + x_n$? This question was originally raised by A. N. Kolmogorov. For $n = 2$, there exist efficient algorithms that provide best-possible bounds for the distribution of $x_1 + x_2$; these algorithms have been implemented as a part of the efficient software for handling probabilistic uncertainty.

The known formulas can be extended to the case $n > 2$. In this paper, we have proven that for $n > 2$, the known formulas are not best-possible. Moreover, we have proven that for $n > 2$, the problem of computing the best-possible bounds is, in general, NP-hard (computationally intractable).

References

[1] G. Bojadziev and M. Bojadziev, *Fuzzy Sets, Fuzzy Logic, Applications*, World Scientific, Singapore, 1995.

[2] S. Ferson, *RAMAS Risk Calc 4.0: Risk Assessment with Uncertain Numbers*, CRC Press, Boca Raton, Florida, 2002.

[3] 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.

[4] S. Ferson, D. Myers, and D. Berleant, *Distribution-free risk analysis: I. Range, mean, and variance*, Applied Biomathematics, Technical Report, 2001.

[5] M. J. Frank, R. B. Nelsen, and B. Schweizer, "Best-possible bounds for the distribution of a sum – a problem of Kolmogorov", *Probability Theory and Related Fields*, 1987, Vol. 74, pp. 199–211.

[6] L. Granvilliers, V. Kreinovich, and N. Müller, "Novel Approaches to Numerical Software with Result Verification", In: R. Alt, A. Frommer, R. B. Kearfott, and W. Luther, editors, *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.

[7] 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.

[8] G. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic: Theory and Applications*, Prentice Hall, Upper Saddle River, New Jersey, 1995.

[9] 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.

[10] V. Kreinovich, A. Lakeyev, J. Rohn, and P. Kahl, *Computational Complexity and Feasibility of Data Processing and Interval Computations*, Kluwer, Dordrecht, 1997.

[11] 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.

[12] G. D. Makarov, "Estimates for the distribution function of a sum of two random variables when the marginal distributions are fixed", *Theory of Probability and its Applications*, 1981, Vol. 26, pp. 803–806.

[13] R. E. Moore and W. A. Lodwick, "Interval Analysis and Fuzzy Set Theory", *Fuzzy Sets and Systems*, 2003, Vol. 135, No. 1, pp. 5–9.

[14] H. T. Nguyen and V. Kreinovich, "Nested Intervals and Sets: Concepts, Relations to Fuzzy Sets, and Applications", In: R. B. Kearfott et al. (eds.), *Applications of Interval Computations*, Kluwer, Dordrecht, 1996, pp. 245–290

[15] H. T. Nguyen and E. A. Walker, *First Course in Fuzzy Logic*, CRC Press, Boca Raton, FL, 1999.

[16] P. V. Novitskii and I. A. Zograph, *Estimating the Measurement Errors*, Energoatomizdat, Leningrad, 1991 (in Russian).

[17] A. I. Orlov, "How often are the observations normal?", *Industrial Laboratory*, 1991, Vol. 57. No. 7, pp. 770–772.

[18] C. H. Papadimitriou, *Computational Complexity*, Addison-Wesley, Reading, Massachusetts, 1994.

[19] S. Rabinovich, *Measurement Errors: Theory and Practice*, Amer. Inst. Phys., N.Y., 1993.

[20] H. Regan, S. Ferson, and D. Berleant, "Equivalence of five methods for bounding uncertainty", *Journal of Approximate Reasoning*, 2004, Vol. 36, No. 1, pp. 1–30.

[21] R. C. Williamson, "An Extreme limit theorem for dependency bounds of normalized sums of random variables", *Information Sciences*, 1991, Vol. 56, pp. 113–141.

[22] R. Williamson and T. Downs, "Probabilistic arithmetic I: numerical methods for calculating convolutions and dependency bounds", *International Journal of Approximate Reasoning*, 1990, Vol. 4, No. 2, pp. 89–158.