# Computing Statistical Characteristics When We Know Probabilities with Interval or Fuzzy Uncertainty: Computational Complexity

Gang Xiang
Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968, USA
gxiang@acm.org

Jim W. Hall
Tyndall Center for Climate Change Research
School of Civil Engineering and Geosciences
University of Newcastle upon Tyne, NE1 7RU, UK
Jim.Hall@newcastle.ac.uk

***Abstract*—In traditional statistics, we usually assume that we know the exact probability distributions. In practice, we often only know the probabilities with interval uncertainty.**

**The main emphasis on taking this uncertainty into account has been on situations in which we know a cumulative distribution function (cdf) with interval uncertainty. However, in some cases, we know the probability density function (pdf) with interval uncertainty. We show that in this situations, the exact range of some statistical characteristics can be efficiently computed. Surprisingly, for some other characteristics, similar statistical problems which are efficiently solvable for interval-valued cdf become computationally difficult (NP-hard) for interval-valued pdf.**

## I. Uncertainty in Probability

In the traditional statistics, we usually assume that we know the exact probability distributions.

In general, a probability distribution can be described by a cumulative probability distribution (cdf) $F(x) = \text{Prob}(t \leq x)$.

A continuous probability distribution can also be described by a probability density function (pdf) $\rho(x)$. A discrete distribution can be similarly described by the probabilities $p(x)$ of individual values; however, there are distributions which cannot be described in this way.

In practice, we usually know the probabilities only with some uncertainty. It is therefore desirable to take this uncertainty into account when we compute the values of the statistical characteristics.

## II. p-Boxes: The Most Computationally Developed Approach to Handling Uncertainty with Which We Know the Probabilities

Since the cdf corresponds to the most general case of a probability distribution, most algorithmic efforts in taking uncertainty into account have been directed towards the case when our knowledge about the probability distribution is represented by a cdf.

In the case of a cdf, uncertainty means that for every $x$, instead of the exact value of $F(x)$, we only know the interval of possible values $[\underline{F}(x), \overline{F}(x)]$.

Situation when we only know the cdf with interval uncertainty $[\underline{F}(x), \overline{F}(x)]$ is known as a *p-box*; see, e.g., [5]. For p-boxes, there are efficient algorithms that compute statistical characteristics such as ranges of moments, ranges of the cdf for the sum $x' + x''$ of two independent random variables in the situation when we know the p-boxes for $x'$ and $x''$, etc.

In some cases, we have *fuzzy* uncertainty, i.e., for every $x$, we have a fuzzy number corresponding to $F(x)$.

## III. From the Computational Viewpoint, It Is Sufficient to Consider Interval Uncertainty

In the fuzzy case, to describe the corresponding uncertainty, for each value $F$ of the probability $F(x)$, we describe the degree $\mu_x(F)$ to which this value is possible.

For each degree of certainty $\alpha$, we can determine the set of values of $F(x)$ that are possible with at least this degree of certainty – the $\alpha$-*cut* $\mathbf{F}_\alpha(x) \stackrel{\text{def}}{=} \{F \mid \mu_x(F) \geq \alpha\}$ of the original fuzzy set. In many practical cases, this $\alpha$-cut is an interval.

Vice versa, if we know $\alpha$-cuts for every $\alpha$, then, for each value $x$ and for each $F$, we can determine the degree of possibility that $F$ belongs to the original fuzzy set for $F(x)$ [4], [11]. A fuzzy set can be thus viewed as a nested family of its $\alpha$-cuts.

A *fuzzy number* can be defined as a fuzzy set for which all $\alpha$-cuts are intervals.

So, if instead of an interval $\mathbf{F}(x)$ of possible values of the probability $F(x)$, we have a fuzzy number $\mu_x(F)$ of possible values, then we can view this information as a family of nested intervals $\mathbf{F}_\alpha(x)$ ($\alpha$-cuts of the given fuzzy sets).

Our objective is then to compute the fuzzy number corresponding to the desired statistical characteristic (moment, pdf of the convolution, etc). In this case, for each level $\alpha$, the corresponding $\alpha$-cut of the desired fuzzy number can be computed based on the $\alpha$-cuts $\mathbf{F}_\alpha(x)$ of the corresponding input fuzzy sets. The resulting nested intervals form the fuzzy number for the desired statistical characteristic.

So, e.g., if we want to describe 10 different levels of uncertainty, then we must solve 10 interval computation problems.

Thus, from the computational viewpoint, it is sufficient to produce an efficient algorithm for the interval case.

## IV. Practical Situation: Bounds on Probabilities

In many practical situations, e.g., in climate modeling, we have bounds on the probability density or, in the discrete case, bounds on the probabilities of individual values; see, e.g., [7], [8]. How can we process this uncertainty?

## V. In Principle, We Can Use p-Boxes

One possible approach to dealing with bounds $[\underline{\rho}(x), \overline{\rho}(x)]$ on the (unknown) probability density $\rho(x)$ is to find corresponding range of the cdf $F(x)$, i.e., to find the (smallest possible) p-box which contains all probability distributions for which $\rho(x) \in [\underline{\rho}(x), \overline{\rho}(x)]$. Once we have this p-box, we can use known methods to estimate the ranges of different statistical characteristics.

## VI. Limitations of Using p-Boxes: General Description

The problem with this approach is that the p-box estimates are based on the assumption that all $F(x) \in [\underline{F}(x), \overline{F}(x)]$ are possible, while we are only interested in cumulative distribution functions $F(x)$ for which $\rho(x) = F'(x)$ is bounded by the given bounds $[\underline{\rho}(x), \overline{\rho}(x)]$. As a result, we often only get an *enclosure* for the desired range, an enclosure which has excess width.

## VII. Limitations of Using p-Boxes: Example

Let us describe a simple example where the use of p-boxes leads to excess width. Let us have a discrete random variable $x$ which can take 3 possible values 1, 2, and 3. We do not know the exact probabilities $p_1$, $p_2$, and $p_3$ of accessing these values; instead, we only know the intervals of possible values of these probabilities

$$\mathbf{p}_1 = \mathbf{p}_2 = \left[\frac{1}{3} - \beta, \frac{1}{3} + \beta\right]$$

for some small positive value $\beta \le \dfrac{1}{6}$. For $p_3$, we do not have separate interval bounds, only bounds which can be inferred from the bounds on $p_1$ and $p_2$ and the fact that $p_1+p_2+p_3 = 1$.

In this simple example, we are interested in the probability that $x = 2$. Of course, based on the known information, we can easily find the interval of possible value of this probability: it is

$$\mathbf{p}_2 = \left[\frac{1}{3} - \beta, \frac{1}{3} + \beta\right].$$

Let us show that if we convert to p-boxes, we will instead get an enclosure with excess width.

In this discrete situation, a cdf-style description means that we need to describe two numbers:

$$F(1) = p_1 = \mathrm{Prob}(x \le 1)$$

and

$$F(2) = \mathrm{Prob}(x \le 2) = p_1 + p_2.$$

These two values uniquely determine the resulting cdf:

- for $x < 1$, we have $F(x) = 0$;
- for $1 < x < 2$, we have $F(x) = F(1)$;
- for $2 < x < 3$, we have $F(x) = F(2)$;
- finally, for $x \ge 3$, we have $F(x) = 1$.

Based on the known intervals $\mathbf{x}_1$ and $\mathbf{p}_2$ of possible values of $p_1$ and $p_2$, we conclude that the resulting bound on $F(1)$ is $\mathbf{F}(1) = \mathbf{p}_1 = \left[\dfrac{1}{3} - \beta, \dfrac{1}{3} + \beta\right]$ and that the resulting bound on $F(2) = p_1 + p_2$ is

$$\mathbf{F}(2) = \mathbf{p}_1 + \mathbf{p}_2 = \left[\frac{1}{3} - \beta, \frac{1}{3} + \beta\right] + \left[\frac{1}{3} - \beta, \frac{1}{3} + \beta\right] = \left[\frac{2}{3} - 2\beta, \frac{2}{3} + 2\beta\right]$$

(see, e.g., [10]). Thus, the resulting p-box $\mathbf{F}(x) = [\underline{F}(x), \overline{F}(x)]$ has the following form:

- for $x < 1$, we have $\mathbf{F}(x) = [0, 0]$;
- for $1 \le x < 2$, we have
$$\mathbf{F}(x) = \left[\frac{1}{3} - \beta, \frac{1}{3} + \beta\right];$$
- for $2 \le x < 3$, we have
$$\mathbf{F}(x) = \left[\frac{2}{3} - 2\beta, \frac{2}{3} + 2\beta\right];$$
- finally, for $x \ge 3$, we have $\mathbf{F}(x) = [1, 1]$.

Based on this p-box information, the probability that $x = 2$ can be found as $F(2) - F(2-0)$, where $F(2-0) \stackrel{\text{def}}{=} \lim F(2-\delta)$, where $\delta > 0$ and $\delta \to 0$. From the p-box information, we conclude that $F(2)$ can take any values from the interval $\left[\dfrac{2}{3} - 2\beta, \dfrac{2}{3} + 2\beta\right]$ and that $F(2-0)$ can take any value from the interval $\left[\dfrac{1}{3} - \beta, \dfrac{1}{3} + \beta\right]$. According to interval computations [10], [**?**], [**?**]:

- the largest possible value of the difference
$$F(2) - F(2-0)$$
is when we subtract the smallest possible value of $F(2-0)$ from the largest possible value of $F(2)$, and
- the smallest possible value of the difference
$$F(2) - F(2-0)$$
is when we subtract the largest possible value of $F(2-0)$ from the smallest possible value of $F(2)$.

Thus, we conclude that the resulting interval of possible values of $\mathrm{Prob}(x = 2)$ is

$$\left[\frac{2}{3} - 2\beta, \frac{2}{3} + 2\beta\right] - \left[\frac{1}{3} - \beta, \frac{1}{3} + \beta\right] = \left[\frac{1}{3} - 3\beta, \frac{1}{3} + 3\beta\right].$$

This interval has the width of $6\beta$ – three times wider than the actual interval of possible values of $p_2$.

This example shows that if we only use p-boxes, we can get estimates with excess width.

## VIII. How Can We Compute Exact Ranges: Formulation of the Problem

It is desirable to compute the exact ranges for such characteristics as mean, central moments, convolution of several distributions (corresponding to the distribution of the sum of two independent variables), etc.

## IX. Computing Moments: Efficient Algorithms

**Formulation of the problem.** In the discrete case, we know the values $x_1 < x_2 < \ldots < x_n$, we know the bounds $[\underline{p}_i, \overline{p}_i]$ on the (unknown) actual probabilities $p_i$, and we are given an integer $m > 0$. Our objective is to find the range for $\sum\limits_{i=1}^{n} p_i \cdot x_i^m$ under the constraints $p_i \in [\underline{p}_i, \overline{p}_i]$ and $\sum\limits_{i=1}^{n} p_i = 1$.

**This problem is a particular case of a more general problem for which efficient algorithms are known.** To compute these ranges, we can use the fact that a linear-time algorithm (i.e., an algorithm with an $O(n)$ running time) is known for a more general problem. Namely, such an algorithm is known for a general case of computing the range $[\underline{a}, \overline{a}]$ of the expected value $a = \sum\limits_{i=1}^{n} p_i \cdot a_i$ of a known variable $(a_1, \ldots, a_n)$ under the constraints $p_i \in [\underline{p}_i, \overline{p}_i]$ and $\sum\limits_{i=1}^{n} p_i = 1$ [2], [9]. The case of moments correspond to $a_i = x_i^m$.

**Computing the upper endpoint $\overline{a}$: analysis.** Let us first consider the problem of computing the maximum $\overline{a}$. One can easily show that if for the maximizing vector $(p_1, \ldots, p_n)$, we have two values $p_i > \underline{p}_i$ and $p_j < \overline{p}_j$ for which $a_i < a_j$, then, by adding a small value $\Delta$ to $p_j$ and subtracting this value from $p_i$, we can get a new vector $p'_i$ for which still $\sum p'_i = 1$ but the value of $a = \sum\limits_{i=1}^{n} p_i \cdot a'_i$ is larger. Thus, for $a_i < a_j$, we cannot have $p_i > \underline{p}_i$ and $p_j < \overline{p}_j$ in the maximizing vector. So, we conclude that we can only have one value $k$ for which $p_k \in (\underline{p}_k, \overline{p}_k)$.

- For all values $a_i$ for which $a_i < a_k$, we have $p_i = \underline{p}_i$.
- For all values $a_j$ for which $a_j > a_k$, we have $p_j = \overline{p}_j$.

So, if we sort the values $a_i$ in increasing order, then we conclude that for some $k$, the maximum is attained for the vector $(\underline{p}_1, \ldots, \underline{p}_{k-1}, p_i, \overline{p}_{k+1}, \ldots, \overline{p}_n)$, where $p_k$ can be determined, from the condition that $\sum\limits_{i=1}^{n} p_i = 1$, as

$$p_k = 1 - \underline{p}_1 - \ldots - \underline{p}_{k-1} - \overline{p}_{k+1} - \ldots - \overline{p}_n.$$

The condition that $\underline{p}_k \le p_k \le \overline{p}_k$ leads to

$$\sum_{i=1}^{k-1} \underline{p}_i + \sum_{j=k}^{n} \overline{p}_j \le 1 \le \sum_{i=1}^{k} \underline{p}_i + \sum_{j=k+1}^{n} \overline{p}_j.$$

This condition uniquely determines the desired value $k$.

The above analysis leads to the following algorithm for computing $\overline{a}$.

**Computing the upper endpoint $\overline{a}$: first algorithm.**

- First, we sort the values $a_i$ in increasing order; sorting can be done in time $O(n \cdot \log(n))$; see, e.g., [3].
- Next, we compute the sums corresponding to $k = 0$; this computation takes linear time.
- Then, for each $k$, we need to change two terms to compute the new sums, so we need linear time for check all possible values of $k$ and find the right one.
- After this, we can compute the maximizing vector $p_i$ and the resulting upper endpoint $\overline{a} = \sum\limits_{i=1}^{n} p_i \cdot a_i$ in linear time.

In general, this algorithm requires $O(n \cdot \log(n)) + O(n) = O(n \cdot \log(n))$ time.

*Comment.* If the values $a_i$ are already sorted, then we only need linear time to compute $\overline{a}$. It turns out that we can have a linear-time algorithm in the general case, when the values $a_i$ are not pre-sorted.

**Computing the upper endpoint $\overline{a}$: linear-time algorithm.** This algorithm is based on the known fact that we can compute the median of a set of $n$ elements in linear time (see, e.g., [3]).

The algorithm is iterative. At each iteration of this algorithm we have three sets:

- the set $I^-$ of all the indices $i$ from 1 to $n$ for which we already know that for the maximizing vector $p$, we have $p_i = \underline{p}_i$;
- the set $I^+$ of all the indices $j$ for which we already know that for the maximizing vector $p$, we have $p_j = \overline{p}_j$;
- the set $I = \{1, \ldots, n\} \setminus (I^- \cup I^+)$ of the indices $i$ for which we are still undecided.

In the beginning, $I^- = I^+ = \emptyset$ and $I = \{1, \ldots, n\}$. At each iteration we also update the values of two auxiliary quantities $E^- \stackrel{\text{def}}{=} \sum\limits_{i \in I^-} \underline{p}_i$ and $E^+ \stackrel{\text{def}}{=} \sum\limits_{j \in I^+} \overline{p}_j$. In principle, we could compute these values by computing these sums. However, to speed up computations on each iteration, we update these two auxiliary values in a way that is faster than re-computing the corresponding two sums. Initially, since $I^- = I^+ = \emptyset$, we take $E^- = E^+ = 0$.

At each iteration we do the following:

- first, we compute the median $m$ of the set $I$ (median in terms of sorting by $a_i$);
- then, by analyzing the elements of the undecided set $I$ one by one, we divide them into two subsets $P^- = \{i : a_i \le a_m\}$ and $P^+ = \{j : a_j > a_m\}$;
- we compute $e^- = E^- + \sum\limits_{i \in P^-} \underline{p}_i$ and $e^+ = E^+ + \sum\limits_{j \in P^+} \overline{p}_j$;
- If $e^- + e^+ > 1$, then we replace $I^-$ with $I^- \cup P^-$, $E^-$ with $e^-$, and $I$ with $P^+$.
- If $e^- + e^+ + 2\Delta_m < 1$, then we replace $I^+$ with $I^+ \cup P^+$, $E^+$ with $e^+$, and $I$ with $P^-$.
- Finally, if $e^- + e^+ \le 1 \le e^- + e^+ + 2\Delta_m$, then we replace $I^-$ with $I^- \cup (P^- - \{m\})$, $I^+$ with $I^+ \cup P^+$, $I$ with $\{m\}$, $E^-$ with $e^- - \underline{p}_m$, and $E^+$ with $e^+$.

At each iteration the set of undecided indices is divided in half. Iterations continue until we have only one undecided index $I = \{k\}$. After this we return, as $\overline{a}$, the value of the linear combination $\sum\limits_{i=1}^{n} p_i \cdot a_i$ for the vector $p$ for which $p_i = \underline{p}_i$ for $i \in I^-$, $x_j = \overline{p}_j$ for $j \in I^+$, and $p_k = 1 - e^- - e^+$ for the remaining value $k$.

**Proof that the second algorithm for computing $\overline{a}$ requires linear time.** At each iteration, computing median requires linear time, and all other operations with $I$ require time $t$ linear in the number of elements $|I|$ of $I$: $t \le C \cdot |I|$ for some $C$. We start with the set $I$ of size $n$. On the next iteration, we have a set of size $n/2$, then $n/4$, etc. Thus, the overall computation time is $\le C \cdot (n + n/2 + n/4 + \ldots) \le C \cdot 2n$, i.e. linear in $n$.

**How to compute $\underline{a}$.** It is known that the smallest possible value $\underline{a}$ of the linear form $\sum\limits_{i=1}^{n} p_i \cdot a_i$ under given constraints is equal to $-\overline{b}$, where $\overline{b}$ is the largest possible value of the form $\sum\limits_{i=1}^{n} p_i \cdot b_i$, with $b_i = -a_i$. Thus, by using the above algorithm, we can compute the lower endpoint as well.

## X. Computing Convolution: A Practically Important Problem

If we know the distributions $\rho'(x)$ and $\rho''(x)$ of two independent random variables $x'$ and $x''$, then the probability density function $\rho(x)$ for their sum $x = x' + x''$ is described by the convolution $\rho(x) = \int \rho'(z) \cdot \rho''(x - z)\, dz$.

## XI. There Exist Efficient Algorithms for Computing Convolution of p-Boxes

Researchers have analyzed the problem of computing the convolution in situations when instead of knowing the exact cumulative distribution functions $F'(x)$ and $F''(x)$, we only know p-boxes $[\underline{F}'(x), \overline{F}'(x)]$ and $[\underline{F}''(x), \overline{F}''(x)]$.

In these situations, we can efficiently compute a p-box for $x = x' + x''$; see, e.g., [5]. This possibility comes from the fact that for every $x$, the value $F(x)$ corresponding to the convolution is a (non-strictly) increasing function of the values $F'(x')$ and $F''(x'')$. Thus:

- to compute the lower endpoint $\underline{F}(x)$ for the resulting cdf $F(x)$, it is sufficient to compute the convolution of the distributions corresponding to $\underline{F}'(x)$ and $\underline{F}''(x)$;
- similarly, to compute the upper endpoint $\overline{F}(x)$ for the resulting cdf $F(x)$, it is sufficient to compute the convolution of the distributions corresponding to $\overline{F}'(x)$ and $\overline{F}''(x)$.

## XII. Convolution of Interval-Valued Probabilities: General Case

In line with our previous discussions, let us now consider the situation in which, instead of the p-boxes (i.e., bounds on the cumulative distribution functions), we know the interval bounds $[\underline{\rho}'(x), \overline{\rho}'(x)]$ and $[\underline{\rho}''(x), \overline{\rho}''(x)]$ of the corresponding probability distribution functions.

In this case, for every $x$, we would like to compute the exact range $[\underline{\rho}(x), \overline{\rho}(x)]$ of the convolution

$$\rho(x) = \int \rho'(z) \cdot \rho''(x - z)\, dz$$

when $\rho'(x) \in [\underline{\rho}'(x), \overline{\rho}'(x)]$ and $\rho''(x) \in [\underline{\rho}''(x), \overline{\rho}''(x)]$.

Let us prove that this problem is computationally difficult even in the discrete case, when each of the two variables $x'$ and $x''$ can only takes finitely many values.

## XIII. Convolution of Interval-Valued Probabilities: Discrete Case

Let us assume that we have two independent discrete random variables $x'$ and $x''$.

- For the variable $x'$, we know its possible values $x'_1, \ldots, x'_{n'}$ and the bounds $\left[\underline{p}'_i, \overline{p}'_i\right]$ on the corresponding (unknown) probabilities $p'_i$.
- Similarly, for the variable $x''$, we know its possible values $x''_1, \ldots, x''_{n''}$ and the bounds $\left[\underline{p}''_j, \overline{p}''_j\right]$ on the corresponding (unknown) probabilities $p''_j$.

For the sum $x = x' + x''$, we have possible values $x_{ij} = x'_i + x''_j$.

The question is to find the ranges $\underline{p}_{ij}$ and $\overline{p}_{ij}$ of possible values of the corresponding probabilities

$$p_{ij} = \sum_{i',j' : x'_{i'} + x''_{j'} = x'_i + x''_j} p'_{i'} \cdot p''_{j'},$$

where the values $p'_i$ and $p''_i$ satisfy the conditions

$$p'_i \in \left[\underline{p}'_i, \overline{p}'_i\right], \quad p''_i \in \left[\underline{p}''_i, \overline{p}''_i\right],$$

$$\sum_{i=1}^{n'} p'_i = 1, \quad \sum_{i=1}^{n''} p''_i = 1.$$

## XIV. New result: For Interval-Valued Probabilities, Computing Convolution Is NP-hard

In this paper, we prove, that, in contrast to the case of p-boxes, computing the (endpoints of) the exact range $\left[\underline{p}_{ij}, \overline{p}_{ij}\right]$ of the convolution probabilities $p_{ij}$ is computationally difficult (namely, NP-hard).

To be more precise, we prove two results:

- that the problem of computing the upper endpoint $\overline{p}_{ij}$ of the convolution if NP-hard, and
- that the problem of computing the lower endpoint $\underline{p}_{ij}$ is also NP-hard.

*Comment.* Our proof is somewhat similar to the proof of NP-hardness from [1].

## XV. PROOF OF NP-HARDNESS

$1^\circ$. Our proof is based on reducing, to this problem, a known NP-hard *subset* problem, where we are given $n$ positive integers $s_1, \ldots, s_n$, and we must find the values $\varepsilon_i \in \{-1, 1\}$ for which $\sum\limits_{i=1}^{n} \varepsilon_i \cdot s_i = 0$.

For precise definitions of NP-hardness, see, e.g., [6].

$2^\circ$. To each instance $s_1, \ldots, s_n$ of the subset problem, we assign the following two interval-valued probability distributions $x'$ and $x''$.

$2.1^\circ$. The variable $x'$ can only take $n' = n$ values $x'_1 = 1, \ldots, x'_i = i, \ldots, x'_n = n$. For each $i$ from 1 to $n$, the corresponding probability $p'_i$ can take any value from the interval

$$\left[\underline{p}'_i, \overline{p}'_i\right] = \left[\frac{1}{n} - \beta \cdot s_i, \frac{1}{n} + \beta \cdot s_i\right].$$

$2.2^\circ$. Similarly, the variable $x''$ can only take $n'' = n$ values $x''_1 = -1, \ldots, x''_i = -i, \ldots, x''_n = -n$. For each $i$ from 1 to $n$, the corresponding probability $p''_i$ can take any value from the interval

$$\left[\underline{p}''_i, \overline{p}''_i\right] = \left[\frac{1}{n} - \beta \cdot s_i, \frac{1}{n} + \beta \cdot s_i\right].$$

$3^\circ$. The value $\beta$ should be selected in such a way as to guarantee that the resulting probabilities are always non-negative, i.e., that $\dfrac{1}{n} - \beta \cdot s_i \geq 0$ for all $i$. This requirement is equivalent to $\beta \cdot s_i \leq \dfrac{1}{n}$, i.e., to $\beta \leq \dfrac{1}{n \cdot s_i}$. This must hold for all $i$, so we must make sure that $\beta$ does not exceed the smallest of these values – i.e., the value corresponding to the largest $s_i$. Thus, we can take

$$\beta = \frac{1}{n \cdot \max\limits_i s_i}.$$

$4.1^\circ$. In this case, for every $i$, the (unknown) actual probability $p'_i$ can be described as

$$p'_i = \frac{1}{n} + \beta \cdot \Delta'_i,$$

where $\Delta'_i \stackrel{\text{def}}{=} p'_i - \dfrac{1}{n}$ can take any value from the interval $[-s_i, s_i]$.

$4.2^\circ$. Similarly, for every $i$, the (unknown) actual probability $p''_i$ can be described as

$$p''_i = \frac{1}{n} + \beta \cdot \Delta''_i,$$

where $\Delta''_i \stackrel{\text{def}}{=} p''_i - \dfrac{1}{n}$ can take any value from the interval $[-s_i, s_i]$.

$5.1^\circ$. In terms of the auxiliary variables $\Delta'_i$, the requirement that $\sum\limits_{i=1}^{n} p'_i = 1$ means that

$$\sum_{i=1}^{n} \left(\frac{1}{n} + \beta \cdot \Delta'_i\right) = 1,$$

i.e., that $1 + \sum\limits_{i=1}^{n} \Delta'_i = 1$ and $\sum\limits_{i=1}^{n} \Delta'_i = 0$.

$5.1^\circ$. Similarly, the requirement that $\sum\limits_{i=1}^{n} p''_i = 1$ means that

$$\sum_{i=1}^{n} \left(\frac{1}{n} + \beta \cdot \Delta''_i\right) = 1,$$

i.e., that $1 + \sum\limits_{i=1}^{n} \Delta''_i = 1$ and $\sum\limits_{i=1}^{n} \Delta''_i = 0$.

$6^\circ$. Let us now find the range of possible values for the probability that the sum $x = x' + x''$ is equal to 0.

The value 0 can be obtained if $x' = i$ and $x'' = -i$ for the same value $i$. Thus, the desired probability is equal to

$$p_{11} = \sum_{i=1}^{n} p'_i \cdot p''_i.$$

Substituting the expressions $p'_i = \dfrac{1}{n} + \beta \cdot \Delta'_i$ and $p''_i = \dfrac{1}{n} + \beta \cdot \Delta''_i$ into this formula, we get

$$p_{11} = \sum_{i=1}^{n} \left(\frac{1}{n} + \beta \cdot \Delta'_i\right) \cdot \left(\frac{1}{n} + \beta \cdot \Delta''_i\right).$$

Here,

$$\left(\frac{1}{n} + \beta \cdot \Delta'_i\right) \cdot \left(\frac{1}{n} + \beta \cdot \Delta''_i\right) =$$

$$\left(\frac{1}{n}\right)^2 + \frac{1}{n} \cdot \beta \cdot \Delta'_i + \frac{1}{n} \cdot \beta \cdot \Delta''_i + \beta^2 \cdot \Delta'_i \cdot \Delta''_i.$$

Therefore, we conclude that

$$p_{11} = \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 + \sum_{i=1}^{n} \frac{1}{n} \cdot \beta \cdot \Delta'_i + \sum_{i=1}^{n} \frac{1}{n} \cdot \beta \cdot \Delta''_i +$$

$$\sum_{i=1}^{n} \beta^2 \cdot \Delta'_i \cdot \Delta''_i.$$

By moving constant factors outside the sum, we get:

$$p_{11} = \left(\frac{1}{n}\right)^2 \cdot \sum_{i=1}^{n} 1 + \frac{1}{n} \cdot \beta \cdot \sum_{i=1}^{n} \Delta'_i + \frac{1}{n} \cdot \beta \cdot \sum_{i=1}^{n} \Delta''_i +$$

$$\beta^2 \cdot \sum_{i=1}^{n} \Delta'_i \cdot \Delta''_i.$$

The first sum is equal to $\left(\frac{1}{n^2}\right)\cdot n = \frac{1}{n}$. The second and the third sums are equal to 0 since $\sum\limits_{i=1}^{n} \Delta'_i = 0$ and $\sum\limits_{i=1}^{n} \Delta''_i = 0$. Thus, we conclude that

$$p_{11} = \frac{1}{n} + \beta^2 \cdot \sum_{i=1}^{n} \Delta'_i \cdot \Delta''_i.$$

7°. Let us prove that the number $\frac{1}{n} + \beta^2 \cdot \sum\limits_{i=1}^{n} s_i^2$ is a possible value of $p_{11}$ if and only if the original instance of a subset problem has a solution.

This will prove that the problem of computing the upper endpoint $\overline{p}_{11}$ of the range of $p_{11}$ is NP-hard.

7.1°. Indeed, if the original instance has a solution $\varepsilon$ for which $\sum\limits_{i=1}^{n} \varepsilon_i \cdot s_i = 0$, then we can take $\Delta'_i = \Delta''_i = \varepsilon_i \cdot s_i$ and get

$$p_{11} = \frac{1}{n} + \beta^2 \cdot \sum_{i=1}^{n} s_i^2.$$

7.2°. Vice versa, let us assume that the number $\frac{1}{n} + \beta^2 \cdot \sum\limits_{i=1}^{n} s_i^2$ is a possible value of $p_{11}$. Let us prove that in this case, the original instance of the subset problem has a solution.

Indeed, since $|\Delta'_i| \le s_i$ and $|\Delta''_i| \le s_i$, we always have $|\Delta'_i \cdot \Delta''_i| \le s_i^2$ and hence $\Delta'_i \cdot \Delta''_i \le s_i^2$.

So, the only possibility to have

$$p_{11} = \frac{1}{n} + \beta^2 \cdot \sum_{i=1}^{n} \Delta'_i \cdot \Delta''_i = \frac{1}{n} + \beta^2 \cdot \sum_{i=1}^{n} s_i^2$$

is to have $\Delta'_i \cdot \Delta''_i = s_i^2$ for all $i$ – otherwise, we would have

$$p_{11} = \frac{1}{n} + \beta^2 \cdot \sum_{i=1}^{n} \Delta'_i \cdot \Delta''_i < \frac{1}{n} + \beta^2 \cdot \sum_{i=1}^{n} s_i^2.$$

If $|\Delta'_i| < s_i$ or $|\Delta''_i| < s_i$, then we have $|\Delta'_i \cdot \Delta''_i| < s_i^2$ and hence $\Delta'_i \cdot \Delta''_i < s_i^2$. So, the only way to have $\Delta'_i \cdot \Delta''_i = s_i^2$ is to have $|\Delta'_i| = s_i$ and $|\Delta''_i| = s_i$. So, we have $\Delta'_i = \pm s_i$, i.e., $\Delta'_i = \varepsilon_i \cdot s_i$ for some value $\varepsilon_i \in \{-1, 1\}$.

The fact that $\sum\limits_{i=1}^{n} \Delta'_i = 0$ implies that $\sum\limits_{i=1}^{n} \varepsilon_i \cdot s_i = 0$. So, the values $\varepsilon_i$ form a solution to the original instance of the subset problem.

7.3°. The reduction is proven, and so the problem of computing the upper endpoint $\overline{p}_{ij}$ of the convolution is indeed NP-hard in case of interval uncertainty.

8°. Let us prove that the number $\frac{1}{n} - \beta^2 \cdot \sum\limits_{i=1}^{n} s_i^2$ is a possible value of $p_{11}$ if and only if the original instance of a subset problem has a solution.

This will prove that the problem of computing the lower endpoint $\underline{p}_{11}$ of the range of $p_{11}$ is also NP-hard.

8.1°. Indeed, if the original instance has a solution $\varepsilon$ for which

$$\sum_{i=1}^{n} \varepsilon_i \cdot s_i = 0,$$

then we can take $\Delta'_i = \varepsilon_i \cdot s_i$ and $\Delta''_i = -\varepsilon_i \cdot s_i$, and get

$$p_{11} = \frac{1}{n} - \beta^2 \cdot \sum_{i=1}^{n} s_i^2.$$

8.2°. Vice versa, let us assume that the number $\frac{1}{n} - \beta^2 \cdot \sum\limits_{i=1}^{n} s_i^2$ is a possible value of $p_{11}$. Let us prove that in this case, the original instance of the subset problem has a solution.

Indeed, since $|\Delta'_i| \le s_i$ and $|\Delta''_i| \le s_i$, we always have $|\Delta'_i \cdot \Delta''_i| \le s_i^2$ and hence $\Delta'_i \cdot \Delta''_i \ge -s_i^2$.

So, the only possibility to have

$$p_{11} = \frac{1}{n} + \beta^2 \cdot \sum_{i=1}^{n} \Delta'_i \cdot \Delta''_i = \frac{1}{n} - \beta^2 \cdot \sum_{i=1}^{n} s_i^2$$

is to have $\Delta'_i \cdot \Delta''_i = -s_i^2$ for all $i$ – otherwise, we would have

$$p_{11} = \frac{1}{n} + \beta^2 \cdot \sum_{i=1}^{n} \Delta'_i \cdot \Delta''_i > \frac{1}{n} + \beta^2 \cdot \sum_{i=1}^{n} s_i^2.$$

If $|\Delta'_i| < s_i$ or $|\Delta''_i| < s_i$, then we have $|\Delta'_i \cdot \Delta''_i| < s_i^2$ and hence $\Delta'_i \cdot \Delta''_i > -s_i^2$. So, the only way to have $\Delta'_i \cdot \Delta''_i = -s_i^2$ is to have $|\Delta'_i| = s_i$ and $|\Delta''_i| = s_i$. So, we have $\Delta'_i = \pm s_i$, i.e., $\Delta'_i = \varepsilon_i \cdot s_i$ for some value $\varepsilon_i \in \{-1, 1\}$.

The fact that $\sum\limits_{i=1}^{n} \Delta'_i = 0$ implies that $\sum\limits_{i=1}^{n} \varepsilon_i \cdot s_i = 0$. So, the values $\varepsilon_i$ form a solution to the original instance of the subset problem.

## ACKNOWLEDGMENTS

This work was supported in part by the Texas Department of Transportation grant No. 0-5453. The authors are thankful to the anonymous referees to valuable suggestions.

## REFERENCES


[1] R. Araiza, G. Xiang, O. Kosheleva, and D. Škulj, "Under Interval and Fuzzy Uncertainty, Symmetric Markov Chains Are More Difficult to Predict", *These Proceedings*.

[2] P. van der Broek and J. Noppen, "Fuzzy weighted average: alternative approach", *Proc. NAFIPS'06*.

[3] Th. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT Press, Cambridge, MA, 2001.

[4] D. Dubois and H. Prade, "Operations on fuzzy numbers", *International Journal of Systems Science*, 1978, Vol. 9, pp. 613–626.

[5] S. Ferson. *RAMAS Risk Calc 4.0*. CRC Press, Boca Raton, Florida, 2002.

[6] M. R. Garey and D. S. Johnson. *Computers and Intractability*, Freeman, San Francisco, California, 1979.

[7] J. W. Hall, "Soft methods in Earth science engineering", In: J. Lawry et al. (eds.), *Soft Methods for Integrated Uncertainty Modeling*, Springer-Verlag, 2006, pp. 7–10.

[8] J. W. Hall, G. Fu, and J. Lawry, "Imprecise probabilities of climate change: aggregation of fuzzy scenarios and model uncertainties", *Climatic Change*, 2007, Vol. 81, No. 3-4, pp. 265–281.

[9] P. Hansen, M. V. P. de Aragao, and C. C. Ribeiro, "Hyperbolic 0-1 programming and optimization in information retrieval", *Math. Programming*, 1991, Vol. 52, pp. 255–263.

[10] L. Jaulin, M. Kieffer, O. Didrit, and E. Walter, *Applied Interval Analysis*, Springer, London, 2001.

[11] H. T. Nguyen and E. A. Walker, *A First Course in Fuzzy Logic*, CRC Press, Boca Raton, Florida, 2006.