Towards Discrete Interval, Set, and Fuzzy Computations

Enrique Portillo, Olga Kosheleva, and Vladik Kreinovich

University of Texas at El Paso
El Paso, Texas 79968, USA
eportillo2@miners.utep.edu, olgak@utep.edu
vladik@utep.edu

1. Need for Data Processing

- We want to understand the world, to know the values of different quantities characterizing the world.
- Some quantities we can directly measure: e.g., we can easily measure the temperature in El Paso.
- Often, we are interested in the value of some quantity y which is difficult to measure directly, e.g.:
  - temperature inside a star,
  - where a close-to-Earth asteroid will be in 20 years.
- Such quantities can only be measured *indirectly*:
  - find easier-to-measure measure quantities $x_1, \ldots, x_n$ related to y by a known relation $y = f(x_1, \ldots, x_n)$;
  - measure x_i and *process* the measurement results $\widetilde{x}_i$, generating $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.

2. Example of Data Processing

- The goals of geosciences are:
  - to better predict earthquakes,
  - to better find minerals, etc.
- For that, we need to know the density at different depths and locations.
- This density is difficult (or even impossible) to measure directly.
- To indirectly measure the density, we can:
  - measure the gravity field at different locations and at different heights, and then
  - use the known algorithms to reconstruct the desired density values.

3. Need to Take Uncertainty into Account

- The result $\widetilde{x}_i$ of measuring (or estimating) a quantity x_i is, in general, different from its actual value:

$$\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i \neq 0.$$

- Since $\widetilde{x}_i \neq x_i$, the result $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ of data processing is different from the actual (unknown) value y:

$$y = f(x_1, \ldots, x_n) \neq f(\widetilde{x}_1, \ldots, \widetilde{x}_n) = \widetilde{y}.$$

- To make decisions based on the estimate $\widetilde{y}$, we need to know the accuracy $\Delta y \stackrel{\text{def}}{=} \widetilde{y} - y$ of this estimate.
- If a geophysical analysis has shown that a natural gas field contains $\widetilde{y} = 10$ trillion cubic feet of gas:
  - if it is 10 ± 1, we should start production;
  - if it is 10 ± 20, there may be no gas at all, so a further analysis is needed.

4. Traditional Probabilistic Approach to Uncertainty

- In the traditional probabilistic approach, we:
  - estimate the probabilities of different values of Δx_i (it is often Gaussian),
  - find correlations (if any) between the corresponding random variables Δx_i, and then
  - use known statistical methods to derive the resulting probability distribution for Δy.
- A usual way of finding the probability distribution for $\Delta x_i = \widetilde{x}_i - x_i$ is to repeatedly compare:
  - the results $\widetilde{x}_i$ obtained by our measuring instrument and
  - the results $\widetilde{x}_i^{\rm st} \approx x_i$ obtained by a much more accurate ("standard") measuring instrument.

5. Need to Go Beyond the Traditional Probabilistic Approach

- The first case if when we have state-of-the-art measurements.
- For example, a geophysicist uses the state-of-the-art instrument for measuring gravity.
- It would be great to also have a more accurate "standard" instrument, but this is the best we have.
- Another case is when we use the measurements as a part of manufacturing process.
- In such situations, in principle, we can calibrate each sensor, but calibration is very expensive.
- Expensive calibration may be necessary when manufacturing a nuclear reactor, but not for toothbrushes.

6. Case of Interval Uncertainty

- Often, we do not know the probabilities of different values of measurement error Δx_i.
- In such situations, we should know at least an upper bound Δ_i on this error.
- Indeed, if we do not even known any upper bound, then this is not a measurement, this is a wild guess.
- For measurements, a manufacturer provides an upper bound on the meas. error, i.e., Δ_i s.t. $|\Delta x_i| \le \Delta_i$.
- In this case, based on the measurement result $\widetilde{x}_i$, we can conclude that the actual value x_i is in the interval

$$\mathbf{x}_i \stackrel{\text{def}}{=} [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i].$$

- For example, when the measured value is $\widetilde{x}_i = 1.0$, the actual value x_i is in $[1.0 - 0.1, 1.0 + 0.1] = [0.9, 1.1]$.

7. Need for Interval Computations

- When we know all the inputs with interval uncertainty, then:
  - for each input x_i,
  - we only know the interval $\mathbf{x}_i$ of possible values of x_i.
- Different combinations of values $x_i \in \mathbf{x}_i$ lead to different values $y = f(x_1, \ldots, x_n)$ of the desired quantity y.
- In such situations, it is desirable to find the set of all possible values y, i.e., the set

$$Y \stackrel{\text{def}}{=} \{f(x_1, \ldots, x_n) : x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\}.$$

- The problem of estimating Y based on known intervals $\mathbf{x}_i$ is known as a problem of *interval computations*.

8. Need for Fuzzy Uncertainty and Fuzzy Computations

- In many practical cases, instead of measuring the values x_i, we ask experts to estimate these values.
- Experts often use imprecise ("fuzzy") words like "small".
- Fuzzy logic is a technique designed to translate such expert statements into computer-understandable form.
- To each word S (like "small") and to each value x, we assign a degree $\mu_S(x)$ to which x is S.
- The resulting *membership function* can be obtained, e.g., by polling experts.
- Thus, we get membership functions $\mu_i(x_i)$ corresponding to different inputs.
- We need to compute the membership function $\mu(y)$ corresponding to $y = f(x_1, \ldots, x_n)$.

9. How to Perform Fuzzy Computations: Towards Zadeh's Extension Principle

- y is a possible value of the desired variable if for some real numbers $x_1, \ldots, x_n$ for which $y = f(x_1, \ldots, x_n)$:
  - the value x_1 is a possible value of the 1st input, ...
  - the value x_n is a possible value of the n-th input.
- In other words, we are interested in the degree to which the following statement holds:

$$\bigvee_{x_i : f(x_1, \ldots, x_n) = y} (x_1 \text{ is possible} \,\&\, \ldots \,\&\, x_n \text{ is possible}).$$

- The degree to which x_i is possible is equal to $\mu_i(x_i)$.
- If we use min for "and" and max for "or", we get

$$\mu(y) = \max_{x_1, \ldots, x_n : f(x_1, \ldots, x_n)} \min(\mu_1(x_1), \ldots, \mu_n(x_n)).$$

10. From the Computational Viewpoint, Fuzzy Computations Can Be Reduced to Interval and Set Computations

- We have $\mu(y) = \max\limits_{x_1,\ldots,x_n: f(x_1,\ldots,x_n)} \min(\mu_1(x_1),\ldots,\mu_n(x_n))$.
- This formula can be rewritten in terms of α-cuts $\mathbf{x}_i(\alpha) \stackrel{\text{def}}{=} \{x_i : \mu_i(x_i) \geq \alpha\}$ and $\mathbf{y}(\alpha) = \{y : \mu(y) \geq \alpha\}$:

$$\mathbf{y}(\alpha) = \{f(x_1,\ldots,x_n) : x_1 \in \mathbf{x}_1(\alpha),\ldots,\mathbf{x}_n(\alpha)\}.$$

- So, by applying interval (or set) computations, we can find the sets $\mathbf{y}(\alpha)$.
- Once we know the sets $\mathbf{y}(\alpha)$, we can reconstruct each value $\mu(y)$ as $\max\{\alpha : y \in \mathbf{y}(\alpha)\}$.
- In other words, fuzzy computations can be reduced to the interval (set) computations for α-cuts.

11. Need for Discrete Computations

- Usually, in interval and fuzzy computations, we assume that the variables are *continuous*.
- In other words, we assume that each variable can take all real values from the corresponding range.
- In practice, sometimes, variables are *discrete*, e.g., x_i can be number of people.
- It is reasonable to develop efficient algorithms for computing the ranges in such discrete cases.
- We have shown that fuzzy computations can be reduced to interval (set) computations.
- Because of this, we will only describe algorithms for the interval (set) case.

12. How Difficult Are the Usual (Continuous) Interval Computation Problems

- For a linear function $f(x_1, \ldots, x_n) = a_0 + \sum\limits_{i=1}^{n} a_i \cdot x_i$, its range $\mathbf{y}$ over intervals $\mathbf{x}_i = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$ is

$$\mathbf{y} = [\widetilde{y} - \Delta, \widetilde{y} + \Delta], \text{ where } \widetilde{y} = a_0 + \sum_{i=1}^{n} a_i \cdot \widetilde{x}_i, \Delta = \sum_{i=1}^{n} |a_i| \cdot \Delta_i.$$

- For a quadratic function computing the range over given intervals is, in general, an NP-hard problem.
- Moreover, it is NP-hard even when we restrict ourselves to such a simple quadratic function as variance

$$V = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n} \cdot \sum_{i=1}^{n} x_i \right)^2.$$

- Crudely speaking, this means that no feasible algorithm is possible for solving all cases of this problem.

13. Discrete Case: Formulation of the Problem

- In the discrete case, each variable x_i can take only finitely many values $x_{i1}, \ldots, x_{in_i}$.
- In practice, these values are usually equally distributed, i.e., $x_i = c_i + x'_i \cdot h_i$ for integer x'_i.
- W.l.o.g., we can thus assume that each variable x_i takes integer values between some bounds $\underline{X}_i$ and $\overline{X}_i$.
- In this case, the range estimation problem takes the following form:
  - for each input i, we know the bounds $\underline{X}_i$ and $\overline{X}_i$ on x_i;
  - we also know a function $f(x_1, \ldots, x_n)$;
  - our objective is to find the range

$$Y = \{f(x_1, \ldots, x_n) : x_i = \underline{X}_i, \underline{X}_i + 1, \ldots, \overline{X}_i\}.$$

14. Discrete Case: The Problem Becomes NP-Hard Even for Linear Functions

- This directly follows from the known fact that the following *subset sum* problem is NP-hard:
  - given integers $a_1, \ldots, a_n$, and a,
  - check whether there exist values $x_i \in \{0,1\}$ for which $\sum\limits_{i=1}^{n} a_i \cdot x_i = a$.
- Thus, for $x_i \in \{0,1\}$, it is NP-hard to check whether the range of $\sum\limits_{i=1}^{n} a_i \cdot x_i$ contains a given integer a.
- In practice, the ranges of a_i are usually bounded.
- We show that in this case, feasible algorithms are possible.
- We also show that feasible algorithms are possible for variance.

15. Main Assumption and First Idea

- We assume that all the values are bounded by some constant A: $|a_i| \le A$.
- It is also reasonable to assume that the possible values of x_i are bounded by some constant X: $|x_i| \le X$.
- There are finitely many values of each input x_i.
- So, in principle, we can enumerate all combinations $(x_1, \ldots, x_n)$; however:
  - if we take at least two different values of each of n variables x_i,
  - we will thus need to consider at least 2^n different combinations $(x_1, \ldots, x_n)$.
- Thus, the above straightforward algorithm requires exponential time.

16. Linear Case: Polynomial-Time Algorithm $A_{\rm lin}$

- We want to compute $Y = \left\{\sum\limits_{i=1}^n a_i\cdot x_i : x_i = \underline{X}_i, \ldots, \overline{X}_i\right\}$.
- Let us sequentially compute the ranges

$$Y_k \stackrel{\rm def}{=} \left\{\sum_{i=1}^k a_i\cdot x_i : x_i = \underline{X}_i, \underline{X}_i+1, \ldots, \overline{X}_i\right\},\quad k=0,1,\ldots,n.$$

- Here, $Y_0=\{0\}$; once we know Y_k, we compute Y_{k+1} as

$$Y_{k+1} = \{y_k + a_{k+1}\cdot x_{k+1} : y_k\in Y_k \;\&\; x_{k+1} = \underline{X}_{k+1}, \ldots, \overline{X}_{k+1}\}.$$

- From $|a_i|\le A$, $|x_i|\le X$, we get $\left|\sum\limits_{i=1}^k a_i\cdot x_i\right| \le n\cdot A\cdot X$.
- On each of n stages, we thus need $O(n)$ computational steps, to the total of $n\cdot O(n) = O(n^2)$.
- $A_{\rm lin}$ computes the *exact* range in polynomial time; so, we gain computation time without sacrificing accuracy.

17. Computing the Range of Variance in Polynomial Time

- We sequentially compute the intermediate sets of pairs

$$P_k = \left\{ \left(\sum_{i=1}^{k} x_i^2, \sum_{i=1}^{k} x_i \right) : x_i = \underline{X}_i, \ldots, \overline{X}_i \right\}.$$

- $P_0 = \{(0,0)\}$; once we know P_k, we compute P_{k+1} as

$\{(y_k + x_{k+1}^2, z_k + x_{k+1}) : (y_k, z_k) \in P_k \ \& \ x_{k+1} = \underline{X}_{k+1}, \ldots, \overline{X}_{k+1}\}.$

- Once we have the set P_n, we compute the desired set Y of possible values of variance as

$$Y = \left\{ \frac{1}{n} \cdot y - \left(\frac{1}{n} \cdot z \right)^2 : (y, z) \in P_n \right\}.$$

- On each of $n+1$ stages, we need $O(n^2)$ computations.
- So overall, we need $n \cdot O(n^2) = O(n^3)$ computational steps.

18. Computing Higher Central Moments $C_m \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E)^k$ in Polynomial Time

- We sequentially compute the set T_k of all possible tuples $(s_1(k), \ldots, s_m(k))$, where $s_\ell(k) = \sum\limits_{i=1}^{k} x_i^\ell$.
- $T_0 = \{(0, \ldots, 0)\}$; based on T_k, we compute T_{k+1} as
$$T_{k+1} = \{((s_1 + x_{k+1}, \ldots, s_m + x_{k+1}^m) : (s_1, \ldots, s_m) \in T_k \,\&$$
$$x_{k+1} = \underline{X}_{k+1}, \ldots, \overline{X}_{k+1}\}.$$
- Once we have T_n, we compute C_m for each of its tuples, since C_m is a linear combination of $s_0(n), \ldots, s_m(n)$.
- On each stage, this computation takes time $O(n^m)$, so overall, we need $n \cdot O(n^m) = O(n^{m+1})$ steps.

19. Computing Covariance in Polynomial Time

- We want to compute the range for covariance

$$C_{xy} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i - \left(\frac{1}{n} \cdot \sum_{i=1}^{n} x_i \right) \cdot \left(\frac{1}{n} \cdot \sum_{i=1}^{n} y_i \right).$$

- We sequentially compute

$$T_k = \left\{ \left(\sum_{i=1}^{k} x_i, \sum_{i=1}^{k} y_i, \sum_{i=1}^{k} x_i \cdot y_i \right) \right\}.$$

- Once we know T_k, we compute

$$T_{k+1} = \{(s_1 + x_{k+1}, s_2 + y_{k+1}, s_3 + x_{k+1} \cdot y_{k+1}) : (s_1, s_2, s_3) \in T_k \,\&$$
$$x_{k+1} = \underline{X}_{k+1}, \ldots, \overline{X}_{k+1} \,\&\, y_{k+1} = \underline{Y}_{k+1}, \ldots, \overline{Y}_{k+1}\}.$$

- Once we know T_n, we can compute Y.
- Each iteration thus takes $O(n^3)$ steps, so the overall computation time is $n \cdot O(n^3) = O(n^4)$.

20. Computing Correlation in Polynomial Time

- Correlation is defined as $\rho_{xy} = \dfrac{C_{xy}}{\sqrt{V_x \cdot V_y}}$, where

$$V_x \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^n x_i^2 - \left(\frac{1}{n} \cdot \sum_{i=1}^n x_i \right)^2, \quad V_y \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^n y_i^2 - \left(\frac{1}{n} \cdot \sum_{i=1}^n y_i \right)^2.$$

- We sequentially compute

$$T_k = \left\{ \left(\sum_{i=1}^k x_i, \sum_{i=1}^k y_i, \sum_{i=1}^k x_i \cdot y_i, \sum_{i=1}^k x_i^2, \sum_{i=1}^k y_i^2 \right) \right\}.$$

$$T_{k+1} = \{(s_1 + x_{k+1}, s_2 + y_{k+1}, s_3 + x_{k+1} \cdot y_{k+1}, s_4 + x_{k+1}^2, s_5 + y_{k+1}^2) :$$
$$(s_1, \ldots, s_5) \in T_k \,\&\, \underline{X}_{k+1} \le x_{k+1} \le \overline{X}_{k+1} \,\&\, \underline{Y}_{k+1} \le y_{k+1} \le \overline{Y}_{k+1}\}.$$

- Once we know T_n, we compute the desired set Y.

- Each iteration thus takes $O(n^5)$ steps, so overall, we need $n \cdot O(n^5) = O(n^6)$ steps.

21. Conclusions

- We are often interested in the quantity y which is related to known x_i by a known dependence $y = f(x_1, \ldots, x_n)$.
- For example:
  - we know the *range* $[\underline{x}_i, \overline{x}_i]$ of each of x_i,
  - we want to find the *range* of possible values of y.
- When we have *fuzzy* information about each x_i, we want to know the resulting *fuzzy* set for Y.
- Usually, we assume that all values from each interval $[\underline{x}_i, \overline{x}_i]$ are possible.
- In practice, we sometimes have an additional information, e.g., that the values x_i are integers.
- We provide feasible algorithms for computing such limited ranges for several functions $f(x_1, \ldots, x_n)$.

22. Acknowledgment

This work was supported in part:

- by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721,
- by Grants 1 T36 GM078000-01 and 1R43TR000173-01 from the National Institutes of Health, and
- by a grant on F-transforms from the Office of Naval Research.