When Is Data Processing Under Interval and Fuzzy Uncertainty Feasible: What If Few Inputs Interact? Does Feasibility Depend on How We Describe Interaction?

Milan Hladík[1], Michal Černý[2], and Vladik Kreinovich[3]
[1]Department of Applied Mathematics, Charles University
Prague, Czech Republic, milan.hladik@matfyz.cz
[2]Department of Econometrics, University of Economics
Prague, Czech Republic, cernym@vse.cz
[3]Department of Computer Science, University of Texas at El Paso
El Paso, Texas 79968, USA, vladik@utep.edu

1. Need for Data Processing

- In many practical situations:
  - we are interested in the value of a quantity y
  - which is difficult or even impossible to measure directly.
- For example, we may be interested in a distance to a faraway star, or in tomorrow's temperature.
- Since we cannot measure y directly, we measure it indirectly: namely,
  - we find easier-to-estimate quantities $x_1, \ldots, x_n$ related to y by a known dependence

$$y = f(x_1, \ldots, x_n),$$

  - and we use the results $\widetilde{x}_i$ of measuring or estimating the quantities x_i to estimate y as $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.

2. Need to Take Uncertainty into Account

- Measurements are never absolutely accurate.
- As a result, the measurement results $\widetilde{x}_i$ are, in general, different from the actual (unknown) values x_i.
- Thus, even if the relation $y = f(x_1, \ldots, x_n)$ is precise,
  - the result $\widetilde{y}$ of applying the algorithm f to the measurement results
  - is, in general, different from the actual value y.
- How accurate is the estimate $\widetilde{y}$?
- In other words, what can we conclude about the measurement errors $\Delta y \stackrel{\text{def}}{=} \widetilde{y} - y$.

3. Taking Uncertainty into Account (cont-d)

- This is definitely important:
  - if we predict tomorrow's temperature as $\widetilde{y} = -2^\circ$ C, and the accuracy of this prediction is $\pm 1^\circ$,
  - then we know that tomorrow will be freezing, with the possibility of ice on the road,
  - so we need to send a warning to the public, put sand (or salt) on the roads, etc.
- On the other hand, if the accuracy is ± 10 degrees:
  - we may still alert the public,
  - but we better wait for more accurate information before placing sand (or salt).

4. Taking Uncertainty into Account (cont-d)

- This is even more important for a spaceship sent to Mars; we want to make sure that:
  - with all the uncertainty taken into account,
  - the spaceship will land in the desired Martian region.

5. Case of Interval Uncertainty

- In many practical situations:
  - the only information that we have about the measurement errors $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$ is
  - the upper bound Δ_i on its absolute value:

$$|\Delta x_i| \le \Delta_i.$$

- In this case, once we know the measurement result $\widetilde{x}_i$:
  - the only information that we have about the actual (unknown) values x_i is
  - that x_i belongs to the interval

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

6. Case of Interval Uncertainty (cont-d)

- Different values x_i from the corresponding intervals lead, in general, to different values $y = f(x_1, \ldots, x_n)$.
- In this case, we would like to find the range of all possible values of y: $[\underline{y}, \overline{y}] = f([\underline{x}_1, \overline{x}_1], \ldots, [\underline{x}_n, \overline{x}_n]) \stackrel{\text{def}}{=}$

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

- The problem of computing this range is known as the problem of *interval computations*.

7. Already for Interval Uncertainty, the Corresponding Problem Is NP-Hard

- Sometimes, the function $f(x_1, \ldots, x_n)$ is linear:

$$f(x_1, \ldots, x_n) = a_0 + \sum_{i=1}^{n} a_i \cdot x_i.$$

- Then, we have explicit formulas for the range:

$$\underline{y} = \widetilde{y} - \Delta, \overline{y} = \widetilde{y} + \Delta, \text{ where } \Delta = \sum_{i=1}^{n} |a_i| \cdot \Delta_i.$$

- However, already for quadratic functions $f(x_1, \ldots, x_n)$, the problem of computing the range $[\underline{y}, \overline{y}]$ is NP-hard.
- This means that, if P$\neq$NP (as most computer scientists believe) then:
  - no feasible algorithms is possible
  - that would solve all particular cases of this problem.

8. Case of Fuzzy Uncertainty

- In many practical situations, in addition to the upper bounds Δ_i on the measurement error:
  - experts also tell us which values from the corresponding interval $[-\Delta_i, \Delta_i]$ are more probable and
  - which values are less probable.
- This information is usually given in terms of imprecise ("fuzzy") words form natural langauge, such as:
  - "somewhat probable",
  - "very probable", etc.
- We need to describe such knowledge in precise computer-understandable terms.
- For this purpose, Zadeh invented the technique of *fuzzy logic*.

9. Case of Fuzzy Uncertainty (cont-d)

- In this technique:
  - to describe each imprecise property like "somewhat probable",
  - we ask the expert to mark, on a scale from 0 to 10, to what extent the corresponding value is possible.
- If an expert marks 7, we take 7/10 as the degree to which the corresponding value is possible.
- As a result, in addition to the interval $[-\Delta_i, \Delta_i]$, we also have:
  - for each value Δx_i from this interval,
  - a degree $\mu_i(\Delta x_i)$ to which this value is possible.
- The function that assigns, to each value Δx_i, the corresponding degree, is known as the *membership function*.

10. Data Processing Under Fuzzy Uncertainty

- A value y is possible if $y = f(x_1, \ldots, x_n)$ for some tuples for which:
  - x_1 is a possible value of the first input and
  - x_2 is a possible value of the second inputs, etc.
- We know the degrees $\mu_i(x_i)$ to which each x_i is a possible value of the i-th input.
- We need to estimate the degree to which x_1 is possible *and* x_2 is possible, etc.
- It is reasonable to use a corresponding "and"-operation $_\&(a, b)$ (t-norm) of fuzzy logic, resulting in

$$f_\&(\mu_1(x_1), \ldots, \mu_n(x_n)).$$

- The simplest such operation is $f_\&(a, b) = \min(a, b)$, in which case the corresponding inputs has the form

$$\min(\mu_1(x_1), \ldots, \mu_n(x_n)).$$

11. Data Processing Under Fuzzy Uncertainty (cont-d)

- We need to find the degree $\mu(y)$ corresponding to the possibility of having either one tuple *or* another.
- So, we can similarly apply an "or"-operation (t-conorm) $f_\vee(a, b)$.
- The simplest "or"-operation is $f_\vee(a, b) = \max(a, b)$.
- Then, we get $\mu(y) =$

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

- This formula wa originally proposed by Zadeh and is thus known as *Zadeh's extension principle.*

12. Data Processing Under Fuzzy Uncertainty: Computational Aspects

- From the computational viewpoint, this formula can be described in terms of α-*cuts*

$$\mathbf{x}_i(\alpha) \stackrel{\text{def}}{=} \{x_i : \mu_i(x_i) \geq \alpha\}.$$

- For every α, $\mathbf{y}(\alpha) = f(\mathbf{x}_1(\alpha), \ldots, \mathbf{x}_n(\alpha))$.
- Thus, from the computational viewpoint:
  - propagation of fuzzy uncertainty can be reduced to
  - several interval computation problems corresponding, e.g., to $\alpha = 0, 01, \ldots, 0.9, 1.0$.
- Because of this reduction, in the following text, we will consider only the case of interval uncertainty.

13. How to Describe the Dependence?

- In some cases, we know the dependence $f(x_1, \ldots, x_n)$ from physics.
- In many other cases, however, we need to determine this dependence experimentally.
- For this, we need to:
  - first select a reasonable finite-parametric family of functions, and
  - then find the parameters from the experiments.
- When we analyze the dependence of the desired quantity y on the auxiliary quantities $x_1, \ldots, x_n$,
  - the first thing we usually do
  - is analyze how y changes if we change only of these inputs.
- For each input x_i, we get dependence $y = f_i(x_i)$.

14. How to Describe the Dependence (cont-d)

- This dependence may be linear, maybe quadratic, etc.
- In some cases, inputs are independent:
  - the changes in y caused by each inputs x_i
  - do not depend on the values of all the other inputs x_j with $j \neq i$.
- In this case, the resulting dependence has the form

$$f(x_1, \ldots, x_n) = \sum_{i=1}^{n} f_i(x_i).$$

- In this case, the range $[\underline{y}, \overline{y}]$ is equal to the sum of the ranges corresponding to each of the inputs:

$$\underline{y} = \underline{y}_1 + \ldots + \underline{y}_n, \quad \overline{y} = \overline{y}_1 + \ldots + \overline{y}_n, \text{ where}$$

$$[\underline{y}_i, \overline{y}_i] \stackrel{\text{def}}{=} \{f_i(x_i) : x_i \in [\underline{x}_i, \overline{x}_i]\}.$$

15. How to Describe the Dependence (cont-d)

- For simple functions $f_i(x_i)$ like linear or quadratic, the range is easy to compute.
- Thus, the corresponding interval computations problem is feasible.
- In practice, inputs often interact.
- A natural idea is to use bilinear terms $x_i \cdot x_j$ to describe such an interaction.
- In this case, we get a general quadratic formula, for which the corresponding problems are NP-hard.

16. Formulation of the Problems

- NP-hardness comes from considering the case when all inputs interact with each other.
- What if only a few inputs interact?
- This is the first question for which we provide an answer in this paper.
- In the NP-hardness result, we assume that the interaction is described by bilinear terms.
- However, other expressions are also possible.
- For example, in chemical kinetics:
  - for small concentrations of the chemicals,
  - the interactions are bilinear $x_i \cdot x_j$.

17. Formulation of the Problems (cont-d)

- However, for very strong concentrations, the interaction is described by a different formula $\min(x_i, x_j)$.
- For intermediate concentration, we get a more complex formula.
- Will the general result remain NP-hard if we consider such interaction?
- This is the second question for which we provide the answer.

18. Answer to the First Question

- Our first result is that:
  - if we have a quadratic form in which only $O(\log(n))$ pairs of interacting inputs,
  - then we have a feasible algorithm for estimating the range $[\underline{y}, \overline{y}]$.

19. Analysis of the Problem and Auxiliary Result

- What if we have more interacting inputs?
- It is known that $\log(n)$ can be viewed as a limit of power functions n^ε when $\varepsilon \to 0$.
- So, a natural next question is: what if we have n^ε interacting inputs, for some small ε?
- **Result:** If we allow n^ε interacting inputs, then the problem of computing the range $[\underline{y}, \overline{y}]$ is NP-hard.

20. Answering the Second Question

- Instead of the usual interaction terms $x_i \cdot x_j$, we allow more general terms $f_{ij}(x_i, x_j)$.
- If one of the inputs is absent, i.e., if $x_i = 0$, then there is usually no interaction.
- So we can safely assume that $f_{ij}(0, x_j) = f_{ij}(x_i, 0) = 0$ for all x_i and x_j.
- Let us make the comparison with the product term (for which $f_{ij}(1,1) = 1$) easier.
- For this purpose, let us divide and multiply the expression $f_{ij}(x_i, x_j)$ by $a_{ij} \stackrel{\text{def}}{=} f_{ij}(1,1)$.
- Then $f_{ij}(x_i, x_j)$ takes the form $a_{ij} \cdot T_{ij}(x_i, x_j)$, where

$$T_{ij}(x_i, x_j) \stackrel{\text{def}}{=} \frac{f_{ij}(x_i, x_j)}{a_{ij}}.$$

21. Answering the Second Question (cont-d)

- It is reasonable to require that small changes in x_i and x_j should lead to small changes in T_{ij}.
- In precise terms, let us require that for some Lipschitz constant L:

$$|T_{ij}(x_i, x_j) - T_{ij}(x'_i, x'_j)| \le L \cdot (|x_i - x'_i| + |x_j - x'_j|).$$

- In this case, for quadratic $f_i(x_i)$, we consider expressions of the type

$$f(x, \ldots, x_n) = \sum_{i=1}^{n} f_i(x_i) + \sum_{i \ne j} a_{ij} \cdot T_{ij}(x_i, x_j).$$

- **Result:** Computing the range $[\underline{y}, \overline{y}]$ of a function of the above type over a given box is NP-hard.

22. Acknowledgments

- M. Hladík was supported by the Czech Science Foundation Grant P403-18-04735S.
- The work of M. Cerny was supported by the Czech Science Foundation under Grant P402/12/G097.
- V. Kreinovich was supported in part by the National Science Foundation grant HRD-1242122.

23. Proof of the First Result

- We know that only $v = O(\log(n))$ many inputs $x_{i_1}, \ldots, x_{i_v}$ are involved in the interaction.
- So, we can describe the desired quadratic function as the sum

$$f(x_1, \ldots, x_n) = \sum_{i \neq i_k} f_i(x_i) + r(x_{i_1}, \ldots, x_{i_v}).$$

- Here $f_i(x_i)$ is a quadratic function of one variable, and $r(x_{i_1}, \ldots, x_{i_v})$ is a quadratic function of v variables.
- Since each of the terms in the above sum depends on each own inputs, we conclude that

$$\underline{y} = \sum_{i \neq i_k} \underline{y}_i + \underline{r} \text{ and } \overline{y} = \sum_{i \neq i_k} \overline{y}_i + \overline{r},$$

$$\underline{r} = \min\{r(x_{i_1}, \ldots, x_{i_v}) : z_{i_1} \in [\underline{x}_{i_1}, \overline{x}_{i_1}], \ldots, x_{i_v} \in [\underline{x}_{i_v}, \overline{x}_{i_v}]\},$$

$$\overline{r} = \max\{r(x_{i_1}, \ldots, x_{i_v}) : x_{i_1} \in [\underline{x}_{i_1}, \overline{x}_{i_1}], \ldots, x_{i_v} \in [\underline{x}_{i_v}, \overline{x}_{i_v}]\}.$$

Need for Data Processing
Need to Take . . .
Case of Interval . . .
Case of Fuzzy Uncertainty
How to Describe the . . .
Formulation of the . . .
Answer to the First . . .
Analysis of the . . .
Answering the Second . . .
Home Page
Title Page
Page 24 of 35
Go Back
Full Screen
Close
Quit

24. Proof of the First Result (cont-d)

- Minima and maxima $\underline{y}_i$ and $\overline{y}_i$ of a quadratic function $f_i(x_i)$ over an interval are easy to compute. Thus:
  - to show that computing $[\underline{y}, \overline{y}]$ is feasible,
  - we need to show how to feasibly compute min and max of $r(x_{i_1}, \ldots, x_{i_v})$ over the box

$$[\underline{x}_{i_1}, \overline{x}_{i_1}] \times \ldots \times [\underline{x}_{i_v}, \overline{x}_{i_v}].$$

- According to calculus, a maximum or a minimum of a function $F(z)$ on an interval $[\underline{z}, \overline{z}]$ is attained:
  - either at a point which is inside the interval $(\underline{z}, \overline{z})$, in which case

$$\frac{dF}{dz} = 0;$$

  - or at the left endpoint $z = \underline{z}$ of the give interval,
  - or at the right endpoint $z = \overline{z}$ of this interval.

25. Proof of the First Result (cont-d)

- Similarly, min and max of $F(z_1, \ldots, z_v)$ on $[\underline{z}_1, \overline{z}_1] \times \ldots \times [\underline{z}_v, \overline{z}_v]$ is attained when:
  - for each of the v variables z_i,
  - one of the following three situations happens:
  - either the corresponding value z_i is inside the interval $(\underline{z}_i, \overline{z}_i)$, in which case $\dfrac{\partial F}{\partial z_i} = 0$;
  - or the optimizing value is at the left end of the corresponding interval $z_i = \underline{z}_i$,
  - or the optimizing value is at the right end of the corresponding interval $z_i = \overline{z}_i$.
- For each variable, we have 3 options.
- So, for two variables, we have $3 \cdot 3 = 9$ possible options.
- For v variables, we have 3^v possible options.

26. Proof of the First Result (cont-d)

- In each of these 3^v options, for each variables z_i, we have either $z_i = \underline{z}_i$, or $z_i = \overline{z}_i$, or $\dfrac{\partial F}{\partial z_i} = 0$.
- The first two equations are clearly linear in z_i.
- In our case, $z_k = x_{i_k}$ and $F(z_1, \ldots, z_v) = r(z_1, \ldots, z_v)$ is quadratic, so derivatives are linear.
- Thus, the equation $\dfrac{\partial F}{\partial z_i} = 0$ is also linear in $z_1, \ldots, z_v$.
- So, in each of the 3^v cases, we have a system of linear equations to find the corresponding values $z_1, \ldots, z_v$.
- Such a system can be feasible solved.

27. Proof of the First Result (cont-d)

- Out of all cases for which each component z_i of the solution is within the corresponding interval, we choose:
  - the smallest as $\underline{r}$ and
  - the largest as $\overline{r}$.
- When $v = O(\log(n))$, i.e., $v \le C \cdot \log(n)$ for some C, we have $3^v \le 3^{C\cdot\log(n)} = n^{\log(3)\cdot C}$ systems.
- Thus, the number of linear system is polynomial in n.
- Hence, the overall time for solving all these systems is also bounded by a polynomial in n.
- This time is thus feasible.
- This proves our main result.

28. Proof of the Auxiliary Result

- Formally, NP-hard means that any problem from a class NP can be reduced to this problem.
- Thus, if we can reduce a known NP-hard problem to a new problem, this means,
  - by transitivity of reduction,
  - that every problem from the class NP can be reduced to the new problem as well, and
  - thus, that the new problem is also NP-hard.
- We know that the problem of estimating the range of a general quadratic function over a given box is NP-hard.
- Let us reduce this known NP-hard problem to our new problem – of estimating
  - the range of a quadratic function
  - in which at most n^ε inputs interact.

29. Proof of the Auxiliary Result (cont-d)

- For this, let us start with any original quadratic form

$$Q(x_1, \ldots, x_m).$$

- Then, we add $M = n^{1/\varepsilon}$ new variables $v_1, \ldots, v_M$, and consider a new quadratic function

$$f(x_1, \ldots, x_m, v_1, \ldots, v_M) = Q(x_1, \ldots, x_m) + \sum_{j=1}^{M} v_j.$$

- For this function, only inputs $x_1, \ldots, x_m$ interact.
- So out of $n = m + M$ variables, only $O(n^\varepsilon)$ interact with each other.

30. Proof of the Auxiliary Result (cont-d)

- On the other hand:
  - since the new function f is the sum of expressions each of which depends only on its own variables,
  - we conclude that its range $[\underline{y}, \overline{y}]$ has the form

$$\underline{y} = \underline{q} + \sum_{j=1}^{M} \underline{v}_j \text{ and } \overline{y} = \overline{q} + \sum_{j=1}^{M} \overline{v}_j.$$

- Here $\underline{q}$ and $\overline{q}$ are min and max of the original quadratic expression $Q(x_1, \ldots, x_m)$ on the corr. box.
- So, if we know the bounds for f, we can easily find the bounds for Q, and vice versa.
- Thus, computing the range of f is indeed feasibly equivalent to computing the range of Q.
- So, we have the desired reduction, and thus, the problem is indeed NP-hard.

31. Proof of the Second Result

- It is known that computing the range of a quadratic function over a box is NP-hard already when:
  - the quadratic form is positive definite (i.e., the function is convex) and
  - the range of each variable is $[\underline{x}_i, \overline{x}_i] = [0, 1]$.
- Reduction to $[0, 1]$ can be easily achieved by a linear transformation of each variable.
- To be more precise, for convex functions:
  - computing the minimum $\underline{y}$ is feasible, but
  - computing the maximum $\overline{y}$ is NP-hard.

32. Proof of the Second Result (cont-d)

- Let us reduce the NP-hard problem of computing this maximum to the new problem.
- Let us start with a general convex quadratic expression

$$f(x_1,\ldots,x_n)=a_0+\sum_{i=1}^n a_i\cdot x_i+\sum_{i,j} a_{ij}\cdot x_i\cdot x_j.$$

- By separating quadratic terms corresponding to $i=j$ and $i\neq j$, we get

$$f(x_1,\ldots,x_n)=a_0+\sum_{i=1}^n a_i\cdot x_i+\sum_{i=1}^n a_{ii}\cdot x_i^2+$$

$$\sum_{i\neq j} a_{ij}\cdot x_i\cdot x_j.$$

33. Proof of the Second Result (cont-d)

- Let us consider a new function for some $\beta > 0$:

$$F(x_1, \ldots, x_n) = a_0 + \sum_{i=1}^n a_i \cdot x_i + \sum_{i=1}^n a_{ii} \cdot x_i^2 +$$

$$\sum_{i \neq j} a_{ij} \cdot T_{ij}(x_i, x_j) + \beta \cdot \sum_{i=1}^n (2x_i - 1)^2.$$

- Due to the Lipschitz condition, for sufficiently large β, the function $F(x_1, \ldots, x_n)$ is convex.
- For a convex function, the maximum $\overline{Y}$ on a convex set $[0,1]^n$ is attained at one of the vertices.
- So, $\overline{Y}$ is attained when each x_i is 0 or 1:

$$\overline{Y} = \max_{x_i \in \{0,1\}} F(x_1, \ldots, x_n).$$

- On each vertex $T_{ij}(x_i, x_j) = x_i \cdot x_j$ and $(2x_i - 1)^2 = 1$.

34. Proof of the Second Result (cont-d)

- So, for vertices $(x_1, \ldots, x_n)$, we have

$$F(x_1, \ldots, x_n) = f(x_1, \ldots, x_n) + \beta \cdot n.$$

- The maximum $\overline{y}$ of the original convex quadratic function $f(x_1, \ldots, x_n)$ is also attained at one of the vertices:

$$\overline{y} = \max_{x_i \in \{0,1\}} f(x_1, \ldots, x_n).$$

- Thus, $\overline{Y} = \max\limits_{x_i \in \{0,1\}} F(x_1, \ldots, x_n) =$

$$\max_{x_i \in \{0,1\}} (f(x_1, \ldots, x_n) + \beta \cdot n) =$$

$$\max_{x_i \in \{0,1\}} f(x_1, \ldots, x_n) + \beta \cdot n = \overline{y} = \beta \cdot n.$$

- So, we get $\overline{Y} = \overline{y} + \beta \cdot n$.
- Thus, the computation of $\overline{y}$ is indeed feasible reduced to computing $\overline{Y}$; so, our problem is also NP-hard.