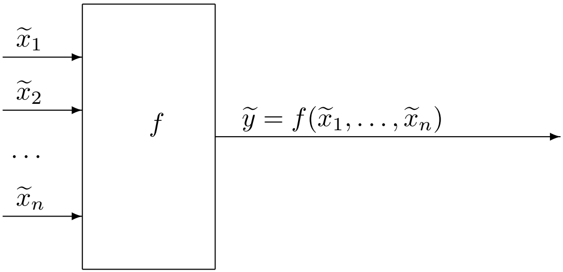
Data Processing Under Imprecise Probabilistic Uncertainty: Computational Challenges

Vladik Kreinovich
Department of Computer Science
University of Texas at El Paso, 500 W. University
El Paso, TX 79968, USA, vladik@utep.edu

http://www.cs.utep.edu/vladik
http://www.cs.utep.edu/interval-comp

1. General Problem of Data Processing under Uncertainty

- *Indirect measurements:* way to measure y that are are difficult (or even impossible) to measure directly.
- *Idea:* $y = f(x_1, \ldots, x_n)$

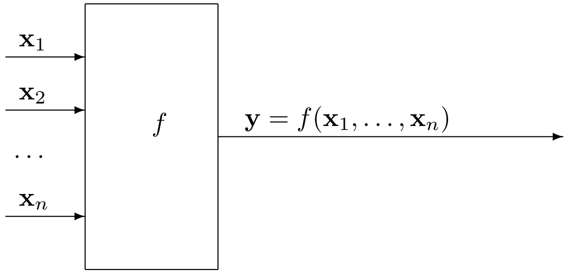


- *Problem:* measurements are never 100% accurate: $\widetilde{x}_i \neq x_i$ ($\Delta x_i \neq 0$) hence

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

What are bounds on $\Delta y \stackrel{\text{def}}{=} \widetilde{y} - y$?

2. General Challenge



- *Given:*
  - an algorithm $y = f(x_1, \ldots, x_n)$ that transforms n real numbers x_i into a number y;
  - n intervals $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$.
- *Compute:* the corresponding range of y:

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

- *Fact:* even for quadratic f, the problem of computing the exact range $\mathbf{y}$ is NP-hard.
- *Challenge:*
  - find classes of problems for which efficient algorithms are possible; and
  - for problems outside these classes, find efficient techniques for *approximating* uncertainty of y.

3. General Approach: Interval-Type Step-by-Step Techniques

- *Problem:*
- *Solution:* compute an enclosure $\mathbf{Y}$ such that $\mathbf{y} \subseteq \mathbf{Y}$.
- *Interval arithmetic:* for arithmetic operations $f(x_1, x_2)$, we have explicit formulas for the range.
- *Examples:* when $x_1 \in \mathbf{x}_1 = [\underline{x}_1, \overline{x}_1]$ and $x_2 \in \mathbf{x}_2 = [\underline{x}_2, \overline{x}_2]$, then:
  - The range $\mathbf{x}_1 + \mathbf{x}_2$ for $x_1 + x_2$ is $[\underline{x}_1 + \underline{x}_2, \overline{x}_1 + \overline{x}_2]$.
  - The range $\mathbf{x}_1 - \mathbf{x}_2$ for $x_1 - x_2$ is $[\underline{x}_1 - \overline{x}_2, \overline{x}_1 - \underline{x}_2]$.
  - The range $\mathbf{x}_1 \cdot \mathbf{x}_2$ for $x_1 \cdot x_2$ is $[\underline{y}, \overline{y}]$, where

$$\underline{y} = \min(\underline{x}_1 \cdot \underline{x}_2, \underline{x}_1 \cdot \overline{x}_2, \overline{x}_1 \cdot \underline{x}_2, \overline{x}_1 \cdot \overline{x}_2);$$

$$\overline{y} = \max(\underline{x}_1 \cdot \underline{x}_2, \underline{x}_1 \cdot \overline{x}_2, \overline{x}_1 \cdot \underline{x}_2, \overline{x}_1 \cdot \overline{x}_2).$$

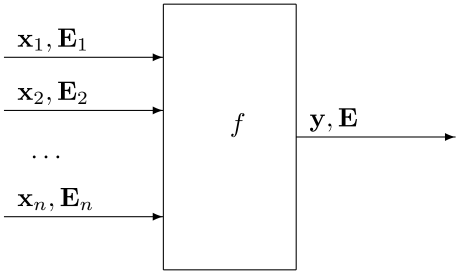
- The range $1/\mathbf{x}_1$ for $1/x_1$ is $[1/\overline{x}_1, 1/\underline{x}_1]$ (if $0 \notin \mathbf{x}_1$).

4. General Approach: Example

- *Example:* $f(x) = (x-2)\cdot(x+2)$, $x \in [1,2]$.
- How will the computer compute it?
  - $r_1 := x - 2$;
  - $r_2 := x + 2$;
  - $r_3 := r_1 \cdot r_2$.
- *Main idea:* do the same operations, but with *intervals* instead of *numbers*:
  - $\mathbf{r}_1 := [1,2] - [2,2] = [-1,0]$;
  - $\mathbf{r}_2 := [1,2] + [2,2] = [3,4]$;
  - $\mathbf{r}_3 := [-1,0] \cdot [3,4] = [-4,0]$.
- *Actual range:* $f(\mathbf{x}) = [-3,0]$.
- *Comment:* this is just a toy example, there are more efficient ways of computing an enclosure $\mathbf{Y} \supseteq \mathbf{y}$.

5. General Approach: Successes

- *Objective:* make decisions $E_x[u(x,a)] \to \max a$.
- For smooth $u(x)$, we have $u(x) = u(x_0) + (x - x_0) \cdot u'(x_0) + \ldots$, so we must know moments to estimate $E[u]$.
- For threshold-type $u(x)$, we need cdf $F(x) = \text{Prob}(\xi \le x)$.
- *General solution:* parse to elementary operations $+$, $-$, $\cdot$, $1/x$, max, min.
- Explicit formulas for arithmetic operations known for intervals, for p-boxes $\mathbf{F}(x) = [\underline{F}(x), \overline{F}(x)]$, for intervals + 1st moments $E_i \stackrel{\text{def}}{=} E[x_i]$:



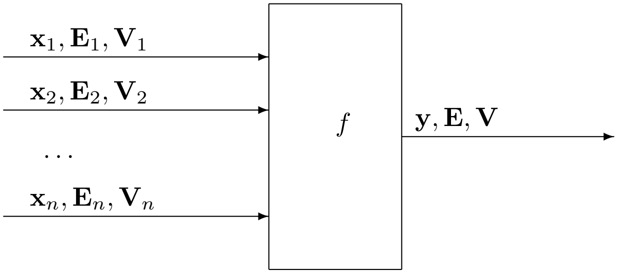
6. Successes (cont-d)

- *Easy cases:* $+$, $-$, product of independent x_i.
- *Example of a non-trivial case:* multiplication $y = x_1 \cdot x_2$, when we have no information about the correlation:
  - $\underline{E} = \max(p_1 + p_2 - 1, 0) \cdot \overline{x}_1 \cdot \overline{x}_2 + \min(p_1, 1 - p_2) \cdot \overline{x}_1 \cdot \underline{x}_2 +$ $\min(1 - p_1, p_2) \cdot \underline{x}_1 \cdot \overline{x}_2 + \max(1 - p_1 - p_2, 0) \cdot \underline{x}_1 \cdot \underline{x}_2$;
  - $\overline{E} = \min(p_1, p_2) \cdot \overline{x}_1 \cdot \overline{x}_2 + \max(p_1 - p_2, 0) \cdot \overline{x}_1 \cdot \underline{x}_2 +$ $\max(p_2 - p_1, 0) \cdot \underline{x}_1 \cdot \overline{x}_2 + \min(1 - p_1, 1 - p_2) \cdot \underline{x}_1 \cdot \underline{x}_2$,

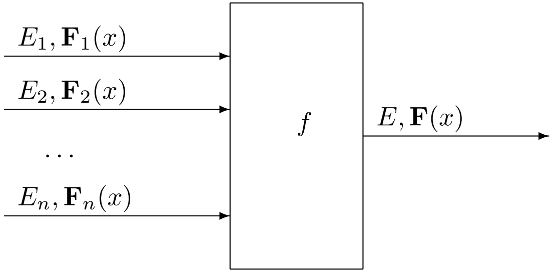
  where $p_i \stackrel{\text{def}}{=} (E_i - \underline{x}_i)/(\overline{x}_i - \underline{x}_i)$.

7. Challenges

- intervals + 2nd moments (e.g. – chip design):



- moments + p-boxes; e.g.:



8. Additional Challenges

- *General problem:* how to efficiently deduce the statistical information from, e.g., interval data.
- *Example:* we know intervals $\mathbf{x}_1 = [\underline{x}_1, \overline{x}_1]$, ..., $\mathbf{x}_n = [\underline{x}_n, \overline{x}_n]$, we want to compute the ranges of possible values of the population mean $\mu = \frac{1}{n}\sum\limits_{i=1}^{n} x_i$, population variance $V = \frac{1}{n}\sum\limits_{i=1}^{n}(x_i - \mu)^2$, etc.
- *Difficulty:* in general, this problem is NP-hard even for the variance.
- *Known:*
  - efficient algorithms for $\underline{V}$,
  - efficient algorithms for $\overline{V}$ for reasonable situations, etc.
- *Challenges:* finding the ranges of covariance, correlation, etc.