# Are Your Computations Accurate, Private, and Secure?

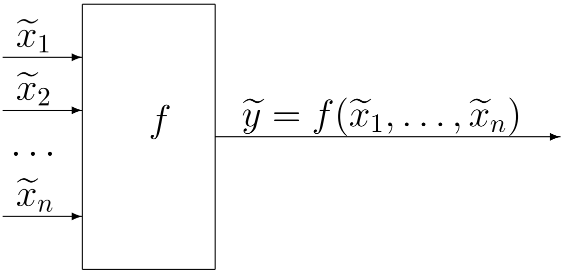
Luc Longpré and Vladik Kreinovich

Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968, USA
longpre@utep.edu, vladik@utep.edu

Interval computations website:
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:* $y = f(x_1, \ldots, x_n)$



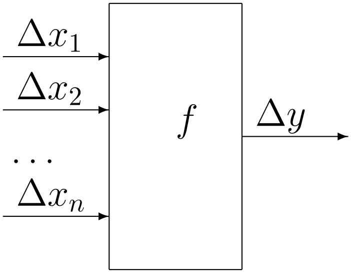

- *Problem:* measurements are never 100% accurate: $\widetilde{x}_i \neq x_i$ $(\Delta x_i \neq 0)$ hence

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

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


General Problem of . . .
Probabilistic and . . .
Interval . . .
Interval Arithmetic: . . .
Case Studies
Need for Statistical . . .
Need for Statistical . . .
Maintaining Privacy is . . .
Maintaining Privacy: . . .
Interval Approach to . . .
Title Page
◀◀ ▶▶
◀ ▶
Page 2 of 11
Go Back
Full Screen
Close
Quit

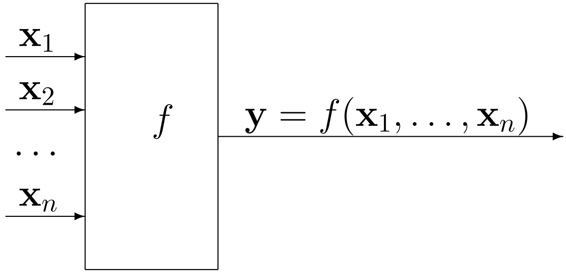
# 2. Probabilistic and Interval Uncertainty




- *Traditional approach:* we know probability distribution for $\Delta x_i$ (usually Gaussian).
- *Where it comes from:* calibration using standard MI.
- *Problem:* calibration is not possible in:
  - fundamental science
  - manufacturing
- *Solution:* we know upper bounds $\Delta_i$ on $|\Delta x_i|$ hence

$$x_i \in [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i].$$

# 3. Interval Computations: A Problem




- *Given:* an algorithm $y = f(x_1, \ldots, x_n)$ and $n$ 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:* NP-hard even for quadratic $f$.
- *Challenge:* when are feasible algorithm possible?
- *Challenge:* when computing $\mathbf{y} = [\underline{y}, \overline{y}]$ is not feasible, find a good approximation $\mathbf{Y} \supseteq \mathbf{y}$.


Title Page
◀◀ ▶▶
◀ ▶
Page 4 of 11
Go Back
Full Screen
Close
Quit

# 4. Interval Arithmetic: Foundations of Interval Techniques

- *Problem:* compute the range

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

- *Interval arithmetic:* for arithmetic operations $f(x_1, x_2)$ (and for elementary functions), 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 $0 \notin \mathbf{x}_1$).

# 5. Case Studies

- *Chip design:* one of the main objectives is to decrease the clock cycle.
- *Bioinformatics:* find genetic difference between cancer cells and healthy cells.
- *Ideal case:* we directly measure concentration $c$ of the gene in cancer cells and $h$ in healthy cells.
- *In reality:* difficult to separate.
- *Solution:* we measure $y_i \approx x_i \cdot c + (1 - x_i) \cdot h$, where $x_i$ is the percentage of cancer cells in $i$-th sample.
- *Outlier Detection Under Interval Uncertainty.* In some practical situations, we only have intervals $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$.
- *Example:* structural integrity – not to miss a fault.
- *Example:* before a surgery, we want to make sure that there is a micro-calcification.

# 6. Need for Statistical Databases

- *Fact:* in many areas, statistics is gathered.
- *Why:* it is useful for many practical situations.
- *Example of gathering statistics:* a census.
- *Information gathered:* data about health, employment, and mortality in different regions.
- *Application:* so that resources can be allocated where they are needed the most.
- *Other applications:* industrial and medical fields.
- *Statistical databases:* databases whose intent is for outside users to compute statistics.

# 7. Need for Statistical Analysis, Need for Privacy

- *What we want to compute:* statistical characteristics such as
  - statistical moments, such as mean $E$, variance $V = M_2$, skewness $S = M_3$, and higher central moments $M_m$,
  - covariance $C_{xy}$, correlation $\rho$, etc.
- *Applications:* these characteristics provide valuable information on the distribution of the data.
- *Need for privacy:* a large part of this data is *sensitive*, such as salaries, medical information, etc.
- *Objective:*
  - outside users *should* be able to perform statistical analysis,
  - but outside users *should not* be able to get sensitive information about individuals.

## 8. Maintaining Privacy is Not Easy

- *Misconception:* anonymity, averaging protect privacy.
- *Main idea of anonymity:* delete the names from all the records.
  - *Toy example:* faculty data, with salary, department, education.
  - *Privacy violation:* ask for the data about a CS Dept. professor with PhD from Russia.
- *Main idea of averaging:* only return averages.
  - *Toy example:* same salaries database.
  - *Privacy violation:* ask for the average salary $E_{\rm all}$ and $E_{\rm nR}$ of all CS professors and all whose PhD is not from Russia:

$$E_{\rm all}=\frac{1}{n}\cdot\sum_{i=1}^{n} s_i,\;\; E_{\rm nR}=\frac{1}{n-1}\cdot\sum_{i\ne i_R} s_i,\;\; s_{i_R}=n\cdot E_{\rm all}-(n-1)\cdot E_{\rm nR}.$$

# 9. Maintaining Privacy: Interval Approach

- *Main idea:* instead of storing the actual values $x_i$, we only store *ranges* $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$.
- *Traditional approach:* we ask a person $i$ for his or her age $x_i$.
- *Interval approach:* we only ask whether the age is between, say, 0 and 10, 10 and 20, 20 and 30, etc.
- *Example:* a 28 years-old person.
- *What we store:* we only store the interval value $[20, 30]$ years in the *age* field of this person's record.
- *Fact:* we do not store the actual data.
- *Result—privacy is preserved:* we cannot reconstruct the actual data, no matter how many queries we ask.

# 10. Interval Approach to Preserving Privacy: Computational Challenges

- *Reminder:* to preserve privacy, instead of the actual values $x_i$, we only store their ranges $\mathbf{x}_i$.
- *New problem:* what to return if a query asks for a statistical characteristic $C(x_1, \ldots, x_n)$ such as variance?
- *Difficulty:* different possible values $x_i \in \mathbf{x}_i$ lead, in general, to different values $C(x_1, \ldots, x_n)$.
- *Possible solution:* return the range of possible values of $C(x_1, \ldots, x_n)$:

$$\mathbf{C} = [\underline{C}, \overline{C}] \stackrel{\text{def}}{=} \{C(x_1, \ldots, x_n) : x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\}.$$

- *Computational problem:* how to compute $\mathbf{C}$?
- *Solution:* this is a particular case of interval computations.