Two Etudes on Combining Probabilistic and Interval Uncertainty: Processing Correlations and Measuring Loss of Privacy

Martine Ceberio, Gang Xiang
Luc Longpré, Vladik Kreinovich
University of Texas at El Paso

Hung T. Nguyen
New Mexico State University

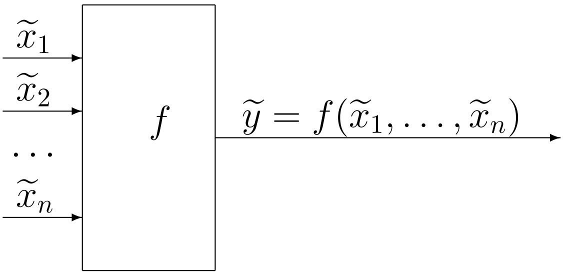
Daniel Berleant
University of Arkansas at Little Rock

1. Overview

- In many practical situations, there is a need to combine interval and probabilistic uncertainty.
- The need for such a combination leads to two types of problems:
  - how to process the given combined uncertainty, and
  - how to gauge the amount of uncertainty and – a related question – how to best decrease this uncertainty.
- In our research, we concentrate on these two types of problems.
- In this talk, we present two examples that illustrate how the corresponding problems can be solved.

2. General Problem of Data Processing under Uncertainty

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

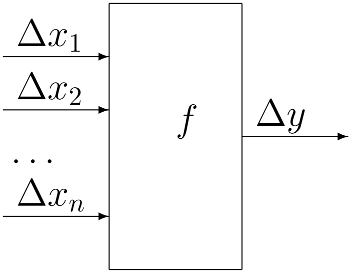


- *Problem:* measurements are never 100% accurate: $\widetilde{x}_i \neq x_i$ ($\Delta x_i \neq 0$) hence

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

- *Question:* what are bounds on $\Delta y \stackrel{\text{def}}{=} \widetilde{y} - y$?

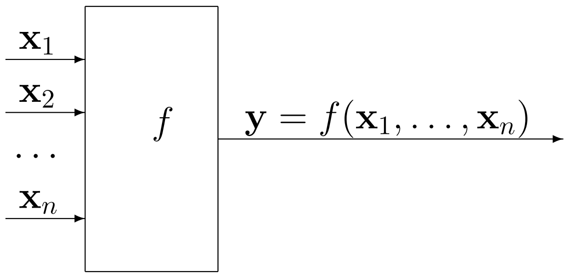
3. Probabilistic and Interval Uncertainty



- *Traditional approach:* we know probability distribution for Δx_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 Δ_i on $|\Delta x_i|$ hence

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

4. 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 $\mathbf{Y} \supseteq \mathbf{y}$.

5. Interval Arithmetic: Foundations of Interval Techniques

- *Problem:* compute the range

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

- *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 $0 \notin \mathbf{x}_1$).

6. Straightforward Interval Computations: 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:* perform 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 $\mathbf{Y} \supseteq \mathbf{y}$.

7. Combining Interval and Probabilistic Uncertainty

- *Situation:* in some cases, in addition to the bounds on each variables, we have partial information about its probability distribution.
- *Problem:* there are many ways to represent a probability distribution.
- *Idea:* look for an objective.
- *Objective:* make decisions $E_x[u(x,a)] \to \max a$.
- *Analysis:* for smooth $u(x)$, we have

$$u(x) = u(x_0) + (x - x_0) \cdot u'(x_0) + \frac{1}{2} \cdot (x - x_0)^2 \cdot u''(x_0) + \ldots$$

so

$$E[u(x)] = u(x_0) + E[x - x_0] \cdot u'(x_0) + \frac{1}{2} \cdot E[(x - x_0)^2] \cdot u''(x_0) + \ldots$$

- *Conclusion:* we must know moments to estimate $E[u]$.

8. Extension of Interval Arithmetic to Probabilistic Case: Successes

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

9. First Result

- *Problem:* the above expression is computationally complicated.
- *New result:* new, equivalent, more computationally efficient expressions for $\underline{E}$ and $\overline{E}$:

$$\underline{E} = E_1 \cdot E_2 - \min((E_1 - \underline{x}_1) \cdot (E_2 - \underline{x}_2), (\overline{x}_1 - E_1) \cdot (\overline{x}_2 - E_2));$$

$$\overline{E} = E_1 \cdot E_2 + \min((E_1 - \underline{x}_1) \cdot (\overline{x}_2 - E_2), (\overline{x}_1 - E_1) \cdot (E_2 - \underline{x}_2)).$$

10. Taking Correlation into Account: A Problem

- *Fact:* the range of $E[x_1 \cdot x_2]$ depends on the ranges of $E[x_i]$ and on the correlation between the x_i.
- *Previously covered:*
  - case when x_1 and x_2 are independent, and
  - case when we have no information about their correlation.
- *Practical situation:* sometimes, we know the interval $[\underline{\rho}, \overline{\rho}]$ of possible values of the correlation ρ:

$$\rho(x_1, x_2) \stackrel{\text{def}}{=} \frac{E[x_1 \cdot x_2] - E_1 \cdot E_2}{\sigma_1 \cdot \sigma_2}.$$

- *Question:* what is the resulting range of $E[x_1 \cdot x_2]$?

11. Taking Correlation into Account: First Result

- *Given:*
  - $[\underline{x}_1, \overline{x}_1]$ and $[\underline{x}_2, \overline{x}_2]$ are given intervals,
  - $E_1 \in [\underline{x}_1, \overline{x}_1]$ and $E_2 \in [\underline{x}_1, \overline{x}_1]$ are given numbers, and
  - ρ is a given number.
- *Find:* the range $[\underline{E}, \overline{E}]$ of possible values $E[x_1 \cdot x_2]$ for all possible distributions for which:
  - x_1 is located in $[\underline{x}_1, \overline{x}_1]$, and x_2 is located in $[\underline{x}_2, \overline{x}_2]$;
  - $E[x_1] = E_1$, and $E[x_2] = E_2$; and
  - $\rho[x_1, x_2] = \rho$.
- *Solution:*
  - for $\rho \ge 0$: $[\underline{E}, \overline{E}] = [E_1 \cdot E_2, E_1 \cdot E_2 + \rho \cdot \sigma]$;
  - for $\rho \le 0$: $[\underline{E}, \overline{E}] = [E_1 \cdot E_2 + \rho \cdot \sigma, E_1 \cdot E_2]$.

12. Taking Correlation into Account: Second Result

- *Given:*
  - $[\underline{x}_1, \overline{x}_1]$ and $[\underline{x}_2, \overline{x}_2]$ are given intervals;
  - $E_1 \in [\underline{x}_1, \overline{x}_1]$ and $E_2 \in [\underline{x}_1, \overline{x}_1]$ are given numbers;
  - $[\underline{\rho}, \overline{\rho}]$ is a given interval.
- *Find:* the range $[\underline{E}, \overline{E}]$ of possible values $E[x_1 \cdot x_2]$ for all possible distributions for which:
  - x_1 is located in $[\underline{x}_1, \overline{x}_1]$, and x_2 is located in $[\underline{x}_2, \overline{x}_2]$;
  - $E[x_1] = E_1$, and $E[x_2] = E_2$; and
  - $\rho[x_1, x_2] \in [\underline{\rho}, \overline{\rho}]$.
- *Solution:*
  - for $0 \le \underline{\rho}$: $[\underline{E}, \overline{E}] = [E_1 \cdot E_2, E_1 \cdot E_2 + \overline{\rho} \cdot \sigma]$;
  - for $\overline{\rho} \le 0$: $[\underline{E}, \overline{E}] = [E_1 \cdot E_2 + \underline{\rho} \cdot \sigma, E_1 \cdot E_2]$;
  - for $\underline{\rho} \le 0 \le \overline{\rho}$: $[\underline{E}, \overline{E}] = [E_1 \cdot E_2 + \underline{\rho} \cdot \sigma, E_1 \cdot E_2 + \overline{\rho} \cdot \sigma]$.

Title Page
Page 13 of 16
Go Back
Full Screen

13. Second Problem: How to Measure Loss of Privacy

- *Measuring loss of privacy is important:* to compare different privacy protection schemes.
- *Natural idea:* gauge the loss of privacy by the resulting worst-case financial loss.
- *Example:* the effect of a person's blood pressure x on this person's insurance payments:
  - let $f(x)$ be average medical expenses for a person with blood pressure x; let α be investment profit;
  - in case of privacy, the insurance payments are
  $$r = (1+\alpha)\cdot E[f(x)];$$
  - if a person's blood pressure is revealed as x_0, with $f(x_0) > E[f(x)]$, then the payments are higher:
  $$r_0 = (1+\alpha)\cdot f(x_0) > r = (1+\alpha)\cdot E[f(x)].$$

14. Loss of Privacy: Main Result

- *Situation:* we knew that $x \in [L, U]$, now we learned that $x \in [l, u] \subseteq [L, U]$.
- *Description:* we knew that $P \in \mathcal{P}$ (all distributions located on $[L, U]$), now we know that $P \in \mathcal{Q}$ (all distributions located on $[l, u]$).
- *Definition:* let $M > 0$ be a real number. The *amount of privacy* $A(\mathcal{P})$ related to $\mathcal{P}$ is the largest possible value of the difference $F(x_0) - \int \rho(x) \cdot F(x)\,dx$ over:
  - all possible values x_0,
  - all possible probability distributions $\rho \in \mathcal{P}$, and
  - all possible f-s $F(x)$ for which $|F'(x)| \le M$ for all x.
- *Result:* the *relative loss of privacy* $\dfrac{A(\mathcal{P}) - A(\mathcal{Q})}{A(\mathcal{P})}$ is equal to $1 - \dfrac{u - l}{U - L}$.

15. Acknowledgments

This work was supported in part by:

- NASA under cooperative agreement NCC5-209,
- NSF grants EAR-0225670 and DMS-0532645,
- Star Award from the University of Texas System, and
- Texas Department of Transportation grant No. 0-5453.