What If We Use Ellipsoids Instead of p-Boxes?

Miguel A. Flores, Joshua Salas, Javier A. Venegas, and Vladik Kreinovich

Department of Computer Science, University of Texas at El Paso
500 W. University, El Paso, Texas 79968, USA
maflores32@miners.utep.edu, jsalas19@miners.utep.edu,
javenegas8@miners.utep.edu, vladik@utep.edu

1. Formulation of the problem: p-boxes and their limitations

- Our knowledge of the world comes from measurements.
- Measurement are never 100% accurate, so we always have uncertainty.
- Uncertainty means, in particular, that for each quantity x:
  - instead of its exact value,
  - we only know the range $[\underline{x}, \overline{x}] = [\widetilde{x} - \Delta, \widetilde{x} + \Delta]$ of possible values, where $\widetilde{x}$ is known approximation and Δ is its accuracy.
- Ideally, we should also know how frequent are different values from this range, i.e., what are the probabilities of different values.
- These probabilities also come from observations and measurements, so we also know them with some uncertainty.
- What is the best way to describe this uncertainty?

2. Formulation of the problem: p-boxes and their limitations (cont-d)

- There are many ways to describe a probability distribution; we can describe it:
  - by its probability density function (pdf) $f(x)$,
  - by its cumulative distribution function (cdf)
$$F(x) = \text{Prob}(X \le x),$$
  - by its moments $E[X^k] = \int x^k \cdot f(x)\,dx$, etc.
- Out of these descriptions, only cdf is universal; e.g.:
  - pdf does not work when we have a distribution located in a single point with probability 1,
  - moments do not work, e.g., for Cauchy distribution with
$$f(x) \sim 1/(1+x^2).$$
- So, a reasonable way to describe uncertainty with which we know probabilities is to describe uncertainty with which we know $F(x)$.

3. Formulation of the problem: p-boxes and their limitations (cont-d)

- As we have mentioned, the basic way to describe uncertainty in a quantity is to have an interval.
- In line with this idea, the mainstream way to represent the uncertainty with which we know $F(x)$ is to have an interval

$$[\underline{F}(x), \overline{F}(x)] = [\widetilde{F}(x) - \Delta(x), \widetilde{F}(x) + \Delta(x)] \text{ for each } x.$$

- This representation is known as the *probability box* (*p-box*, for short).
- p-boxes have many successful applications.
- However, they have a problem:
  - a p-box means a large number of intervals, and
  - it is known that, in general, processing data with many intervals is NP-hard.
- How can we make corresponding computations faster?

4. Idea

- Normal (Gaussian) distributions are ubiquitous.
- Suppose that we have several independent Gaussian distributions describing measurement uncertainty $\widetilde{x}_i \neq x_i$:
  - with 0 mean and
  - with standard deviation σ_i.
- Then the probability density depends on the expression

$$J \stackrel{\text{def}}{=} \sum (x_i - \widetilde{x}_i)^2 \cdot \sigma_i^{-2}.$$

- Thus, if we dismiss improbable regions, with very small probability, we thus get a set $J \leq C$ for some constant C.
- This set is an ellipsoid.
- Its advantage is that we have only one inequality constraint, and thus computations are often much faster.

5. Idea (cont-d)

- A natural idea is, instead of intervals for $F(x)$, to consider an ellipsoid, which in continuous case takes the form
$$J \stackrel{\text{def}}{=} \int (F(x) - \widetilde{F}(x))^2 \cdot \Delta^{-2}(x)\, dx \le C.$$
- In this case, a simple Lagrange multiplied method can help find bounds on many statistical characteristics.
- For example:
  - for the mean $E[X] = T - \int_{-\infty}^{T} F(x)\, dx$,
  - this method means optimizing the expression $E[X] + \lambda \cdot (J - C)$ for an appropriate Lagrange multiplier λ.
- Differentiating this expression with respect to $F(x)$ and equating derivative to 0, we get
$$-1 + \lambda \cdot 2 \cdot (F(x) - \widetilde{F}(x)) \cdot \Delta^{-2}(x) = 0.$$
- So $1 = \lambda \cdot 2 \cdot (F(x) - \widetilde{F}(x)) \cdot \Delta^{-2}(x)$.

6. Idea (cont-d)

- Dividing both sides by the coefficient at $F(x) - \widetilde{F}(x)$, we get

$$F(x) - \widetilde{F}(x) = c \cdot \Delta^2(x), \text{ where we denoted } c \stackrel{\text{def}}{=} 1/(2\lambda).$$

- Thus $F(x) = \widetilde{F}(x) + \Delta F(x)$, where $\Delta F(x) = c \cdot \Delta^2(x)$.
- The constant c must be determined from the condition

$$J = \int (\Delta F(x))^2 \cdot \Delta^{-1}(x)\, dx = C.$$

- So $c = \sqrt{C/(\int \Delta^2(x)\, dx)}$.
- Hence, the range of the mean is $[\widetilde{E} - \Delta, \widetilde{E} + \Delta]$, where $\widetilde{E}$ is computed based on $\widetilde{F}(x)$, and $\Delta = \sqrt{C \cdot \int \Delta^2(x)\, dx}$.
- Similar explicit expressions can be obtained for second moment and for other statistical characteristics.

7. Acknowledgments

This work was supported in part by:

- National Science Foundation grants 1623190, HRD-1834620, HRD-2034030, and EAR-2225395;
- AT&T Fellowship in Information Technology;
- program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478, and
- a grant from the Hungarian National Research, Development and Innovation Office (NRDI).