How Accurately Can We Determine the Coefficients: Case of Interval Uncertainty

Michal Cerny[1] and Vladik Kreinovich[2]
[1]Faculty of Informatics and Statistics
University of Economics
Prague, Czech Republic, cernym@vse.cz
[2]University of Texas at El Paso
El Paso TX 79968, USA
vladik@utep.edu

1. Need to Determine the Dependence Between Different Quantities

- One of the main objectives of science is to find the dependencies $y = f(x_1, \ldots, x_n)$ between values of different quantities.
- Once we know such dependencies, we can then use them to predict the future values of different quantities.
- E.g., Newton's laws describe how the acceleration y of a celestial body depends on the location and masses x_i of this and other bodies.
- Thus, these laws enable us to predict how these bodies will move.
- Another important case is when we want to estimate the value of a quantity y which is difficult to directly measure.
- In such cases, it is often possible to find easier-to-measure quantities $x_1, \ldots, x_n$ knowing which we can determine y.
- For example, it is difficult to directly measure the distance y between two faraway locations on the Earth.
- However, we can determine this distance if we use astronomical observations – or, nowadays, signals from the GPS satellites.

2. How Can We Determine This Dependence

- In some cases, we can use the known physical laws to derive the desired dependence.
- However, in most other cases, this dependence needs to be determined empirically:
  - we measure the values $x_1, \ldots, x_n$, and y in different situations;
  - we use the measurement results to find the desired dependence.
- In many cases, we know the general form of the desired dependence.
- So, we know that $y = F(x_1, \ldots, x_n, c_0, c_1, \ldots, c_m)$, where F is known function, and the coefficients c_i need to be determined.
- For example, we may know that the dependence is linear, i.e., that

$$y = c_0 + c_1 \cdot x_1 + \ldots + c_n \cdot x_n.$$

- This is a typical situation:
  - when the values x_i have a narrow range $[\underline{X}_i, \overline{X}_i]$,
  - we can expand the function $f(x_1, \ldots, x_n)$ in Taylor series over $x_i - \underline{X}_i$ and ignore quadratic (and higher order) terms in this expansion.

3. Need to Take uncertainty into Account – in Particular, Interval Uncertainty

- Measurements are never absolutely accurate: the measurement result $\widetilde{x}$ is, in general, different from the actual (unknown) value x.
- In many practical situations, the only information that we have about the measurement error $\Delta x \stackrel{\text{def}}{=} \widetilde{x} - x$ is the upper bound Δ on its absolute value: $|\Delta x| \le \Delta$.
- In this case, after each measurement, the only information that we have about the actual value x is that this value is somewhere in the interval
$$[\widetilde{x} - \Delta, \widetilde{x} + \Delta].$$
- Because of this fact, this case is known as the case of *interval uncertainty.*
- There exist many algorithms for dealing with such uncertainty.

4. Measurement Uncertainty Leads to Uncertainty in Coefficients

- We can only measure the values x_i and y with some uncertainty.
- So, we can therefore only determine the coefficients c_i with some uncertainty.
- It is therefore important to determine how accurate are the values c_i that we get as a result of these measurements.

5. Which Uncertainty Should Be Taken into Account

- Strictly speaking, there are measurement uncertainties both:
  - when we measure easier-to-measure quantities $x_1, \ldots, x_n$, and
  - when we measure the desired difficult-to-measure quantity y.
- Usually, y is much more difficult to measure than x_i.
- Thus, the measurement errors Δy corresponding to measuring y are much larger than the measurement errors of measuring x_i.
- So, we can usually safely ignore the measurement errors of measuring x_i and assume that these values are known exactly.
- So, in the linear case, we can safely assume that for each measurement k:
  - we know the exact values $x_1^{(k)}, \ldots, x_n^{(k)}$, but we only know $y^{(k)}$ with uncertainty,
  - i.e., based on the measurement result $\widetilde{y}^{(k)}$ and the known accuracy $\Delta > 0$, we know that the actual value

$$\underline{y}^{(k)} = \widetilde{y}^{(k)} - \Delta \le y^{(k)} = c_0 + \sum_{i=1}^{n} c_i \cdot x_i^{(k)} \le \overline{y}^{(k)} = \widetilde{y}^{(k)} + \Delta.$$

6. Once we Perform the Measurements, we can feasibly find the accuracy

- Suppose that have the measurement results.
- Then, we can find the bounds on each of the coefficients c_i by solving the following linear programming problems:
  - minimize (maximize) c_i
  - under the constraints that for all the measurements $k = 1, \ldots, K$:

$$\underline{y}^{(k)} \le c_0 + \sum_{i=1}^{n} c_i \cdot x_i^{(k)} \le \overline{y}^{(k)}$$

7. Remaining Question

- Before we start spending our resources on measurements, it is desirable to check how accurately we can determine the coefficients c_i.
- If the resulting accuracy is not enough for us – then we should not waste time performing the measurements.
- Instead we should invest in a more accurate y-measuring instrument.
- Of course, we can answer the above question by simulating measurement errors.
- However, it would be great to have simple analytical expressions that would not require extensive simulation-related computations.

8. Discussion

- The range of each physical quantity is usually bounded:
  - coordinates of Earth locations are bounded by the Earth's size,
  - velocities are bounded by the speed of light, etc.
- Thus, we can safely assume that for each variable x_i, we know the interval $[\underline{X}_i, \overline{X}_i]$ of its possible values.
- Thus, we arrive at the following formulation of the problem.

9. Formulation of the Problem

- Let us assume that we are given the value $\Delta > 0$ and n intervals $[\underline{X}_i, \overline{X}_i]$, $i = 1, 2, \ldots, n$.
- We say that a tuple $(\Delta c_0, \Delta c_1, \ldots, \Delta c_n)$ is *within the possible uncertainty* if for each tuple c_i and for all $x_i \in [\underline{X}_i, \overline{X}_i]$, we have:

$$|y' - y| \le \Delta, \text{ where } y \stackrel{\text{def}}{=} c_0 + \sum_{i=1}^{n} c_i \cdot x_i \text{ and } y' \stackrel{\text{def}}{=} c_0' + \sum_{i=1}^{n} c_i' \cdot x_i,$$

$$\text{where } c_i' \stackrel{\text{def}}{=} c_i + \Delta c_i.$$

- Because of the measurement uncertainty, after the measurement, the range of possible values of the corresponding quantity x is $[\widetilde{x} - \Delta, \widetilde{x} + \Delta]$.
- It may be therefore convenient to represent the intervals $[\underline{X}_i, \overline{X}_i]$ in the same form, as $[\underline{X}_i, \overline{X}_i] = [\widetilde{X}_i - \Delta_i, \widetilde{X}_i + \Delta_i]$.
- For this, we need to take $\widetilde{X}_i = \dfrac{\underline{X}_i + \overline{X}_i}{2}$ and $\Delta_i = \dfrac{\overline{X}_i - \underline{X}_i}{2}$.

10. Main Results

- For each Δ and $[\underline{X}_i, \overline{X}_i]$, a tuple $(\Delta c_0, \Delta c_1, \ldots, \Delta c_n)$ is within the possible uncertainty if and only if $|\Delta c_0'| + \sum\limits_{i=1}^{n} |\Delta c_i| \cdot \Delta_i \le \Delta$.
- Here $\Delta c_0' \stackrel{\text{def}}{=} \Delta c_0 + \sum\limits_{i=1}^{n} \Delta c_i \cdot \widetilde{X}_i$.
- Based on this result, we can find bounds on each of the coefficient Δc_i:

$$\Delta c_i \in \left[-\frac{\Delta}{\Delta_i}, \frac{\Delta}{\Delta_i}\right].$$

- Thus:
  - if we can measure y with accuracy Δ, and
  - we can use any value x_i from the interval $[\widetilde{X}_i - \Delta_i, \widetilde{X}_i + \Delta_i]$,
  - then we can determine the coefficient c_i that describes the dependence of y on x_i with accuracy $\dfrac{\Delta}{\Delta_i}$.
- This accuracy is what we can *guarantee* if we perform sufficiently many measurements.
- Even with a primitive y-measuring device, we can get lucky and get even absolutely accurate values of c_i.

11. Main Results (cont-d)

- For example, for the actual value $y = c_0$, measurement results can be $c_0 - \Delta$ and $c_0 + \Delta$.
- Thus, the actual value is in $[c_0 - 2\Delta, c_0] \cap [c_0 + 2\Delta] = \{c_0\}$.
- When 0 is a possible value of each variable x_i, then possible values of Δc_0 form the interval $[-\Delta, \Delta]$.
- When for some i, $0 \notin [\underline{X}_i, \overline{X}_i]$, then all values $\Delta c_0 \in [-\Delta, \Delta]$ are still possible.
- However, some values outside this interval are possible too.
- For $n = 1$, the range of possible values of Δc_0 is $[-\Delta', \Delta']$, where $\Delta' = \Delta + \dfrac{\Delta}{\Delta_1} \cdot m_1$ and:

  – $m_1 = 0$ if $0 \in [\underline{X}_1, \overline{X}_1]$;

  – $m_1 = \underline{X}_1$ if $\underline{X}_1 > 0$, and

  – $m_1 = |\overline{X}_1|$ when $\overline{X}_1 < 0$.

12. Discussion: Probabilistic Case

- Often, in addition to the upper bound, we also have some information about the probability of different values of measurement error.
- In such cases, it is convenient to represent the measurement error as the same of two components:
  - its mean, which is called *systematic error*, and
  - the difference between the measurement error and its mean, which is called the *random error*.
- Usually, we know the upper bound Δ_i on the absolute value of the systematic error, and we know some characteristics of the random error.
- With what accuracy can we then determine c_i? Interestingly, we get the same answer as in the interval case.
- Indeed, if for the same example, we measure y several times, the arithmetic average of the measurement results tends to its mean value.
- So, it tends to the actual value y plus the systematic error s_i.
- Thus, in measurement results obtained this way, the random error disappears and we get, in effect, the interval case.

13. What If We Consider Quadratic Dependencies

- So far, we considered the case when we could ignore quadratic and higher order terms.
- Thus, we assumed that the dependence of y on x_i is linear.
- If we want a more accurate description, we consider quadratic terms:

$$y = c_0 + \sum_{i=1}^{n} c_i \cdot x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} \cdot x_i \cdot x_j.$$

- In this case, even for a single tuple $(\Delta c_i, \Delta c_{ij})$, it is NP-hard (= intractable) to check whether this tuple is within the accuracy.
- So, it is NP-hard to check whether for all $x_i \in [\underline{X}_i, \overline{X}_i]$, we have

$$|\Delta y| = \left|\Delta c_0 + \sum_{i=1}^{n} \Delta c_i \cdot x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} \Delta c_{ij} \cdot x_i \cdot x_j\right| \le \Delta.$$

14. What If We Have an Ellipsoid

- Instead of requiring that possible values of $(x_1, \ldots, x_n)$ form a box, we can consider the case when this set is an ellipsoid.
- In this case, the range of a linear expression $\Delta c_0 + \sum\limits_{i=1}^{n} \Delta c_i \cdot x_i$ can also be explicitly computed.
- Thus, we also have an analytical expression describing tuples $(\Delta c_0, \Delta c_1, \ldots)$ which are within the possible uncertainty.

15. What If We Also Have Relative Measurement Error

- So far, we assumed that the measurement accuracy Δ is the same for all y.
- In measurement terms, this means that we have an *absolute* error.
- In practice, we often also have *relative* error component, in which cases the upper bound $\Delta(y)$ on the y-measurement error depends on y as:

$$\Delta(y) = \Delta_0 + c \cdot |y| \text{ for some } \Delta_0 > 0 \text{ and } c > 0.$$

- Once we have measurement results, we can still use linear programming to find the accuracy with which we can determine the coefficients c_i.
- However, it is not clear how to come up with an analytical expression for the tuples $(\Delta c_0, \Delta c_1, \ldots)$ within the possible uncertainty.

15. Acknowledgments

- This work was supported in part by the US National Science Foundation grant HRD-1242122 (Cyber-ShARE Center of Excellence).
- M. Cerny acknowledges the support of the Czech Science Foundation (project 19-02773S).