From Quantifying and Propagating Uncertainty to Quantifying and Propagating Both Uncertainty and Reliability: Practice-Motivated Approach to Measurement Planning and Data Processing

Niklas R. Winnewisser[1], Michael Beer[1], Olga Kosheleva[2], and Vladik Kreinovich[2]

[1]Leibniz University Hannover, Hannover, Germany
{winnewisser,beer}@irz.uni-hannover.de
[2]University of Texas at El Paso, El Paso, Texas 79968, USA
{olgak,vladik}@utep.edu

1. Data processing is ubiquitous

- The main objectives of science and engineering are:
  - to know the current state of the world,
  - to predict what will happen, and
  - to make sure – by using appropriate devices and/or controls – that the future world is as beneficial for us as possible.
- Knowing the current state of the world means, in particular, to know the values of the physical quantities that characterize this state.
- Some of these quantities we can directly measure, in the sense that there is a measuring instrument that returns the value of this quantity.
- For example, we can measure the current temperature by using a thermometer.
- We can directly measure the wind speed, the distance between two nearby buildings, etc.
- Other quantities y we cannot measure directly in this sense.

2. Data processing is ubiquitous (cont-d)

- E.g., we cannot directly measure the temperature on the surface of the Sun or the distance from the Earth to the Sun.
- Since we cannot measure these quantities directly, we have to measure them *indirectly*.
- We find easier-to-directly-measure auxiliary quantities $x_1, \ldots, x_n$ that are related to y by a known relation $y = f(x_1, \ldots, x_n)$.
- This relation can be known from some physical theory and/or it can be obtained from empirical data – e.g., by using machine learning.
- We measure the values of these auxiliary quantities x_i.
- We get an estimate for the desired quantity y by applying the algorithm $f(x_1, \ldots, x_n)$ to the results of measuring the quantities

$$x_1, \ldots, x_n.$$

- And, of course, at the present moment of time, we cannot directly measure the future value of a physical quantity y.

3. Data processing is ubiquitous (cont-d)

- These future values must also be measured indirectly, by following the same steps.
- In general, this procedure – of applying an algorithm to measurement results – is known as *data processing*.
- In many cases, the data processing algorithm consists of several distinct stages, each processing:
  - the measurement results
  - and/or the results of preceding stages.
- This is how, for example, deep neural networks handle data.

4. Uncertainty and reliability are ubiquitous

- Most information about the real world comes – directly or indirectly – from measurements.
- Measurements are never 100% accurate.
- For each physical quantity x:
  - the measurement results $\widetilde{x}$ is, in general, different from
  - the actual (unknown) value x of the corresponding quantity.
- In most practical situation, the difference $\Delta x \stackrel{\text{def}}{=} \widetilde{x} - x$ is reasonably small.
- This difference is usually called the *measurement uncertainty*.
- Sometimes, a measuring instrument malfunctions.
- It generates a result which is far off from the actual value of the corresponding quantity.

5. Uncertainty and reliability are ubiquitous (cont-d)

- The probability of the measuring instrument functioning well is known as its *reliability*.
- From the purely *mathematical* viewpoint:
  - outliers corresponding to malfunctioning can be viewed
  - as part of the overall probability distribution of measurement uncertainty.
- However, *in practice*, when manufacturers of measuring instruments provide the probabilities of different values of Δx:
  - they usually mean *conditional* probabilities under the condition that we only consider small values Δx,
  - and ignore much larger outliers.

6. Measurement uncertainty affects the results of data processing

- When we process data:
  - we apply an appropriate algorithm $y = f(x_1, \ldots, x_n)$ to the results $\widetilde{x}_1, \ldots, \widetilde{x}_n$ of measuring the quantities $x_1, \ldots, x_n$,
  - i.e., we compute the value $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.
- The measurement results $\widetilde{x}_i$ are, in general, different from the actual values x_i.
- So, the value $\widetilde{y}$ is, in general, different from the ideal value $y = f(x_1, \ldots, x_n)$ – that we would have got if we knew the exact values x_i.

7. Measurement uncertainty affects the results of data processing (cont-d)

- By the way, the relation $y = f(x_1, \ldots, x_n)$ may be only approximate, so our estimate may be even more different from the true value y.
- It is therefore desirable to understand:
  - how the measurement uncertainty propagates through the data processing algorithm,
  - i.e., what is the resulting uncertainty $\Delta y \stackrel{\text{def}}{=} \widetilde{y} - y$.

8. How to take uncertainty into account: what s known and what is the remaining problem

- Several methods have been developed in measurement theory to take uncertainty into account, both:
  - when we plan measurements and
  - when we process data.
- The problem is that most of these methods do not take into account the fact that measurement instruments are also not perfectly reliable.
- Sometimes, they malfunction – and generate results which are far off from the actual value of the corresponding quantity.
- It is therefore important to also take into account this finite reliability when planning measurements and processing data.

9. How to take uncertainty into account: what s known and what is the remaining problem (cont-d)

- In this talk, we describe several practical scenarios, depending on what information we have.
- For each scenario, we show how to take into account both uncertainty and reliability, both:
  - when planning experiments and
  - when processing data.

10. First scenario: journey to the unknown

- Let us start with the case when we have no prior information at all.
- This is typical when we design a new state-of-the-art measuring instrument:
  - a new more powerful space telescope,
  - a new more powerful particle accelerator, etc.
- In many such situations, we do not fully know what to expect.
- We do not fully know what exactly objects we will measure.
- We do not know what uncertainty level (and what reliability level) we will need.

11. First scenario: journey to the unknown (cont-d)

- But we still design the corresponding instrument, because in the past, similar instruments led to important discoveries.
- In this case:
  - if within a given cost limit, we have several designs,
  - a natural idea is to select the design that will provide us with the largest amount of information.

12. Second scenario: working by specifications

- The second scenario is the opposite to the first one.
- We know exactly what uncertainty level (and what reliability level) we need.
- We just need to find the least costly way to achieve these specifications.

13. General case

- In practice:
  - we rarely know nothing about the appropriate values of accuracy and reliability, and
  - we rarely have full information about them.
- Such situations are well-studied in decision theory.
- In decision theory, it is known that:
  - decisions of a rational decision maker – who, e.g., prefers A to C if he/she prefers A to B and B to C
  - can be described by maximizing the expected value of a special function called *utility*.
- This is the framework that we consider in this talk.

14. General case (cont-d)

- This framework can be divided into two scenarios, that we will call third and fourth:
- In the third scenario, we consider a general optimization problem without any constraints.
- The fact that some values are undesirable is described not by a constraint, but by a highly negative utility assigned to these situations.
- In the fourth – rather typical scenario – we consider a limited problem, in which:
  - we only take into account a few quantities, and
  - the analysis of all other aspects is described in terms of constraints.
- For example:
  - when we design a chemical plant,
  - we need to satisfy a constraint that the concentration of undesired chemicals in the air should not exceed some threshold.

15. General case (cont-d)

- This threshold that has already been determined by taking into account potential benefits and limitations of these types of plants.
- *Comment:* decision theory described decisions by *ideal* decision makers.
- It is well known that our *actual* decisions differ from this idealized framework.
- It is therefore desirable to extend our results to realistic non-utility-based decision techniques.

16. How do we describe uncertainty

- In the ideal case, we should know:
  - which values of measurement uncertainty Δx are possible and
  - with what frequency different possible values appear,
  - what is the probability distribution of the measurement uncertainty.
- In practice, often, we only have the upper bound Δ on the absolute value $|\Delta x|$ of the measurement uncertainty: $|\Delta x| \leq \Delta$.
- We will call this value the *accuracy* of the measuring instrument.
- Knowing this upper bound is a must: if we do not know any upper bound, this means that:
  - no matter what value we measure,
  - the actual value can be as far off from it as mathematically possible.
- This is not what we would call a measuring instrument.

17. How do we describe uncertainty (cont-d)

- The mean value of the measurement error can be determined:
  - after several comparison with the "standard" (= much more accurate) measuring instrument,
  - as the arithmetic average of the measurement uncertanties.
- Once we know this mean, we can subtract this value – known as *bias* – from all measurement results.
- Thus, we can conclude that the mean becomes 0.
- We can also estimate the second moment.
- Since the mean is 0, it is equal to the variance V, or, which is equivalent, estimate the standard deviation $\sigma \stackrel{\text{def}}{=} \sqrt{V}$.
- Usually, the measurement uncertainty comes as a joint effect of many relatively small reasonably independent factors.
- In this case, according to the Central Limit Theorem, the resulting distribution is close to Gaussian (normal).

18. How do we describe uncertainty (cont-d)

- Of course, in reality, there may be dependence between factors, and some of these factors may not be that small.
- However, empirical data shows that indeed, for the majority of measuring instruments, the probability distribution of measurement uncertainty is close to normal.
- For a normal distribution, with very high confidence, all the values of the measurement uncertainty are located within an interval

$$[-k \cdot \sigma, k \cdot \sigma].$$

- For $k = 2$ we have confidence 95%.
- For $k = 3$, we have confidence 99.9%.
- For $k = 6$, we have confidence $1 - 10^{-8}$.
- It is then natural to identify Δ as the upper bound of this interval:

$$\Delta = k \cdot \sigma.$$

19. How do we describe uncertainty (cont-d)

- The actual distribution may be different from normal.
- However, for many other distributions, we still have a similar relation $\Delta = k \cdot \sigma$ for some constant k.
- So this is what we will assume in this talk.

20. How to estimate the amount of information

- In the discrete case, we have finitely many possible outcomes.
- Then, a natural measure of the amount of information is:
  - the average number of "yes"-"no" questions
  - that we need to ask to uniquely determine the outcome.
- If we know the probabilities $p_1, \ldots, p_N$ of different outcomes, then the average number of questions is equal to Shannon's entropy

$$S \stackrel{\text{def}}{=} -\sum_{i=1}^{N} p_i \cdot \log_2(p_i).$$

- Sometimes, we do not know the exact values of the probabilities p_i, i.e., several different probability distributions are possible.
- In this case, a natural idea is to take the largest amount of information corresponding to all possible distributions.

21. How to estimate the amount of information (cont-d)

- It is known that if we have no information about the probabilities at all, the largest entropy corresponds to the uniform distribution

$$p_1 = \ldots = p_N.$$

- In the continuous case, we cannot determine the actual value by asking a finite number of "yes"-"no" questions, since:
  - this way we only get finitely many possible combinations of answers, while
  - there are infinitely many real numbers within the interval $[\underline{x}, \overline{x}]$ of possible values of the measured quantity x.
- What we *can* do is determine x with some accuracy δ.

22. How to estimate the amount of information (cont-d)

- This means we should have several values x', x'', ..., so that each value from the range $[\underline{x}, \overline{x}]$:
  - should be close to one of these values,
  - i.e., should be in one of the intervals $[x' - \delta, x' + \delta]$, $[x'' - \delta, x'' + \delta]$ of width 2δ.
- In other words, we divide the range $[\underline{x}, \overline{x}]$ into subintervals of width 2δ and take into account probabilities $p_1, \ldots, p_N$, of x being in different subintervals.
- *Comment.*
  - The smallest amount of information corresponds to the uniform distribution.
  - But this does *not* mean that the actual distribution is uniform,.
  - In the no-information case, we can have many different probability distributions on thr interval.

23. What if we have several measurements of the same quantity: what is the resulting uncertainty

- Suppose that we have m results $\widetilde{x}_1, \ldots, \widetilde{x}_m$ of measuring the same quantity x by different measuring instruments.
- All these measurements have the mean measurement uncertainty 0.
- For each measuring instrument, we know the corresponding standard deviations σ_i.
- It is then desirable to combine these results into a single more accurate estimate $\widetilde{x} = f(\widetilde{x}_1, \ldots, \widetilde{x}_m)$.
- We want to find a combination which is the most accurate.
- I.e., for which the standard deviation of the resulting uncertainty Δx is the smallest possible:

$$\Delta x = f(\widetilde{x}_1, \ldots, \widetilde{x}_m) - f(x_1, \ldots, x_m) = f(\widetilde{x}_1, \ldots, \widetilde{x}_m) -$$

$$f(\widetilde{x}_1 - \Delta x_1, \ldots, \widetilde{x}_m - \Delta x_m)$$

24. What if we have several measurements of the same quantity: what is the resulting uncertainty (cont-d)

- We can:
  - expand the above expression in Taylor series in terms of Δx_i and
  - take into account that measurement uncertainty is usually reasonably small.
- So terms which are quadratic or of higher order in terms of this uncertainty can be, in the first approximation, safely ignored.
- E.g., for a not very accurate measurement with 10% accuracy, the square of this value is $1\% \ll 10\%$.
- This linearization is a usual techniques in physics.
- Thus, we get $\Delta y = c_1 \cdot \Delta x_1 + \ldots + c_m \cdot \Delta x_m$ for some coefficient c_i.
- If all the instruments show the same result, this is the result we should return.
- This means, in particular, that $\sum c_i = 1$.

25. What if we have several measurements of the same quantity: what is the resulting uncertainty (cont-d)

- Uncertainty of different measurements usually comes from different independent causes.
- For the sum of independent random variables, the variance is equal to the sum of the variances.
- So, for the variance σ^2 of Δx, we have

$$\sigma^2 = c_1^2 \cdot \sigma_1^2 + \ldots + c_m^2 \cdot \sigma_m^2.$$

- We want to generate the most accurate estimate, i.e., we want to minimize σ^2 under the above constraint $\sum c_i = 1$.
- To solve this constraint optimization problem, we can use the Lagrange multiplier method.
- As a result, we get $\sum c_i^2 \cdot \sigma_i^2 + \lambda \cdot (\sum c_i - 1) \to \min$; thus, by differentiating, $2c_i \cdot \sigma_i^2 + \lambda = 0$, so $c_i = \text{const} \cdot \sigma_i^{-2}$.
- By using the equation $\sum c_i = 1$, we conclude that $c_i = \sigma_i^{-2} / (\sum \sigma_j^{-2})$.

26. What if we have several measurements of the same quantity: what is the resulting uncertainty (cont-d)

- Substituting these values into the formula for σ^2, we get

$$\sigma^2 = \left(\sum \sigma_i^{-2}\right)/\left(\sum \sigma_i^{-2}\right)^2, \text{ i.e., } \sigma^2 = 1/\left(\sum \sigma_i^{-2}\right) \text{ and}$$

$$\sigma^{-2} = \sum_{i=1}^{m} \sigma_i^{-2}.$$

- We assumed that the bounds Δ_i are proportional to the standard deviations σ_i.
- Thus, for the overall bound Δ, we get a similar formula

$$\Delta^{-2} = \sum_{i=1}^{m} \Delta_i^{-2}.$$

27. What if we have several measurements of the same quantity: what is the resulting reliability

- Suppose that we have m measurements of the same quantity.
- In each measurement i, the probability that we have an outlier is p_i.
- In this case, the only case when we miss the actual value is when all m measurement are outliers.
- It is reasonable to assume that the measurements are independent.
- So, the probability that all m measurements are outliers is equal to the product of the given probabilities: $p = p_1 \cdot \ldots \cdot p_m$.

28. First Scenario: Journey to the Unknown

- Now we are ready to start analyzing specific scenarios.
- Let us start with the first scenario, when:
  - we do not have any information about probabilities and
  - we are, thus, interested in getting as much information as possible.
- In this scenario, we may need to answer the following natural questions.
- Sometimes, we can only employ one measuring instrument.
- In this case, it is desirable to select the most informative instrument.
- In other cases, we can, in principle, employ several measuring instruments.
- The only limitation is the overall measurement cost.
- In this case, it is desirable to find the arrangement that – within the given cost – will bring us the maximum amount of information.

29. First Scenario: Journey to the Unknown (cont-d)

- In yet other cases, our goal is to extract a certain amount of information.
- We want to find the arrangement with the minimal cost that will provide the required amount of information.
- Let us formulate all these problems in precise terms.
- Then, we will be able to use usual numerical techniques to solve the corresponding problems.
- To be able to formulate these problems, let us describe what is known.
- For each type of measuring instrument, let us denote its accuracy by Δ_i, its probability of an outlier by p_i, and the cost of each measurement by c_i.
- To formalize the second and third questions, let us also denote the number of instruments of type i that we will use by n_i.

30. Preliminary analysis

- Suppose that we have a measuring instrument with accuracy Δ and outlier probability p.
- What is the number of bits that we are still missing after a single measurement by this instrument?
- To answer this question, let us pick some value δ. Then:
  - with probability p, the actual value is somewhere in the original range $[\underline{x}, \overline{x}]$ of width $w \stackrel{\text{def}}{=} \overline{x} - \underline{x}$, and
  - with probability $1-p$, it is in the interval $[\widetilde{x}-\Delta, \widetilde{x}+\Delta]$ of width 2δ.
- As we mentioned earlier, to find the largest amount of information, we need to use uniform distribution.
- So, in the range $[\widetilde{x}-\Delta, \widetilde{x}+\Delta]$ of width 2Δ, we have $(2\Delta)/(2\delta) = \Delta/\delta$ intervals with probability $(1-p)/(\Delta/\delta)$.

31. Preliminary analysis (cont-d)

- Here, $p \ll 1$, so in the first approximation, $1 - p \approx 1$, and these intervals have probability $1/(w/(\Delta/\delta)$.
- The remaining part of the range $[\underline{x}, \overline{x}]$ is of width $\overline{x} - \underline{x} - 2\Delta$.
- Here, $\Delta \ll \overline{x} - \underline{x}$, so in the first approximation, we can safely assume that this part has width $w = \overline{x} - \underline{x}$.
- In this part, we gave $w/(2\delta)$ intervals of probability

$$p/(w/(2\delta)) = (2\delta \cdot p)/w.$$

- The resulting entropy has the form

$$S = -\frac{\Delta}{\delta} \cdot \frac{1}{\Delta/\delta} \cdot \log_2\left(\frac{1}{\Delta/\delta}\right) - \frac{w}{2\delta} \cdot \frac{2p \cdot \delta}{w} \cdot \log_2\left(\frac{2p \cdot \delta}{w}\right).$$

- This expression can be simplified into

$$S = -\log_2(\delta) + \log_2(\Delta) - p\log_2(p) - p \cdot log_2(\delta) - p + p \cdot \log_2(w).$$

32. Preliminary analysis (cont-d)

- Here, $p \ll 1$, thus, $|\log_2(p)| \gg 1$, and hence, the term p can be safely ignored in comparison with $p \cdot \log_2(p)$.
- Thus, the number of missing bits is equal to

$$\log_2(\Delta) - p \cdot \log_2(p) + p \cdot \log_2(w) - p \cdot \log_2(\delta) + \ldots$$

- Here the three dots indicate terms that do not depend on the selection of the measuring instrument.
- Now, we are ready to start answering the questions.

33. How to select the most informative measuring instrument

- In line with the above computations, we need to select the measuring instrument:
  - with the smallest value of the above-mentioned quantity

$$v \stackrel{\text{def}}{=} \log_2(\Delta) - p \cdot \log_2(p) + p \cdot \log_2(w) - p \cdot \log_2(\delta),$$

  - i.e., equivalently, with the smallest value of e^v:

$$e^v = \Delta \cdot \left(\frac{p \cdot w}{\delta}\right)^p.$$

34. How to select the most informative combination of measurements within a given cost

- If we use n_i measuring instruments of type i, then, as stated previously:
  - the resulting outlier probability p is equal to

$$p = p_1^{n_1} \cdot \ldots \cdot p_k^{n_k},$$

  - the resulting uncertainty Δ is equal to

$$\Delta = \left(n_1 \cdot \Delta_1^{-2} + \ldots + n_k \cdot \Delta_k^{-2}\right)^{-1/2},$$

  - and the resulting cost c is equal to

$$c = n_1 \cdot c_1 + \ldots + n_k \cdot c_k.$$

- Thus, if we limit cost to some value c_0, the problem is:
  - among all the tuples $(n_1, \ldots, n_k)$ that satisfy the inequality $c \le c_0$,
  - we need to find the tuples with the smallest value of the quantity v.

35. How to find the least expensive way to get the desired amount of information

- In this case:
  - we minimize the cost
  - under the constraint that the amount of information is larger than or equal to the desired value v_0: $v \geq v_0$.

36. Second Scenario: Working By Specifications

- Suppose that the requirements come in the form of the thresholds Δ_0 on accuracy and p_0 on the outlier probability.
- I.e., we should have $\Delta \le \Delta_0$ and $p \le p_0$.
- Among all tuples $(n_1, \ldots, n_k)$ that satisfy both constraints, we need to find the least expensive one.

37. Analysis of the problem and its resulting formal description

- The inequality $\Delta \le \Delta_0$ is equivalent to $\Delta^{-2} \ge \Delta_0^{-2}$.
- Substituting the expression for Δ into this inequality, we get

$$n_1 \cdot \Delta_1^{-2} + \ldots + n_k \cdot \Delta_k^{-2} \ge \Delta_0^{-2}.$$

- Similarly, the inequality $p \le p_0$ is equivalent to $\ln(p) \le \ln(p_0)$.
- Substituting the expression for p into this inequality, we get

$$n_1 \cdot \ln(p_1) + \ldots + n_k \cdot \ln(p_k) \le \ln(p_0).$$

- In these terms, the problem is to find:
  - among all the tuples $(n_1, \ldots, n_k)$ that satisfy the given inequalities,
  - the tuple with the smallest value of the overall cost c.

38. How can we solve this optimization problem

- We need to optimize a linear expression under linear constraints.
- This is an integer-valued version of the linear programming problem.
- There are algorithms for solving such problems.
- One of the simplest of such algorithms is to solve the corresponding continuous optimization problem:
  - when we allow arbitrary non-negative values n_i, not just integer ones,
  - and then round up each value n_i to the nearest integer.
- With two constraints, the solution to a continuous linear programming problem will have only two non-zero values n_i.
- So in this case, we use only two types of measuring instruments.

39. Third Scenario: Optimization Problem Without Any Constraints

- In this scenario, we know the ideal (optimal) value of the parameter x_0 that we want to reach.
- For example, we want an airplane to follow the speed at which its fuel consumption per unit of distance is the smallest.
- To maintain this value x_0, we need to perform measurements.
- The problem is that:
  - even if we make sure that the measuring instrument returns the desired value x_0,
  - it does not mean that the actual value of the corresponding quantity x is equal to x_0.
- Due to measurement uncertainty, the actual value can take any value from the interval $[x_0 - \Delta, x_0 + \Delta]$.

40. Third Scenario: Optimization Problem Without Any Constraints (cont-d)

- Also, with some small probability p, the measurement result is an outlier that has nothing to do with reality.
- In this case, x can be anywhere within the general range $[\underline{x}, \overline{x}]$ of the quantity x.
- When x deviates from the optimal value x_0, we have a loss.
- The more accurately and the more reliably we measure, the smaller this loss.
- However, at the same time, the larger the measurement expenses.
- What is the measurement strategy that minimizes the overall costs – including both:
  - the additional costs of filtering and
  - the measurement expenses.

41. Let us formulate this problem in precise terms

- Let D be the expected cost of the situation when the measured value x_0 is an outlier – and thus, the actual value x can be anything.
- For an airplane, this may lead to a disaster, so we denoted this cost by D.
- The value x_0 minimizes expenses, i.e., minimizes the expression $E(x)$ that describes how expenses depend on x.
- In a small vicinity of x_0, we can:
  - expand the expression $E(x) = E(x_0 + \Delta x)$ in Taylor series, and
  - keep only the first few terms in this expansion.
- Since the function $E(x)$ attains its minimum at x_0, its linear term is equal to 0.
- Thus, the first non-constant term in the Taylor expansion is quadratic: $E(x_0 + \Delta) = E(x_0) + K \cdot (\Delta x)^2$, for some constant K.

42. Let us formulate this problem in precise terms (cont-d)

- So, the additional expenses caused by the measurement uncertainty are equal to $K \cdot (\Delta x)^2$.
- According to the decision theory, we need to select the decision in which the expected value of the utility is the largest – i.e., equivalently, in which the expected loss is the smallest.
- To find the expected loss, we need to know the probabilities of different uncertainty values from the interval $[-\Delta, \Delta]$.
- As we have mentioned, in practice, we often do not have any information about these probabilities.
- However, according to the utility-based decision-making paradigm, we need to select one of the possible probability distribution.
- We do not have any reason to believe that some probabilities are larger than others.

43. Let us formulate this problem in precise terms (cont-d)

- So, it makes sense to select the distribution for which all the probabilities are the same, i.e., the uniform distribution on this interval.
- One can show that for the uniform distribution on the interval $[-\Delta, \Delta]$, the average value of the expression $K \cdot (\Delta x)^2$ is equal to

$$(K/3) \cdot \Delta^2.$$

- Thus, the overall loss caused by the measurement imperfection is equal to $p \cdot D + (K/3) \cdot \Delta^2$.
- The overall cost can be computed as the sum of this loss and the measurement cost.
- Thus, we arrive at the following formulation of the problem:
- Find the tuple $(n_1, \ldots, n_k)$ that minimizes the expression

$$p \cdot D + (K/3) \cdot \Delta^2 + c.$$

- Here p, Δ, and c are determined by the known formulas.

44. Fourth Scenario: Optimization Under Constraints

- In this scenario, we assume that there is a threshold x_0 that we cannot overcome – otherwise, we get a huge penalty.
- E.g., for a chemical plant, the concentration x of some chemical in the surrounding air cannot exceed a given threshold x_0.
- Decreasing the concentration x to the desired level invokes costs.
- The smaller this level, the larger this cost.
- If we could measure x with absolute accuracy, then the best solution would be:
  - to apply the minimal necessary filtering,
  - i.e., to keep the value x exactly at the largest allowed value x_0.
- In practice, there is measurement uncertainty.
- If we measure with some accuracy Δ, this means that the actual value x may differ from the measurement result by Δ.

45. Fourth Scenario: Optimization Under Constraints (cont-d)

- We want to make sure that we never exceed the value x_0.
- So, we need to make sure that the measured value never exceeds

$$x_0 - \Delta.$$

- In other words, we need additional filtering.
- The smaller Δ, the less costly the filtering – but the more expensive the measurements.
- So, we want to minimize the overall expenses on filtering and on measurement.
- We also need to take into account the possibility that the measurement result is an outlier.

46. Let us formulate this problem in precise terms

- In this case, similar to the third scenario, we can also:
  - expand the expression $E(x) = E(x_0 - \Delta)$ (that describes how the expenses depend on x), and
  - keep only the first non-constant terms in this expansion.
- In this case, the function $E(x)$ does not attain its minimum for $x = x_0$.
- So we have non-constant linear terms: $E(x_0 - \Delta) = E(x_0) + K \cdot \Delta$ for some constant K.
- Thus, the overall loss caused by the measurement imperfection is equal to $p \cdot D + K \cdot \Delta$.
- The overall cost can be computed as the sum of this loss and the measurement cost.

47. Let us formulate this problem in precise terms (cont-d)

- Thus, we arrive at the following formulation of the problem.
- Find the tuple $(n_1, \ldots, n_k)$ that minimizes the expression

$$p \cdot D + K \cdot \Delta + c.$$

- Here p, Δ, and c are determined by the known formulas.

48. How This Affects Data Processing

- In data processing, we apply the algorithm $f(x_1, \ldots, x_n)$ to the results $\widetilde{x}_i$ of measuring the quantities $x_1, \ldots, x_n$.
- The measurement results are, in general, somewhat different from the corresponding actual values x_i; so:
  - the result $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ of data processing is, in general, different from
  - the ideal value $y = f(x_1, \ldots, x_n)$ that we would have gotten if we knew the exact values x_i.
- What can we say about the difference $\Delta y \stackrel{\text{def}}{=} \widetilde{y} - y$?
- We know the standard deviation σ_i of each measurement uncertainty

$$\Delta x_i = \widetilde{x}_i - x_i.$$

49. How This Affects Data Processing (cont-d)

- We know the probability p_i that the i-th measurement result is an outlier.
- Based on this information, we want to know the standard deviation σ of the value Δy, and the probability p that the value $\widetilde{y}$ is an outlier.

50. How to solve this problem

- To find σ, we can expand the expression Δy in Taylor series in terms of Δx_i and keep only linear terms in this expansion:

$$\Delta y = f(\widetilde{x}_1, \ldots, \widetilde{x}_n) - f(x_1, \ldots, x_n) = f(\widetilde{x}_1, \ldots, \widetilde{x}_n) - f(\widetilde{x}_1 - \Delta x_1, \ldots, \widetilde{x}_n - \Delta x_n).$$

- Then, we get $\Delta y = s_1 \cdot \Delta x_1 + \ldots + s_n \cdot \Delta x_n$, where we denoted

$$s_i \stackrel{\text{def}}{=} \frac{\partial f}{\partial x_i}.$$

- Thus, $\sigma^2 = s_1^2 \cdot \sigma_1^2 + \ldots + s_n^2 \cdot \sigma_n^2$.
- To estimate p, the main idea is that:
  - if one of the values $\widetilde{x}_i$ is very different from x_i,
  - then the result $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ of data processing is also very different from the desired value y.
- Thus, the only case when the value $\widetilde{y}$ is not an outlier is when none of the inputs are outliers.

51. How to solve this problem (cont-d)

- For each i, the probability that the i-th measurement result is not an outlier is equal to $1 - p_i$.
- Since the measurements are independent, the probability that all measurement results are not outliers is equal to the product

$$(1 - p_1) \cdot \ldots \cdot (1 - p_n).$$

- So, the probability p that $\widetilde{y}$ *is* an outlier is equal to 1 minus this probability, i.e., to $p = 1 - (1 - p_1) \cdot \ldots \cdot (1 - p_n)$.
- Usually, the values p_i are small.
- So we can expand this expression in Taylor series in terms of p_i and keep only the first terms in this expansion.
- This leads to a simplified formula $p = p_1 + \ldots + p_n$.

52. Conclusions

- Most of the data that we process comes from measurements, and measurements are never 100% accurate.
- There is always *measurement uncertainty*: the non-zero difference between the measurement result and the actual value of the quantity.
- This uncertainty affects the result of data processing.
- Measurement theory has developed many effective methods for quantifying and propagating measurement uncertainty.
- These methods allow us to gauge how the result of processing the measurement results differs from :
  - what we would have computed in the idealized case,
  - when we could apply the data processing algorithm to the actual values of the corresponding quantities.
- However, many of these methods do not take into account the issue of *reliability*.

53. Conclusions (cont-d)

- Sometimes, the measuring instruments malfunction and produce the results which are far off from the actual values of the quantities.
- In such situation, the results of data processing may also be far off from the desired values.
- In this talk, on several realistic scenarios, we show how both uncertainty and reliability can be taken into account in data processing.
- In this talk, we mostly concentrate on situations in which we know the probabilities of all situations.
- In practice, we often only have partial knowledge of these probabilities.
- This information may come from measurements and observations – or from expert estimates.

54. Conclusions (cont-d)

- It is therefore desirable to extend our ideas to such imprecise probability case, starting with the two simplest situations of this type:
- The case of nterval uncertainty, when:
  - we only know bounds on the corresponding values, and
  - we do not have any information about the probability of different values within these bounds.
- The case of fuzzy uncertainty, when we only have expert estimates described in natural-language terms.

55. Acknowledgments

This work was supported in part by:

- Deutsche Forschungsgemeinschaft Focus Program SPP 100+ 2388, Grant Nr. 501624329;
- 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).