2-1-2024

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
Leibniz University Hannover, winnewisser@irz.uni-hannover.de

Vladik Kreinovich
The University of Texas at El Paso, vladik@utep.edu

Olga Kosheleva
The University of Texas at El Paso, olgak@utep.edu

Comments:
Technical Report: UTEP-CS-24-05

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][0000−0003−0694−1349], Michael Beer[1], Vladik Kreinovich[2][0000−0002−1244−1650], and Olga Kosheleva[2][0000−0003−2587−4209]

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

Abstract. When we process data, it is important to take into account that data comes with uncertainty. There exist techniques for quantifying uncertainty and propagating this uncertainty through the data processing algorithms. However, most of these techniques do not take into account that in real world, measuring instruments are not 100% reliable – they sometimes malfunction and produce values which are far off from the measured values of the corresponding quantities. How can we take into account both uncertainty and reliability? In this paper, we consider several possible scenarios, and we show, for each scenario, what is the natural way to plan the measurements and to quantify and propagate the resulting uncertainty and reliability.

Keywords: Data processing · Measurement uncertainty· Measurement reliability.

1 Formulation of the Problem

Data processing is ubiquitous. The main objectives of science and engineering are to understand the current state of the world, to predict what will

* This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Focus Program SPP 100+ 2388, Grant Nr. 501624329, by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), HRD-1834620 and HRD-2034030 (CAHSI Includes), EAR-2225395 (Center for Collective Impact in Earthquake Science C-CIES), and by the AT&T Fellowship in Information Technology.
It was also supported by a grant from the Hungarian National Research, Development and Innovation Office (NRDI).

happen, and to make sure – by using appropriate devices and/or controls – that the future world is as beneficial for us as possible.

Understanding the current state of the world means, in particular, to understand the values of the physical quantities that characterize this state. Some of these quantities we can directly measure. For example, we can measure the current temperature, wind speed, distance between the two nearby buildings, etc. Other quantities y we cannot measure directly – 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-measure auxiliary quantities $x_1, \ldots, x_n$ that are related to y by a known relation $y = f(x_1, \ldots, x_n)$,
- we measure the values of these auxiliary quantities, and
- 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 values of 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 – these future values must be also measured indirectly:

- we find quantities $x_1, \ldots, x_n$ whose current values are related to the desired future value y by a known relation $y = f(x_1, \ldots, x_n)$,
- we measure the values of these current quantities, and
- we get an estimate for the future value y of the desired quantity by applying the algorithm $f(x_1, \ldots, x_n)$ to results of measuring the current values of the quantities $x_1, \ldots, x_n$.

The process of applying an algorithm to measurement results is known as *data processing*.

Comment. In many cases, the data processing algorithm consists of several distinct stages:

- first, we directly process the measurement results to estimate the values of some other auxiliary quantities;
- then, we apply some algorithm to process the measurement results and/or the results of the first stage of data processing;
- then, we apply some algorithm to process both the measurement results, the results of the first stage, and/or the results of the second stage of data processing, etc.

This is how, for example, deep neural networks process data; see, e.g., [5].

Uncertainty is ubiquitous. As we have mentioned, 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 other words, we have a measurement uncertainty $\Delta x \stackrel{\text{def}}{=} \widetilde{x} - x$.

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., compute the value $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$. Since the measurement results $\widetilde{x}_i$ are, in general, different from the actual values x_i, 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 – and even in this case, the dependence $y = f(x_1, \ldots, x_n)$ may be only approximate.

It is therefore desirable to understand how the measurement uncertainty propagates through the data processing algorithm, i.e., what is the resulting uncertainty $\Delta y \stackrel{\text{def}}{=} \widetilde{y} - y$.

How to take uncertainty into account: what is known. Several methods have been developed in measurement theory (see, e.g., [11]) to take uncertainty into account when we plan measurements and when we process data.

Remaining problem. The problem is that most of these methods do not take into account the fact that measurement instruments are not only not 100% accurate, they are also not 100% reliable. Sometimes, they malfunctions – and generate results which are far away 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.

What we do in this paper. In this paper, 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 when planning experiments and processing data.

2 Possible Scenarios

First scenario: journey to the unknown. As we have mentioned, the scenarios depend on what we know about the situation. 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, be it 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 accuracy (and what reliability) we will need – but we still design the corresponding instrument because in the past, such 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.

Second scenario: working by specifications. The second scenario is the opposite to the first one: we know exactly what accuracy (and what reliability) we need, we just need to find the least costly way to achieve these specifications.

Intermediate scenarios. In practice, we rarely know nothing about the appropriate values of accuracy and reliability, and we rarely have full information about them. Such situations of uncertainty are well-studied in decision theory (see, e.g., [3, 4, 7–10, 12]). 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 case 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 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 – a threshold that has already been determined by taking into account potential benefits and limitations of these types of plants.

Comment. These two scenarios look similar, but, as we will see, they lead to different optimization problems and thus, different solutions.

3 Analysis of the Problem

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 – i.e., what is the probability distribution of the measurement uncertainty; see, e.g., [11].

In practice, often, we only have the upper bound Δ on the absolute value $|\Delta x|$ of the measurement uncertainty: $|\Delta x| \le \Delta$. 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 away from it as mathematically possible – this is not what we would call a measuring instrument.

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 corresponding values of the measurement error. Once we know this mean, we can extract this value – known as *bias* – from all measurement results, and thus, conclude that the mean becomes 0.

We can also estimate the second moment – which, since the mean is 0, 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 factors. In this case, according to the Central Limit Theorem (see, e.g., [13]), the resulting distribution is close to Gaussian (normal). For the 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 $k = 3$, 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: $\Delta = k \cdot \sigma$.

The actual distribution may be different, but 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 paper.

How to estimate the amount of information. In the discrete case, when we have finitely many possible outcomes, 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. It is known (see, e.g., [1, 10]) that 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

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

In situations, when we do not know the exact values of the probabilities p_i, i.e., when several different probability distributions are possible, a natural idea is to take the largest amount of information corresponding to all possible distributions. It is known that if we have no information about the probabilities at all, the largest entropy corresponds to the uniform distribution $p_1 = \ldots = p_N$; see, e.g., [6].

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. This means we should have several values $x', x'', \ldots$, so that each 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.

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. All these measurements have the mean measurement uncertainty 0, and we know the corresponding standard deviations $\sigma_1, \ldots, \sigma_m$. 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 corresponding uncertainty

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

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

is the smallest possible.

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; see, e.g., [2, 14]. 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 $\sum c_i = 1$.

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 variance. So, for the variance σ of Δx, we have

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

We can use Lagrange multiplier method to minimize this value under the constraint $c_1 + \ldots + c_m = 1$. 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 $c_i = \sigma_i^{-2} / (\sum \sigma_j^{-2})$. Substituting these values into the formula for σ^2, we get $\sigma^2 = (\sum \sigma_i^{-2}) / (\sum \sigma_i^{-2})^2$, i.e., $\sigma^2 = 1 / (\sum \sigma_i^{-2})$ and

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

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

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

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, and 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. Since, as we have mentioned, it is reasonable to assume that the measurements are independent, the probability that all m measurements are outliers is equal to the product of the given probabilities, i.e., to $p = p_1 \cdot \ldots \cdot p_m$.

4 First Scenario: Journey to the Unknown

Possible questions. 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.
– In yet other cases, our goal is to extract a certain amount of information, and we want to find the arrangement with the minimal cost that will provide the required amount of information.

In this section, we will formulate all these problems in precise terms – so that one can 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.

Preliminary analysis. If 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, in line with the above analysis, let us pick some value δ. Then, with probability p, the actual value is somewhere in the original range $[\underline{x}, \overline{x}]$ of with $w \stackrel{\text{def}}{=} \overline{x} - \underline{x}$, 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 $(1-p)/(\Delta/\delta)$. 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

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

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

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

where 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.

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

$$v \stackrel{\text{def}}{=} \log_2(\Delta) - p \cdot \log_2(p) + p \cdot \log_2(w) - p \cdot \log_2(\delta), \tag{1}$$

i.e., equivalently, with the smallest value of the exponent of this quantity, i.e., the value

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

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 we have mentioned:

- the resulting outlier probability p is equal to

$$p = p_1^{n_1} \cdot \ldots \cdot p_k^{n_k}, \tag{2}$$

- the resulting uncertainty Δ is equal to

$$\Delta = \left(n_1 \cdot \Delta_1^{-2} + \ldots + n_k \cdot \Delta_k^{-2}\right)^{-1/2}, \tag{3}$$

- and the resulting cost c is equal to

$$c = n_1 \cdot c_1 + \ldots + n_k \cdot c_k. \tag{4}$$

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 (1), where c, p, and Δ are determined by the formulas (2)–(4).

How to find the least expensive way to get the desired amount of information. In this case, we minimize the cost (4) under the constraint that the amount of information (1) is larger than or equal to the desired value v_0: $v \ge v_0$.

5 Second Scenario: Working By Specifications

Formulation of the practical problem. 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.

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 (3) for Δ into this inequality, we get

$$n_1 \cdot \Delta_1^{-2} + \ldots + n_k \cdot \Delta_k^{-2} \ge \Delta_0^{-2}. \tag{5}$$

Similarly, the inequality $p \le p_0$ is equivalent to $\ln(p) \le \ln(p_0)$. Substituting the expression (2) for p into this inequality, we get

$$n_1 \cdot \ln(p_1) + \ldots + n_k \cdot \ln(p_k) \le \ln(p_0). \tag{6}$$

In these terms, the problem is to find, among all the tuples $(n_1, \ldots, n_k)$ that satisfy the inequalities (5) and (6), the tuple with the smallest value of the overall cost (4).

How can we solve this optimization problem. The above problem – of optimizing a linear expression under linear constraints – is an integer-valued version of the linear programming problem (see, e.g., [15]). 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.

6 Third Scenario: Optimization Problem Without Any Constraints

Formulation of the practical problem. In this scenario, we know the ideal (optimal) value of the parameter x_0 that we want to reach – e.g., 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 $[x_0 - \Delta, x_0 + \Delta]$. 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 – but at the same time, the larger the measurement expenses. What is the measurement strategy that minimizes the overall loss?

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 $E(x)$ describing how expenses depend on x. In a small vicinity of x_0, we can expand the dependence $E(x) = E(x_0 + \Delta x)$ 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 and 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. So, the additional expenses caused by the measurement uncertainty are equal to $K \cdot (\Delta x)^2$.

As we have mentioned, 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. Since we do not have any reason to believe that some probabilities are larger than others, it makes sense to assume that all the probabilities are the same, i.e., that we have a uniform distribution on this interval; see, e.g., [6]. 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 (4).

Thus, we arrive at the following formulation of the problem: find the tuple $(n_1, \ldots, n_k)$ that minimizes the expression $p \cdot D + (K/3) \cdot \Delta^2 + c$, where p, Δ, and c are determined by the formulas (2)–(4).

7 Fourth Scenario: Optimization Under Constraints

Formulation of the practical problem. In this scenario, we assume that there is a threshold x_0 that we cannot overcome – otherwise, we get a huge penalty. An example that we mentioned above is a chemical plant, for which 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, and 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. So, to make sure that we never exceed the value x_0, 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.

Let us formulate this problem in precise terms. In this case, similar to the third scenario, we can also expand the dependence $E(x) = E(x_0 - \Delta)$ of

expenses 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 (4).

Thus, we arrive at the following formulation of the problem: find the tuple $(n_1, \ldots, n_k)$ that minimizes the expression $p \cdot D + K \cdot \Delta + c$, where p, Δ, and c are determined by the formulas (2)–(4).

8 How This Affects Data Processing

Formulation of the practical problem. In data processing, we apply the algorithm $f(x_1, \ldots, x_n)$ to the results $\widetilde{x}_i$ of measuring the quantities $x_1, \ldots, x_n$.

Since the measurement results are, in general, somewhat different from the corresponding actual values x_i, 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 $\Delta x_i = \widetilde{x}_i - x_i$, and 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.

How to solve this problem. To find σ, we can – as above – expand the expression

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

in Taylor series in terms of Δx_i and keep only linear terms in this expansion. Then, we get $\Delta y = s_1 \cdot \Delta x_1 + \ldots + s_n \cdot \Delta x_n$, where we denoted

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

and thus [11],

$$\sigma^2 = s_1^2 \cdot \sigma_1^2 + \ldots + s_n^2 \cdot \sigma_n^2.$$

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. 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 $(1 - p_1) \cdot \ldots \cdot (1 - p_n)$ of these probabilities. 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

$$p = p_1 + \ldots + p_n.$$

References

1. B. Chokr and V. Kreinovich. "How far are we from the complete knowledge: complexity of knowledge acquisition in Dempster-Shafer approach." In R. R. Yager, J. Kacprzyk, and M. Pedrizzi (Eds.), *Advances in the Dempster-Shafer Theory of Evidence*, Wiley, N.Y., 1994, pp. 555–576.
2. R. Feynman, R. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Addison Wesley, Boston, Massachusetts, 2005.
3. P. C. Fishburn, *Utility Theory for Decision Making*, John Wiley & Sons Inc., New York, 1969.
4. P. C. Fishburn, *Nonlinear Preference and Utility Theory*, The John Hopkins Press, Baltimore, Maryland, 1988.
5. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, Cambridge, Massachusetts, 2016.
6. E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.
7. V. Kreinovich, "Decision making under interval uncertainty (and beyond)", In: P. Guo and W. Pedrycz (eds.), *Human-Centric Decision-Making Models for Social Sciences*, Springer Verlag, 2014, pp. 163–193.
8. R. D. Luce and R. Raiffa, *Games and Decisions: Introduction and Critical Survey*, Dover, New York, 1989.
9. H. T. Nguyen, O. Kosheleva, and V. Kreinovich, "Decision making beyond Arrow's 'impossibility theorem', with the analysis of effects of collusion and mutual attraction", *International Journal of Intelligent Systems*, 2009, Vol. 24, No. 1, pp. 27–47.
10. H. T. Nguyen, V. Kreinovich, B. Wu, and G. Xiang, *Computing Statistics under Interval and Fuzzy Uncertainty*, Springer Verlag, Berlin, Heidelberg, 2012.
11. S. G. Rabinovich, *Measurement Errors and Uncertainty: Theory and Practice*, Springer Verlag, New York, 2005.
12. H. Raiffa, *Decision Analysis*, McGraw-Hill, Columbus, Ohio, 1997.
13. D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.
14. K. S. Thorne and R. D. Blandford, *Modern Classical Physics: Optics, Fluids, Plasmas, Elasticity, Relativity, and Statistical Physics*, Princeton University Press, Princeton, New Jersey, 2021.
15. R. J. Vanderbei, *Linear Programming: Foundations and Extensions*, Springer, New York, 2014.