10-2020

What If We Use Almost-Linear Functions Instead of Linear Ones as a First Approximation in Interval Computations

Martine Ceberio
University of Texas at El Paso, mceberio@utep.edu

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

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

Comments:
Technical Report: UTEP-CS-20-99a
Published in: Franco Pavese, Alistair B. Forbes, Nien Fan Zhang, and Anna G. Chunovkina (eds.), *Advanced Mathematical and Computational Tools in Metrology and Testing XII*, World Scientific, Singapore, 2021, pp. 149-166.

What If We Use Almost-Linear Functions Instead of Linear Ones as a First Approximation in Interval Computations

M. Ceberio

Department of Computer Science
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
E-mail: mceberio@utep.edu

O. Kosheleva

Department of Teacher Education
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
E-mail: olgak@utep.edu

V. Kreinovich

Department of Computer Science
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
E-mail: vladik@utep.edu

In many practical situations, the only information that we have about measurement errors is the upper bound on their absolute values. In such situations, the only information that we have after the measurement about the actual (unknown) value of the corresponding quantity is that this value belongs to the corresponding interval: e.g., if the measurement result is 1.0, and the upper bound is 0.1, then this interval is $[1.0-0.1, 1.0+0.1] = [0.9, 1.1]$. An important practical question is what is the resulting interval uncertainty of indirect measurements, i.e., in other words, how interval uncertainty propagates through data processing. There exist feasible algorithms for solving this problem when data processing is linear, but for quadratic data processing techniques, the problem is, in general, NP-hard. This means that (unless P=NP) we cannot have a feasible algorithm that always computes the exact range, we can only find good approximations for the desired interval. In this paper, we propose two new metrologically motivated approaches (and algorithms) for computing such approximations.

Keywords: interval computations, measurement uncertainty, NP-hard prob-

lems, monotonicity, indirect measurements, uncertainty quantification

1. Why Interval Computations

Measurement uncertainty is ubiquitous. Measurements are never absolutely accurate: the measurement result $\widetilde{x}$ is, in general, different from the actual (unknown) unknown value x of the corresponding quantity; see, e.g., [8].

Case of interval uncertainty. Traditional metrological techniques assumes that we know the probability distribution of the measurement error

$$\Delta x \stackrel{\text{def}}{=} \widetilde{x} - x.$$

However, in many real-life situations, the only information that we have about the measurement error is the upper bound Δ on its absolute value: $|\Delta x| \le \Delta$. In these situations, we do not have any information about the probability of different values, we do not even know which values are more probable and which are less probable. In principle, any probability distribution on the interval $[-\Delta, \Delta]$ is possible.

In this case, after we get the measurement result $\widetilde{x}$, the only information that we have about the actual value x is that x belongs to the interval $[\widetilde{x} - \Delta, \widetilde{x} + \Delta]$; see, e.g., [8].

Why interval uncertainty. The usual way of determining the probability distribution of the measurement errors is to *calibrate* the measuring instrument, i.e., to compare, several times, the results of measuring the same value by this instrument and by a much more accurate ("standard") one. Since the measurement error of the standard instrument is much smaller that the measurement error of our instrument, we can safely ignore this standard measurement error and assume that the values $\widetilde{x}_{st}$ measured by the standard instrument represent the actual values of the corresponding quantity. Under this assumption, the difference $\widetilde{x} - \widetilde{x}_{st}$ is approximately equal to the measurement error $\Delta x = \widetilde{x} - x$. After we perform this comparison several times, we get a sample of values of measurement error – and from this sample, we can determine the desired probability distribution.

This procedure is reasonable, and it is often implemented, but there are two important classes of situations in which this calibration is not performed. First, this procedure is not done for cutting-edge measurements, when the measuring instrument that we use is among the most accurate, and thus, there is no much more accurate instrument that could be used as a standard. Second, this procedure is often not done in practice simply

because calibration is an expensive procedure: e.g., when high school kids build robots, they can buy very cheap sensors, but calibrating each sensor would require the use of expensive high-accuracy measuring instrument and would thus cost much more than sensors themselves. As a result, in manufacturing, often, instead of calibrating a sensor, practitioners simply use the upper bound on the measurement error – bound provided by the manufacturer of the measuring instrument.

Need for indirect measurements. Some quantities we can measure directly. Other quantities y are difficult to measure directly. To estimate these quantities:

- we find (and measure) easier-to-measure quantities $x_1, \ldots, x_n$ which are related to the desired quantity y by a known dependence

$$y = f(x_1, \ldots, x_n),$$

- and then we plug in the measurement results $\widetilde{x}_i$ into this formula, producing an estimate $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.

This estimation process is known as *indirect measurement* or, alternatively, *data processing*.

Need for take measurement uncertainty into account in indirect measurements. The measurement results $\widetilde{x}_i$ are, in general, somewhat different from the actual (unknown) values x_i of the corresponding quantities. As a result, the estimate $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ is, in general, different from the actual value $y = f(x_1, \ldots, x_n)$ of the desired quantity. A natural metrological equation is: how big is the difference $\Delta y \stackrel{\text{def}}{=} \widetilde{y} - y$? What can we say about the measurement error Δy of the indirect measurement?

Why not use uniform distributions? At first glance, the situation can be covered by the traditional probabilistic methods. Indeed, since we do not know which values of a measurement error Δx_i are more probable and which are less probable, a reasonable idea is to assume that all these values are equally probable, i.e., that we have a uniform distribution on the corresponding interval $[-\Delta_i, \Delta_i]$. However, it is easy to show that this seemingly natural idea may lead to a drastic overestimation of the accuracy of the indirect measurement.

Indeed, let us consider a simple situation when $f(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i$, all the measurement results are zeros, i.e., $\widetilde{x}_1 = \ldots = \widetilde{x}_n = 0$, and all the

upper bounds on the measurement errors are equal to 1:

$$\Delta_1 = \ldots = \Delta_n = 1.$$

In this situation, the result of data processing is 0: $\widetilde{y} = \sum_{i=1}^{n} \widetilde{x}_i = 0$, so $\Delta y = y$. The value $\Delta y = y$ attains its largest possible value when all the terms x_i attain their largest possible value $x_i = 1$. In this case, $y = n$, so the largest possible value of Δy is equal to n.

But what if we assume that all measurement errors Δx_i are uniformly distributed on the interval $[-\Delta_i, \Delta_i] = [-1, 1]$? In this case, for large n, the value y is the sum of a large number of independent identically distributed random variables. Due to the Central Limit Theorem, the distribution of the sum y is thus close to normal. The mean of the sum is equal to the sum of the means, i.e., to 0, and the variance σ^2 of the sum is equal to the sum of n variances, i.e., to $n/3$. Thus, the distribution of y is close to a normal distribution with mean 0 and standard deviation $\sqrt{n}/\sqrt{3}$. In practice, deviations larger than 6σ are so improbable that they are ignored. So, we conclude that all the values of y are bounded by $6\sigma = (6/\sqrt{3}) \cdot \sqrt{n}$.

For large n, this value const$\cdot\sqrt{n}$ is much smaller than the actual possible value n of the measurement error – a drastic underestimation of the error of indirect measurement (and thus, a drastic overestimation of accuracy). In many critical situations, an underestimation of the measurement error can lead to a disaster: e.g., when, based on the measurement results, we think that we are still within the safe zone, but in reality, we have already crossed the threshold to a danger zone.

Summarizing: from the metrological viewpoint, we cannot simply replace the interval uncertainty with a uniform distribution, we have to consider interval uncertainty – i.e., in effect, consider all possible probability distributions on each interval.

Need for interval computations. As we have mentioned earlier, often, we have the case of interval uncertainty, when the only information that we have about each value x_i is that this value belongs to the corresponding interval $[\underline{x}_i, \overline{x}_i]$. In this case, the only information that we have about the value $y = f(x_1, \ldots, x_n)$ is that this value belongs to the range $[y^-, y^+]$ of all possible values $f(x_1, \ldots, x_n)$ when each x_i is in the corresponding interval:

$$[y^-, y^+] = f([\underline{x}_1, \overline{x}_1], \ldots, [\underline{x}_n, \overline{x}_n]) \stackrel{\text{def}}{=}$$

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

Computation of this range is known as *interval computations*; see, e.g.,[2,4,5].

2. Interval Computations – Successes and Challenges: A Very Brief Overview

Measurement errors are usually small. Each interval of possible values of x_i has the form $[\underline{x}_i, \overline{x}_i] = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$. Each value x_i from this interval has the form $x_i = \widetilde{x}_i - \Delta x_i$, where $|\Delta x_i| \le \Delta_i$. Thus, the actual value $y = f(x_1, \ldots, x_n)$ has the form $y = f(\widetilde{x}_1 - \Delta x_1, \ldots, \widetilde{x}_n - \Delta x_n)$, and the measurement error Δy of the indirect measurement has the form

$$\Delta y = \widetilde{y} - y = f(\widetilde{x}_1, \ldots, \widetilde{x}_n) - f(\widetilde{x}_1 - \Delta x_1, \ldots, \widetilde{x}_n - \Delta x_n). \qquad (1)$$

The dependence $f(x_1, \ldots, x_n)$ is usually analytical. Thus, the expression (1) can be expanded into power series in terms of the unknown values Δx_i, i.e., represented as the sum of terms which are linear in Δx_i, terms quadratic in terms of Δx_i, terms cubic in terms of Δx_i, etc. The measurement errors Δx_i are usually relatively small: usually, no more than 20% (and in most cases, much smaller than that). In this case, terms quadratic in Δx_i are of order $(20\%)^2 = 4\%$, terms cubic in Δx_i are of order $(20\%)^3 = 0.8\%$, etc. The higher the order, the smaller corresponding terms. Thus, from the practical viewpoint, we can safely ignore higher order terms in this expansion and only keep terms up to a certain power. For example, we can keep:

- only linear terms or
- only linear and quadratics terms.

Comment. If we want a more accurate estimate, then, instead of ignoring higher order terms, we can add one of the known bounds for the remaining terms.

What if we only keep linear terms. If we only keep linear terms, then we get a formula

$$\Delta y = \sum_{i=1}^{n} c_i \cdot \Delta x_i, \qquad (2)$$

where $c_i \stackrel{\text{def}}{=} \dfrac{\partial f}{\partial x_i}$.

In this case, each term $c_i \cdot \Delta x_i$ in the sum is independent – in the sense that each terms depend only on its own variable Δx_i. So the sum attains its largest possible value Δ when each of the terms is the largest possible. Here:

- If $c_i > 0$, then the expression $c_i \cdot \Delta x_i$ is increasing in Δx_i, so its largest possible value is attained when the value Δx_i is the largest, i.e., when $\Delta x_i = \Delta_i$. In this case, the expression has the value $c_i \cdot \Delta_i$.
- If $c_i < 0$, then the expression $c_i \cdot \Delta x_i$ is decreasing in Δx_i, so its largest possible value is attained when the value Δx_i is the smallest, i.e., when $\Delta x_i = -\Delta_i$. In this case, the expression has the value $-c_i \cdot \Delta_i$.

In both case, the largest possible value of each term $c_i \cdot \Delta x_i$ is $|c_i| \cdot \Delta_i$. Thus, the largest possible value Δ of the sum Δy of these n terms is equal to the sum of these largest value:

$$\Delta \stackrel{\text{def}}{=} \sum_{i=1}^{n} |c_i| \cdot \Delta_i. \tag{3}$$

Similarly, the sum attains its smallest possible value Δ when each of the terms is the smallest possible. Here:

- If $c_i > 0$, then the expression $c_i \cdot \Delta x_i$ is increasing in Δx_i, so its smallest possible value is attained when the value Δx_i is the smallest, i.e., when $\Delta x_i = -\Delta_i$. In this case, the expression has the value $-c_i \cdot \Delta_i$.
- If $c_i < 0$, then the expression $c_i \cdot \Delta x_i$ is decreasing in Δx_i, so its smallest possible value is attained when the value Δx_i is the largest, i.e., when $\Delta x_i = \Delta_i$. In this case, the expression has the value $c_i \cdot \Delta_i$.

In both case, the smallest possible value of each term $c_i \cdot \Delta x_i$ is $-|c_i| \cdot \Delta_i$. Thus, the smallest possible value of Δy is equal to

$$-\sum_{i=1}^{n} |c_i| \cdot \Delta_i,$$

i.e., to $-\Delta$, where Δ is the expression (3). Thus, the range of Δy is the interval $[-\Delta, \Delta]$.

The expression (3) is easy to compute – it requires $O(n)$ steps, i.e., linear time.

What if we also keep quadratic terms: general case. In this case, we have a quadratic expression

$$\Delta y = \sum_{i=1}^{n} c_i \cdot \Delta x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} \cdot \Delta x_i \cdot \Delta x_j, \tag{4}$$

where $c_{ij} = c_{ji} \stackrel{\text{def}}{=} \dfrac{1}{2} \cdot \dfrac{\partial^2 f}{\partial x_i \partial x_j}$. We want to find the minimum $\underline{y}$ and the maximum $\overline{y}$ of the expression (4) when $\Delta x_i \in [-\Delta_i, \Delta_i]$.

It turns out that this problem is already NP-hard; see, e.g., [6]. This means that (unless P = NP) no feasible algorithm can always compute the bounds y^- and y^+; see, e.g., [3,7].

Since we cannot compute the exact bounds, we need to provide estimates for these bounds – to be more precise, upper bounds, since one of the main purposes of metrology is to provide guaranteed upper bounds on the measurement errors.

Using monotonicity. One of the main ideas in interval computations is that if a function is monotonic with respect to one of the variables Δx_i, then to compute its range, it is sufficient to consider only the endpoints of the range $[-\Delta_i, \Delta_i]$.

Specifically, if the expression (4) is increasing in Δx_i, then:

- to find y^+, it is sufficient to consider the value $\Delta x_i = \Delta_i$, and
- to find y^-, it is sufficient to consider the value $\Delta x_i = -\Delta_i$.

Similarly, if the expression (4) is decreasing in Δx_i, then:

- to find y^+, it is sufficient to consider the value $\Delta x_i = -\Delta_i$, and
- to find y^-, it is sufficient to consider the value $\Delta x_i = \Delta_i$.

How can we check whether the expression (4) is increasing or decreasing with respect to x_i? According to calculus:

- an expression is increasing on some domain if and only if the partial derivative with respect to x_i is non-negative for all the points from this domain, and
- an expression is increasing on some domain if and only if the partial derivative with respect to x_i is non-negative for all the points from this domain.

For a quadratic expression, the partial derivative is linear, it is equal to

$$c_i + 2 \sum_{j=1}^{n} c_{ij} \cdot \Delta x_j.$$

We already know how to compute the range of a linear function. So, the

range of the partial derivative is equal to:

$$\left[c_i - 2\sum_{j=1}^{n} |c_{ij}| \cdot \Delta_j, c_i + 2\sum_{j=1}^{n} |c_{ij}| \cdot \Delta_j\right].$$

Here:

- If the lower endpoint of this interval is non-negative, this means that the derivative is always non-negative, so the expression is increasing in Δx_i.
- If the upper endpoint of this interval is non-positive, this means that the derivative is always non-positive, so the expression is decreasing in Δx_i.

Thus, for each i, we compute the value $s_i \stackrel{\text{def}}{=} 2\sum\limits_{j=1}^{n} |c_{ij}| \cdot \Delta_j$. If $c_i - s_i \geq 0$, then:

- to compute y^+, we replace Δx_i in the expression (4) with Δ_i; and
- to compute y^-, we replace Δx_i in the expression (4) with $-\Delta_i$.

Similarly, if $c_i + s_i \leq 0$, then:

- to compute y^+, we replace Δx_i in the expression (4) with $-\Delta_i$; and
- to compute y^-, we replace Δx_i in the expression (4) with Δ_i.

After we do this for all variables, we get – for each of the two problems of computing y^- and of computing y^+ – an expression of the similar type (4), but with fewer variables – namely, only with variables x_i with respect to which the original expression was neither everywhere increasing not everywhere decreasing.

So, we end with the same problem of computing the range of the expression (4), but with fewer variables than originally. How can we compute it?

Straightforward approach and a natural question. If we find upper bounds for each term in the expression (4), then by adding them, we clearly get an upper bound for the expression (4). When $|\Delta x_i| \leq \Delta_i$ and $|\Delta x_j| \leq \Delta_j$, then we have $|\Delta x_i \cdot \Delta x_j| \leq \Delta_i \cdot \Delta_j$. Similarly, we conclude that the range of possible values of $(\Delta x_i)^2$ is the interval $\left[0, (\Delta_i)^2\right]$. Thus, for expression

(4), we get the following upper bound $\overline{Y}$ and lower bound $\underline{Y}$:

$$\overline{Y} = \sum_{i=1}^{n} |c_i| \cdot \Delta_i + \sum_{i:c_{ii}>0} c_{ii} \cdot (\Delta_i)^2 + \sum_{i \neq j} |c_{ij}| \cdot \Delta_i \cdot \Delta_j;$$

$$\underline{Y} = -\sum_{i=1}^{n} |c_i| \cdot \Delta_i - \sum_{i:c_{ii}<0} |c_{ii}| \cdot (\Delta_i)^2 - \sum_{i \neq j} |c_{ij}| \cdot \Delta_i \cdot \Delta_j.$$

These formulas are clearly feasible while the problem of computing the exact range is, as have mentioned, NP-hard. Thus, these formulas do not always produce exact ranges.

So, a natural question is: can we – and if yes, how – find more accurate bounds, i.e., bounds which are closer to the actual difficult-to-compute ranges?

Two natural simplifications. Computations can be somewhat less cumbersome if:

- instead of the original variables Δx_i – which take any values from $-\Delta_i$ to Δ_i,
- we consider auxiliary variables $z_i \stackrel{\text{def}}{=} \text{sign}(c_i) \cdot \dfrac{\Delta x_i}{\Delta_i}$ for which $z_i \in [-1, 1]$ and $\Delta x_i = \Delta_i \cdot z_i$.

Substituting the expression for Δx_i in terms of z_i into the formula (4), we get an expression

$$\sum_{i=1}^{n} a_i \cdot z_i + \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \cdot z_i \cdot z_j, \tag{5}$$

where $a_i \stackrel{\text{ref}}{=} c_i \cdot \text{sign}(c_i) \cdot \Delta_i = |c_i| \cdot \Delta_i \geq 0$ and

$$a_{ij} \stackrel{\text{def}}{=} c_{ij} \cdot \text{digns}(c_i) \cdot \text{sign}(c_j) \cdot \Delta_i \cdot \Delta_j.$$

The task it then to find the largest and the smallest value of the expression (5) when $z_i \in [-1, 1]$.

This was the first simplification. The second simplification is that to compute the minimum of an expression, it is sufficient to consider the maximum of minus this expression. Since minus quadratic expression is still a quadratic expression, it is therefore sufficient to learn how to compute the maximum.

What we do in this paper. In this paper, we describe two approaches to compute such more accurate bounds.

3. First Approach: Taking Major Inputs into Account

Idea. In many cases, one input z_i is the most influential. In such cases, it is reasonable to assume that the effect of this input provides the largest contribution to the quadratic part of the expression (5). The corresponding quadratic terms is $a_{ii} \cdot z_i^2$.

We may have another input which is almost as influential, for which the corresponding major term is $a_{jj} \cdot z_j^2$.

Since we cannot estimate the exact range of the expression (5) that contains *all* the quadratic terms, a natural idea is:

- to estimate the expression that have all linear terms and the above-mentioned major quadratic terms, and then
- to use the straightforward estimation method to take all other terms into account – hoping that these other terms are smaller and thus, their overestimation will be smaller.

Resulting problem. We want to find the largest possible value $\overline{e}$ of the expression

$$e \stackrel{\text{def}}{=} \sum_{i=1} a_i \cdot z_i + \sum_{i=1}^{n} a_{ii} \cdot z_i^2 \tag{6}$$

when $z_i \in [-1, 1]$.

How to solve this problem: idea and resulting algorithm. The expression (6) is the sum of n independent expressions, each of which depends only on one of the variables z_i. Thus, the desired maximum $\overline{e}$ is simply equal to sum of the maxima of the corresponding n expressions.

For a quadratic expression $a_i \cdot z_i + a_{ii} \cdot z_i^2$ of one variable, its maximum on the interval $[-1, 1]$ is attained:

- either at one of the endpoints, i.e., for $z_i = -1$ or for $z_i = 1$,
- or at the point where the derivative of this expression is equal to 0, i.e., at the point $z_i = -\dfrac{a_i}{2a_{ii}}$ – provided that this point is located inside the interval $[-1, 1]$.

For each i, the values at these three (or two) points can be easily computed, their largest of these three (or two) points can also be easily computed. The sum of these n maxima is the desired bound $\overline{e}$.

For each of n values of the index i, we need a fixed number of computational steps, so overall, this algorithm requires linear time to compute.

4. Second Approach: Taking Major Combinations of Inputs into Account

Idea. In the previous section, we considered the case when one of the inputs makes a major contribution to the measurement error. In practice, however, we may have a situation in which the major contributor is not *one* of the inputs, but rather a *linear combination* of such inputs.

Let us clarify what we mean. In the linear approximation, the measurement error is equal to $\sum\limits_{i=1}^{n} a_i \cdot z_i = \sum\limits_{i=1}^{n} c_i \cdot \Delta x_i$. This value is equal to the difference Δx between the estimated and actual values of the linear combination $x \stackrel{\text{def}}{=} \sum\limits_{i=1}^{n} c_i \cdot x_i$. This linear combination is responsible for all the linear terms. Since the quadratic terms are usually much smaller than the linear terms, this means that the linear combination x is the major contributor to the measurement error.

If we only take this linear combination into account, but allow quadratic dependence, then we get an expression of the type

$$\sum_{i=1}^{n} a_i \cdot z_i + C \cdot \left(\sum_{i=1}^{n} a_i \cdot z_i \right)^2,$$

for some constant C.

Of course, there may be other ("secondary") linear combinations

$$\sum_{i=1}^{n} b_{kj} \cdot x_j \quad (k = 1, 2, \ldots)$$

whose contribution is smaller than the contribution of x but which we would like to also take into account. We are considering quadratic expressions, so we may have quadratic terms of two types:

- terms proportional to the product of the major linear combination and one of the secondary linear combinations, and
- terms proportional to the square of a secondary combination or to a product of two secondary combinations.

Since the major combination has the largest effect on the measurement error, it is reasonable to assume that the joint effect of two combinations is larger when one of these combinations is the major one. Since we cannot take *all* the terms into account – this will make the problem NP-hard – it makes sense to only take such larger terms into account, i.e., to consider

the expression of the type:

$$\sum_{i=1}^{n} a_i \cdot z_i + C \cdot \left(\sum_{i=1}^{n} a_i \cdot z_i\right)^2 + \sum_k C_k \cdot \left(\sum_{i=1}^{n} a_i \cdot z_i\right) \cdot \left(\sum_{j=1}^{n} b_{kj} \cdot z_j\right),$$

for some coefficients C_k. All non-linear terms in this formula have the same factor $\sum\limits_{i=1}^{n} a_i \cdot z_i$. By combining these terms, we get the following expression:

$$e \stackrel{\text{def}}{=} \sum_{i=1}^{n} a_i \cdot z_i + \left(\sum_{i=1}^{n} a_i \cdot z_i\right) \cdot \left(\sum_{j=1}^{n} b_j \cdot z_j\right), \tag{7}$$

where $b_j \stackrel{\text{def}}{=} C \cdot a_j + \sum\limits_k C_k \cdot b_{kj}$.

Resulting problem. We want to find the largest possible value $\overline{e}$ of the expression (7) when $z_i \in [-1, 1]$.

Natural simplification. If for some i and j, we have $\dfrac{b_i}{a_i} = \dfrac{b_j}{a_j}$, then

$$b_i \cdot z_i + b_j \cdot z_j = c \cdot (a_i \cdot z_i + a_j \cdot z_j),$$

where

$$c \stackrel{\text{def}}{=} \frac{b_i}{a_i} = \frac{b_j}{a_j}.$$

Thus:

- instead of two independent variables z_i and z_j,
- we have, in effect, a single variable $a_i \cdot z_i + a_j \cdot z_j$.

When $z_i \in [-1, 1]$ and $z_j \in [-1, 1]$, this variables takes all possible values from the interval $[-(a_i + a_j), a_i + a_j]$. We can therefore simplify the problem if follow the same idea that we used to introduce the variables a_i; namely:

- in the "a-term", we replace the expression $a_i \cdot z_i + a_j \cdot z_j$
- with the expression $(a_i + a_j) \cdot z_{ij}$ for a new variable $z_{ij} \in [-1, 1]$.

Correspondingly, in the "b-term":

- we replace the sum $b_i \cdot z_i + b_j \cdot z_j$
- with the expression $[c \cdot (a_i + a_j)] \cdot z_{ij}$.

By applying this simplification, we will end up with a problem with fewer variables, in which all the ratios $\frac{b_i}{a_i}$ are different.

How to solve our problem: idea. According to calculus, the maximum of a function $f(x_i)$ on an interval – in particular, on the interval $[-1, 1]$ is attained in one of the three cases:

- The maximum can be attained at the left endpoint – in our case, at the point $z_i = -1$. In this case, at this point, we must have $\frac{\partial f}{\partial z_i} \le 0$. Indeed, if this partial derivative was positive, the value of the function would increase when we slightly increase z_i from -1, so we would not have the maximum at the point -1.
- The maximum can be attained at the right endpoint – in our case, at the point $z_i = 1$. In this case, at this point, we must have $\frac{\partial f}{\partial z_i} \ge 0$. Indeed, if this partial derivative was negative, the value of the function would increase when we slightly decrease z_i from 1, so we would not have the maximum at the point 1.
- The maximum can also be attained inside the interval. In this case, the partial derivative should equal to 0: $\frac{\partial f}{\partial z_i} \le 0$.

So:

- If at the point where the maximum is attained, we have $\frac{\partial f}{\partial z_i} > 0$, this would mean that at this point, $z_i = 1$ – otherwise, the partial derivative would be non-positive.
- Similarly, if at the point where the maximum is attained, we have $\frac{\partial f}{\partial z_i} < 0$, this would mean that at this point, $z_i = -1$ – otherwise, the partial derivative would be non-negative.

For the expression (7), the partial derivative is equal to

$$\frac{\partial f}{\partial z_i} = a_i \cdot B + b_i \cdot A,$$

where

$$A \stackrel{\text{def}}{=} \sum_{j=1} a_j \cdot z_j \text{ and } B \stackrel{\text{def}}{=} 1 + \sum_{j=1}^{n} b_j \cdot z_j.$$

To decide for which values z_i the maximum is attained, we need to analyze the sign of this partial derivative.

Here, $a_i \geq 0$, so the inequality $a_i \cdot B + b_i \cdot A > 0$ is equivalent to $B + \dfrac{b_i}{a_i} \cdot A > 0$. Let us consider two possible cases: $A > 0$ and $A < 0$.

Case when $A > 0$. If $A > 0$, then the above inequality $B + \dfrac{b_i}{a_i} \cdot A > 0$ is, in its turn, equivalent to

$$\frac{b_i}{a_i} > -\frac{B}{A}.$$

We know that all the values $\dfrac{b_i}{a_i}$ are different. Let us order the indices i so that these ratios are increasing:

$$\frac{b_1}{a_1} < \frac{b_2}{a_2} < \ldots < \frac{b_n}{a_n}. \tag{8}$$

There exists a threshold $-\dfrac{B}{A}$ so that:

- for all indices for which the ratio $\dfrac{b_i}{a_i}$ is larger than this threshold, the partial derivative $\dfrac{\partial f}{\partial z_i}$ is positive and thus, $z_i = 1$;
- similarly, for all indices i for which the ratio $\dfrac{b_i}{a_i}$ is smaller than this threshold, the partial derivative $\dfrac{\partial f}{\partial z_i}$ is negative and thus, $z_i = -1$.

There can be no more than one value i for which the ratio $\dfrac{b_i}{a_i}$ is exactly equal to the threshold. For this index i, we cannot say anything about the value z_i. So, when $A > 0$, in the order of indices, the optimal sequence of values z_i should be

$$(-1, -1, \ldots, -1, z_i, 1, \ldots, 1) \tag{9}$$

Case when $A < 0$. If $A > 0$, then the inequality $B + \dfrac{b_i}{a_i} \cdot A > 0$ is, in its turn, equivalent to

$$\frac{b_i}{a_i} < -\frac{B}{A}.$$

So, in this case:

- for all indices for which the ratio $\dfrac{b_i}{a_i}$ is smaller than this threshold, the partial derivative $\dfrac{\partial f}{\partial z_i}$ is positive and thus, $z_i = 1$;
- similarly, for all indices i for which the ratio $\dfrac{b_i}{a_i}$ is larger than this threshold, the partial derivative $\dfrac{\partial f}{\partial z_i}$ is negative and thus, $z_i = -1$.

There can be no more than one value i for which the ratio $\dfrac{b_i}{a_i}$ is exactly equal to the threshold. For this index i, we cannot say anything about the value z_i. So, when $A < 0$, in the order of indices, the optimal sequence of values z_i should be

$$(1, 1, \ldots, 1, z_i, -1, \ldots, -1). \tag{10}$$

General comment. A priori, we do not know whether the maximum will be attained when $A > 0$ or when $A < 0$. So, a reasonable idea is to try all possible combinations (9) and (10).

Thus, we arrive at the following algorithm.

Resulting algorithm. First, we sort all the indices in the increasing order (8) of the ratios $\dfrac{b_i}{a_i}$. Then, for each i_0 from 1 to n, we substitute the following values into the expression (7):

- $z_i = -1$ for $i < i_0$, and
- $z_i = 1$ for $i > i_0$.

The value z_{i_0} is obtained from the condition that the resulting quadratic function of the only remaining variable z_{i_0} attains the largest possible value.

Similarly, for each i_0 from 1 to n, we substitute the following values into the expression (7):

- $z_i = 1$ for $i < i_0$, and
- $z_i = -1$ for $i > i_0$.

The value z_{i_0} is obtained from the condition that the resulting quadratic function of z_{i_0} attains the largest possible value.

This way, we get $2n$ different values of the expression (7). The largest of these values is the desired maximum $\overline{e}$ of the expression (7).

Comment. Sorting requires time $O(n \cdot \log(n))$; see, e.g.,[1]. The remaining part of the algorithm requires computing $1n$ values, each of which requires $O(n^2)$ operations, so the overall time is $O(n^3)$ – very feasible.

How to find the values b_j? The above algorithm assumes that we know the values b_j. But in general, all we know are the values a_i and a_{ij}. How can we find the values b_j based on this information?

Our objective is to minimize the remaining part of the quadratic form – i.e., the terms which are not covered by the algorithm. For each i and j:

- The original coefficient at $z_i \cdot z_j$ was a_{ii} for $i = j$ and $2a_{ij}$ for $i \neq j$.
- In the form (7), we have the coefficient $a_i \cdot b_i$ for $i = j$ and $a_i \cdot b_j + a_j \cdot b_i$ for $i \neq j$.
- Thus, in the remaining part, the coefficient at $z_i \cdot z_j$ is $a_{ii} - a_i \cdot b_i$ for $i = j$ and $2a_{ij} - (a_i \cdot b_j + a_j \cdot b_i)$ for $i \neq j$.

For this remainder, the above-mentioned straightforward estimate of the value of this term is (taking into account that the upper bound for each $|z_i|$ is 1):

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \left| a_{ij} - \frac{a_i \cdot b_j + a_j \cdot b_i}{2} \right|. \tag{11}$$

We want to find the values b_j that minimize this estimate. The expression (11) is a convex function, and there exist feasible algorithms for minimizing convex functions. Thus, computing the values b_j is also feasible.

5. Conclusions

In many practical situations, the only information that we have about each measurement error Δx_i is the upper bound Δ_i on its absolute value: $|\Delta x_i| \leq \Delta_i$. In such situations, we do not know which values Δx_i are more probable and which are less probable. In this case, once we know the result $\widetilde{x}_i$ of measuring the corresponding quantity, the only conclusion that we can make about the actual (unknown) value x_i of this quantity is that this value is located in the interval $[\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$. Since the measurement results $\widetilde{x}_i$ are, in general, different from the actual values x_i, the result $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ of processing the measured values is different from the value $y = f(x_1, \ldots, x_n)$ that we would get if we could process the actual values of the corresponding quantities. A practically important question is to gauge the difference $\Delta y \stackrel{\text{def}}{=} \widetilde{y} - y$, i.e., the error of the corresponding indirect measurement.

In most real-life situations, the measurement errors are relatively small, so in the Taylor expansion of the formula for Δy, we can safely ignore terms which are of higher order in Δx_i, and keep only linear and quadratic terms. Sometimes, the measurement errors are so small that even the quadratic terms can be safely ignored. For such cases, there exist feasible algorithms for gauging Δy.

However, in general, when quadratic terms cannot be ignored, the problem of gauging Δy is NP-hard. This means that, in general, it is not possible to have a feasible algorithm that would always compute the exact upper bound on $|\Delta y|$; at best, we can find upper bounds which are not always exact.

In this paper, we describe two metrologically motivated methods for improving the resulting upper bounds.

Acknowledgments

This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes).

The authors are thankful to the anonymous referees for valuable suggestions.

References

1. T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT Press (Cambridge, Massachusetts, 2009).
2. L. Jaulin, M. Kiefer, O. Didrit, and E. Walter, *Applied Interval Analysis, with Examples in Parameter and State Estimation, Robust Control, and Robotics* (Springer, London, 2001).
3. V. Kreinovich, A. Lakeyev, J. Rohn, and P. Kahl, *Computational Complexity and Feasibility of Data Processing and Interval Computations* (Kluwer, Dordrecht, 1998).
4. G. Mayer, *Interval Analysis and Automatic Result Verification* (de Gruyter, Berlin, 2017).
5. R. E. Moore, R. B. Kearfott, and M. J. Cloud, *Introduction to Interval Analysis* (SIAM, Philadelphia, 2009).
6. H. T. Nguyen, V. Kreinovich, B. Wu, and G. Xiang, *Computing Statistics under Interval and Fuzzy Uncertainty* (Springer Verlag, Berlin, Heidelberg, 2012).

7. C. Papadimitriou, *Computational Complexity* (Addison-Wesley, Reading, Massachusetts, 1994).
8. S. G. Rabinovich, *Measurement Errors and Uncertainties: Theory and Practice* (Springer, New York, 2005).