In System Identification, Interval (and Fuzzy) Estimates Can Lead to Much Better Accuracy than the Traditional Statistical Ones: General Algorithm and Case Study

Sergey I. Kumkov
Institute of Mathematics and Mechanics
Ural Branch, Russian Academy of Sciences, and
Ural Federal University
Ekaterinburg, Russia, kumkov@imm.uran.ru

Vladik Kreinovich and Andrzej Pownuk
Computational Science Program
University of Texas at El Paso, El Paso, TX 79968, USA
ampownuk@utep.edu, vladik@utep.edu

***Abstract*—In many real-life situations, we know the upper bound of the measurement errors, and we also know that the measurement error is the joint result of several independent small effects. In such cases, due to the Central Limit Theorem, the corresponding probability distribution is close to Gaussian, so it seems reasonable to apply the standard Gaussian-based statistical techniques to process this data – in particular, when we need to identify a system. Yes, in doing this, we ignore the information about the bounds, but since the probability of exceeding them is small, we do not expect this to make a big difference on the result. Surprisingly, it turns out that in some practical situations, we get a much more accurate estimates if we, vice versa, take into account the bounds – and ignore all the information about the probabilities. In this paper, we explain the corresponding algorithms. and we show, on a practical example, that using this algorithm can indeed lead to a drastic improvement in estimation accuracy.**

I. Formulation of the Problem

System identification: a general problem. In many practical situations, we are interested in a quantity y which is difficult – or even impossible – to measure directly. This difficulty and/or impossibility may be technical: e.g.:

- while we can directly measure the distance between the two buildings by simply walking there,
- there is no easy way to measure the distance to a nearby start by flying there.

In other cases, the impossibility comes from the fact that we are interested in predictions – and, of course, today we cannot measure tomorrow's temperature.

To estimate the value of such a difficult-to-directly-measure quantity y, a natural idea is:

- to find easier-to-measure quantities $x_1, \ldots, x_n$ that are related to y by a known dependence $y = f(x_1, \ldots, x_n)$, and then
- to use the results $\widetilde{x}_i$ of measuring these auxiliary quantities to estimate y as $\widetilde{y} \stackrel{\text{def}}{=} f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.

For example:

- We can find the distance to a nearby star by measuring the direction to this star in two seasons, when the Earth is at different sides of the Sun, and the angle is thus slightly different.
- To predict tomorrow's temperature, we can measure the temperature and wind speed and direction at different locations today, and use the general equations for atmospheric dynamics to estimate tomorrow's temperature.

In some cases, we already know the dependence $y = f(x_1, \ldots, x_n)$. In many other situations, we know the general form of this dependence, but there are some parameters that we need to determine experimentally. In other words, we know that

$$y = f(a_1, \ldots, a_m, x_1, \ldots, x_n) \tag{1}$$

for some parameters $a_1, \ldots, a_m$ that need to be experimentally determined.

For example, we may know that the dependence of y on x_1 is linear, i.e., $y = a \cdot x_1 + b$, but we do not know the exact values of the corresponding parameters a and b.

In general, the problem of finding the parameters a_j is known as the problem of *system identification*.

What information we use for system identification. To identify a system, i.e., to find the values of the parameters a_j, we can use the results $\widetilde{y}_k$ and $\widetilde{x}_{ki}$ of measuring the quantities y and x_i in several different situations $k = 1, \ldots, K$.

How do we identify the system: need to take measurement uncertainty into account. Most information comes from measurements, but measurements are not 100% accurate: in general, the measurement result $\widetilde{x}$ is somewhat different from the actual (unknown) value x of the corresponding quantity: $\Delta x \stackrel{\text{def}}{=} \widetilde{x} - x \neq 0$; see, e.g., [12].

As a result, while we know that for every k, the corresponding (unknown) exact values y_k and x_{ki} are related by the dependence (1):

$$y_k = f(a_1, \ldots, a_m, x_{k1}, \ldots, x_{kn}), \tag{2}$$

a similar relation between the approximate values $\widetilde{y}_k \approx y_k$ and $\widetilde{x}_{ki} \approx x_{ki}$ is only approximate:

$$\widetilde{y}_k \approx f(a_1, \ldots, a_m, \widetilde{x}_{k1}, \ldots, \widetilde{x}_{kn}).$$

It is therefore important to take this uncertainty into account when estimating the values of the parameters $a_1, \ldots, a_m$.

How can we describe the uncertainty? In all the cases, we should know the bound Δ on the absolute value of the measurement error: $|\Delta x| \le \Delta$; see, e.g., [12]. This means that only values Δx from the interval $[-\Delta, \Delta]$ are possible.

If this is the only information we have then, based on the measurement result $\widetilde{x}$, the only information that we have about the unknown actual value x is that this value belongs to the interval $[\widetilde{x} - \Delta, \widetilde{x} + \Delta]$. There are many techniques for processing data under such interval uncertainty; this is known as *interval computations*; see, e.g., [3], [9].

Ideally, it is also desirable to know how frequent are different values Δx within this interval. In other words, it is desirable to know the probabilities of different values $\Delta x \in [-\Delta, \Delta]$.

A usual way to get these probabilities is to take into account that in many measurement situations, the measurement uncertainty Δx comes from many different independent sources. It is known that for large N, the distribution of the sum of N independent random variables becomes close to the normal (Gaussian) distribution – and tends to it when $N \to \infty$. This result – known as the Central Limit Theorem (see, e.g., [14]) – explain the ubiquity of normal distributions. It is therefore reasonable to assume that the actual distribution is Gaussian – and this is what most practitioners do in such situations [12].

Two approximations, two options. A seemingly minor problem with the Gaussian distribution is that it is, strictly speaking, *not* located on any interval: for this distribution, the probability of measurement error Δx to be in any interval – no matter how far away from Δ – is non-zero.

From this viewpoint, the assumption that the distribution is Gaussian is an approximation. It seems like a very good approximation, since for normal distribution with means 0 and standard deviation σ:

- the probability to be outside the 3σ interval $[-3\sigma, 3\sigma]$ is very small, approximately 0.1%, and
- the probability for it to be outside the 6σ interval is about 10^{-8}, practically negligible.

Yes, when we use Gaussian distributions, we ignore the information about the bounds, but, at first glance, since the difference is small, this should not affect the measurement results.

At first glance, the opposite case – when we keep the bounds but ignore all the information about probabilities, maybe add imprecise (fuzzy) expert information about possible values of Δx (see, e.g., [4], [10], [16]) – should be much worse.

What we found. Our results show, somewhat surprisingly, that the opposite is true: that if ignore the probabilistic information and use only interval (or fuzzy) information, we get much more accurate estimates for the parameters a_j than in the usual statistical methodology.

This may not be fully surprising, since there are theoretical results showing that asymptotically, interval bounds can be better; see, e.g., [15]. However, the drastic improvement in accuracy was somewhat unexpected.

The structure of the paper. First, we describe the algorithm that we used, both the general algorithm and the specific algorithm corresponding to the linear case. After that, we show the results of applying this algorithm.

II. System Identification under Interval Uncertainty: General Algorithm

Formulation of the problem in the interval case. For each pattern $k = 1, \ldots, K$, we know the measurement results $\widetilde{y}_k$ and $\widetilde{x}_{ki}$, and we know the accuracies Δ_k and Δ_{ki} of the corresponding measurements. Thus, we know that:

- the actual (unknown) value y_k belongs to the interval

$$[\underline{y}_k, \overline{y}_k] = [\widetilde{y}_k - \Delta_k, \widetilde{y}_k + \Delta_k];$$

and

- the actual (unknown) value x_{ki} belongs to the interval

$$[\underline{x}_{ki}, \overline{x}_{ki}] = [\widetilde{x}_{ki} - \Delta_{ki}, \widetilde{x}_{ki} + \Delta_{ki}].$$

We need to find the values $a_1, \ldots, a_m$ for which, for every k, some values $x_{ki} \in [\underline{x}_{ki}, \overline{x}_{ki}]$, the quantity $f(a_1, \ldots, a_m, x_{k1}, \ldots, x_{kn})$ belongs to the interval $[\underline{y}_k, \overline{y}_k]$.

Specifically, for each j from 1 to m, we would like to find the range $[\underline{a}_j, \overline{a}_j]$ of all possible values of the corresponding parameter a_j.

What happens in the statistical case. In the statistical case, we use the Least Squares method [14] and find the values $\widetilde{a}_1, \ldots, \widetilde{a}_m$ that minimize the sum of the squares of all the discrepancies:

$$\sum_{k=1}^{K} (\widetilde{y}_k - f(a_1, \ldots, a_m, \widetilde{x}_{k1}, \ldots, \widetilde{x}_{kn}))^2 \to \min_{a_1, \ldots, a_m}.$$

Possibility of linearization. Let us denote $\Delta a_j \stackrel{\text{def}}{=} \widetilde{a}_j - a_j$, where $\widetilde{a}_j$ are the least-squares estimates. In these terms, we have $a_j = \widetilde{a}_j - \Delta a_j$ and $x_{ki} = \widetilde{x}_{ki} - \Delta x_{ki}$. Thus, the corresponding value y_k has the form

$$y_k = f(a_1, \ldots, a_n, x_{k1}, \ldots, x_{kn}) = \tag{3}$$

$$f(\widetilde{a}_1 - \Delta a_1, \ldots, \widetilde{a}_m - \Delta a_m, \widetilde{x}_{k1} - \Delta x_{k1}, \ldots, \widetilde{x}_{kn} - \Delta x_{kn}).$$

The measurement errors Δx_{ki} are usually relatively small. As a result, the difference between the least-squared values $\widetilde{a}_j$ and the actual (unknown) values a_j is also small. Thus, we can expand the expression (3) in Taylor series and keep only linear terms in this expansion. This results in:

$$y_k = Y_k - \sum_{j=1}^{m} b_j \cdot \Delta a_j - \sum_{i=1}^{n} b_{ki} \cdot \Delta x_{ki}, \tag{4}$$

where we denoted

$$Y_k \stackrel{\text{def}}{=} f(\widetilde{a}_1, \ldots, \widetilde{a}_m, \widetilde{x}_{k1}, \ldots, \widetilde{x}_{kn}), \tag{5}$$

$$b_j \stackrel{\text{def}}{=} \frac{\partial f}{\partial a_j}_{|a_1=\widetilde{a}_1,\ldots,a_m=\widetilde{a}_m,x_{k1}=\widetilde{x}_{k1},\ldots,x_{km}=\widetilde{x}_{kn}}, \tag{6}$$

and

$$b_{ki} \stackrel{\text{def}}{=} \frac{\partial f}{\partial x_{ki}}_{|a_1=\widetilde{a}_1,\ldots,a_m=\widetilde{a}_m,x_{k1}=\widetilde{x}_{k1},\ldots,x_{kn}=\widetilde{x}_{kn}}. \tag{7}$$

We want to make sure that for some $\Delta x_{ki} \in [-\Delta_{ki}, \Delta_{ki}]$, the value (4) belongs to the interval $[\underline{y}_k, \overline{y}_k]$. Thus, we want to make sure that for each k, the range $[\underline{Y}_k, \overline{Y}_k]$ of all possible values of the expression (4) when $\Delta x_{ki} \in [-\Delta_{ki}, \Delta_{ki}]$ has a non-empty intersection with the interval $[\underline{y}_k, \overline{y}_k]$.

Let us thus find the expression for the range $[\underline{Y}_k, \overline{Y}_k]$. One can easily see that when $\Delta x_{ki} \in [-\Delta_{ki}, \Delta_{ki}]$, the value of the product $b_{ki} \cdot \Delta x_{ki}$ ranges from $-|b_{ki}| \cdot \Delta_{ki}$ to $|b_{ki}| \cdot \Delta_{ki}$. Thus, the smallest possible value $\underline{Y}_k$ and the largest possible value $\overline{Y}_k$ of the expression (4) are equal to:

$$\underline{Y}_k = Y_k - \sum_{j=1}^{m} b_j \cdot \Delta a_j - \sum_{i=1}^{n} |b_{ki}| \cdot \Delta_{ki}, \tag{8}$$

and

$$\overline{Y}_k = Y_k - \sum_{j=1}^{m} b_j \cdot \Delta a_j + \sum_{i=1}^{n} |b_{ki}| \cdot \Delta_{ki}. \tag{9}$$

One can easily check that the two intervals $[\underline{y}_k, \overline{y}_k]$ and $[\underline{Y}_k, \overline{Y}_k]$ intersect if and only if:

- the lower endpoint of the first interval does not exceed the upper endpoint of the second interval, and
- the lower endpoint of the second interval does not exceed the upper endpoint of the first interval,

i.e., if $\underline{y}_k \le \overline{Y}_k$ and $\underline{Y}_k \le \overline{y}_k$.

These equalities are linear in terms of the unknowns. So, the corresponding problem of finding the smallest and largest possible values of a_j becomes a particular case of optimizing a linear function under linear inequalities. For this class of problems – known as *linear programming* problems – there are known efficient algorithms; see, e.g., [8].

Thus, we arrive at the following algorithm.

Resulting algorithm. We are given:

- the expression $f(a_1, \ldots, a_m, x_1, \ldots, x_n)$ with unknown parameters a_j, and
- K measurement patterns.

For each pattern k, we know:

- the measurement results $\widetilde{y}_k$ and $\widetilde{x}_{ki}$, and
- the accuracies Δ_k and Δ_{ki} of the corresponding measurements.

Based on these inputs, we first use the Least Squares method to find the estimates $\widetilde{a}_1, \ldots, \widetilde{a}_m$. Then, we compute the values $\underline{y}_k = \widetilde{y}_k - \Delta_k$, $\overline{y}_k = \widetilde{y}_k + \Delta_k$, and the values (5), (6), and (7).

After that, for each j_0, we find the desired value $\underline{a}_{j_0}$ as the solution to the following linear programming problem: minimize a_{j_0} under the constraints that for all k, we have

$$\underline{y}_k \le Y_k - \sum_{j=1}^{m} b_j \cdot \Delta a_j + \sum_{i=1}^{n} |b_{ki}| \cdot \Delta_{ki}$$

and

$$Y_k - \sum_{j=1}^{m} b_j \cdot \Delta a_j - \sum_{i=1}^{n} |b_{ki}| \cdot \Delta_{ki} \le \overline{y}_k.$$

The value $\overline{a}_{j_0}$ can be found if we maximize a_{j_0} under the same $2K$ constraints.

How to use these formulas to estimate y? What if we now need to predict the value y corresponding to given values $x_1, \ldots, x_m$? In this case,

$$y = f(a_1, \ldots, a_m, x_1, \ldots, x_n) =$$

$$f(\widetilde{a}_1 - \Delta a_1, \ldots, \widetilde{a}_m - \Delta a_m, x_1, \ldots, x_n) =$$

$$\widetilde{y} - \sum_{j=1}^{M} B_j \cdot \Delta a_j,$$

where we denoted

$$\widetilde{y} = f(\widetilde{a}_1, \ldots, \widetilde{a}_m, x_1, \ldots, x_n)$$

and

$$B_j \stackrel{\text{def}}{=} \frac{\partial f}{\partial a_j}_{|a_1=\widetilde{a}_1,\ldots,a_m=\widetilde{a}_m,x_1,\ldots,x_n}.$$

In this case:

- the smallest possible value $\underline{y}$ of y can be found by minimizing the linear combination $\widetilde{y} - \sum\limits_{j=1}^{M} B_j \cdot \Delta a_j$ under the above constraints; and
- the largest possible value $\overline{y}$ of y can be found by maximizing the same linear combination $\widetilde{y} - \sum\limits_{j=1}^{M} B_j \cdot \Delta a_j$ under the above constraints.

What if we underestimated the measurement inaccuracy? When we applied this algorithm to several specific situations, in some cases, to our surprise, it turned out that the constraints were inconsistent. This means that we underestimated the measurement inaccuracy.

Since measuring y is the most difficult part, most probably we underestimated the accuracies of measuring y. If we denote the ignored part of the y-measuring error by ε, this means that, instead of the original bounds Δ_k on $|\Delta y_k|$, we should have bounds $\Delta_k + \varepsilon$. In this case:

- instead of the original values $\underline{y}_k = \widetilde{y}_k - \Delta_k$ and $\overline{y}_k = \widetilde{y}_k + \Delta_k$,
- we should have new bounds $\widetilde{y}_k - \Delta_k - \varepsilon$ and $\widetilde{y}_k + \Delta_k + \varepsilon$.

It is reasonable to look for the smallest possible values $\varepsilon > 0$ for which the constrains will become consistent. Thus, we

arrive at the following linear programming problem: minimize $\varepsilon > 0$ under the constraints

$$\widetilde{y}_k - \Delta_k - \varepsilon \le Y_k - \sum_{j=1}^{m} b_j \cdot \Delta a_j + \sum_{i=1}^{n} |b_{ki}| \cdot \Delta_{ki}$$

and

$$Y_k - \sum_{j=1}^{m} b_j \cdot \Delta a_j - \sum_{i=1}^{n} |b_{ki}| \cdot \Delta_{ki} \le \widetilde{y}_k + \Delta_k + \varepsilon.$$

III. System Identification under Interval Uncertainty: Simplest Case of Linear Dependence on One Variable

Description of the simplest case. Let us consider the simplest case when there is only one variable x (i.e., $n = 1$), and the dependence on this variable is linear, i.e.,

$$y = a \cdot x + b.$$

In this case:

- we have K measurement results $\widetilde{x}_k$ with accuracy Δ_k, resulting in intervals $[\underline{x}_k, \overline{x}_k]$, and similarly,
- we have intervals $[\underline{y}_k, \overline{y}_k]$ of possible values of y_k.

Based on this information, we need to find the set of all possible values of the pairs (a, b). In particular, we need to find the ranges of possible values of a and b.

Why we need to consider this case separately. Linear programming is feasible, but its algorithms are intended for a general case and thus, for the case when we have few unknowns, usually run for too long. In such situations, it is often possible to find faster techniques.

Finding bounds on a: analysis of the problem. Let us first consider the cases when $a > 0$. (The case when $a < 0$ can be handled similarly – or the same by replacing a with $-a$ and x_k with $-x_k$.)

In this case, the set of possible values of $a \cdot x_k + b$ when $x_k \in [\underline{x}_k, \overline{x}_k]$ has the form $[a \cdot \underline{x}_k + b, a \cdot \overline{x}_k + b]$. We want to make sure that this interval intersects with $[\underline{y}_k, \overline{y}_k]$, i.e., that for every k, we have

$$a \cdot \underline{x}_k + b \le \overline{y}_k \text{ and } \underline{y}_k \le a \cdot \overline{x}_k + b.$$

Thus, once we know a, we have the following lower bounds and upper bounds for b:

$$\underline{y}_k - a \cdot \overline{x}_k \le b \text{ and } b \le \overline{y}_k - a \cdot \underline{x}_k.$$

Such a value b exists if and only if every lower bound for b is smaller than or equal to every upper bound for b, i.e., if and only if, for every k and ℓ, we have

$$\underline{y}_k - a \cdot \overline{x}_k \le \overline{y}_\ell - a \cdot \underline{x}_\ell,$$

i.e., equivalently,

$$\overline{y}_\ell - \underline{y}_k \ge a \cdot (\underline{x}_\ell - \overline{x}_k).$$

- When the difference $\underline{x}_\ell - \overline{x}_k$ is positive, we divide the above inequality by this difference and get an upper bound on a:
$$a \le \frac{\overline{y}_\ell - \underline{y}_k}{\underline{x}_\ell - \overline{x}_k};$$
- When this difference is negative, after division, we get a lower bound on a:
$$a \ge \frac{\overline{y}_\ell - \underline{y}_k}{\underline{x}_\ell - \overline{x}_k}.$$

Thus, the range $[\underline{a}, \overline{a}]$ for a goes from the largest of the lower bounds to the smallest of the upper bounds. So, we arrive at the following formulas.

Resulting range for a. The resulting range for a is $[\underline{a}, \overline{a}]$, where:

$$\underline{a} = \max_{k,\ell:\ \underline{x}_\ell < \overline{x}_k} \frac{\overline{y}_\ell - \underline{y}_k}{\underline{x}_\ell - \overline{x}_k};$$

$$\overline{a} = \min_{k,\ell:\ \underline{x}_\ell > \overline{x}_k} \frac{\overline{y}_\ell - \underline{y}_k}{\underline{x}_\ell - \overline{x}_k}.$$

Range for b: analysis of the problem. For $a > 0$, we need to satisfy, for each k, the inequalities

$$a \cdot \underline{x}_k + b \le \overline{y}_k \text{ and } \underline{y}_k \le a \cdot \overline{x}_k + b.$$

Equivalently, we get

$$a \cdot x_k \le \overline{y}_k - b \text{ and } \overline{y}_k - b \le a \cdot \overline{x}_k.$$

By dividing these inequalities by a coefficient at a, we have the following bounds for a:

- for all k for which $\underline{x}_k > 0$, we get an upper bound
$$a \le \frac{\overline{y}_k}{\underline{x}_k} - \frac{1}{\underline{x}_k} \cdot b;$$
- for all k for which $\underline{x}_k < 0$, we get a lower bound
$$\frac{\overline{y}_k}{\underline{x}_k} - \frac{1}{\underline{x}_k} \cdot b \le a;$$
- for all k for which $\overline{x}_k > 0$, we get a lower bound
$$\frac{\underline{y}_k}{\overline{x}_k} - \frac{1}{\overline{x}_k} \cdot b \le a;$$
- for all k for which $\overline{x}_k > 0$, we get an upper bound
$$a \le \frac{\underline{y}_k}{\overline{x}_k} - \frac{1}{\overline{x}_k} \cdot b.$$

Thus, we get lower bounds $A_p + B_p \cdot b \le a$ and upper bounds $a \le C_q + D_q \cdot b$. These inequalities are consistent if every lower bound is smaller than or equal than every upper bound, i.e., when $A_p + B_p \cdot b \le C_q + D_q \cdot b$, or, equivalently, when $(D_q - B_p) \cdot b \ge A_p - C_q$. So, similarly to the a-case, we arrive at the following formulas:

Resulting range for b. The range for b is equal to $[\underline{b}, \overline{b}]$, where

$$\underline{b} = \max_{p,q:\ D_q > B_p} \frac{A_p - C_q}{D_q - B_p};$$

$$\overline{b} = \max_{p,q:\ D_q < B_p} \frac{A_p - C_q}{D_q - B_p}.$$

What if we underestimated the measurement inaccuracy. In this case, instead of the original bounds $\underline{y}_k$ and $\overline{y}_k$, we get the new bounds $\underline{y}_k - \varepsilon$ and $\overline{y}_k + \varepsilon$. Thus, instead of the original difference $\overline{y}_\ell - \underline{x}_k$, we get a new difference $(\overline{y}_\ell - \underline{y}_k) - \varepsilon$. The lower and upper bounds for a are thus as follows:

- When the difference $\underline{x}_\ell > \overline{x}_k$, we get
$$a \le \frac{\overline{y}_\ell - \underline{y}_k}{\underline{x}_\ell - \overline{x}_k} + \frac{2}{\underline{x}_\ell - \overline{x}_k} \cdot \varepsilon;$$
- When this difference is negative, after division, we get a lower bound on a:
$$a \ge \frac{\overline{y}_\ell - \underline{y}_k}{\underline{x}_\ell - \overline{x}_k} + \frac{2}{\underline{x}_\ell - \overline{x}_k} \cdot \varepsilon;.$$

Thus, we get lower bounds $A_p + B_p \cdot \varepsilon \le a$ and upper bounds $a \le C_q + D_q \cdot \varepsilon$. These inequalities are consistent if every lower bound is smaller than or equal than every upper bound, i.e., when $A_p + B_p \cdot \varepsilon \le C_q + D_q \cdot \varepsilon$, or, equivalently, when $(D_q - B_p) \cdot \varepsilon \ge A_p - C_q$. So, similarly to the a- an b-cases, we arrive at the following formulas.

The desired lower bound for ε for b is equal to the largest of the lower bounds, i.e., to

$$\varepsilon = \max_{p,q:\ D_q > B_p} \frac{A_p - C_q}{D_q - B_p}.$$

IV. Case Study

Description of the case study. One of the important engineering problems is the problem of storing energy. For example, solar power and wind turbines provide access to large amounts of renewable energy, but this energy is not always available – the sun goes down, the wind dies – and storing it is difficult. Similarly, electric cars are clean, but the need to store energy forces us to spend a lot of weight on the batteries.

Therefore, it is desirable to develop batteries with high energy density. One of the most promising directions is using molten salt batteries, including liquid metal batteries. These batteries offer high energy density and high power density.

To properly design these batteries, we need to analyze how the heat of fusion – i.e., the energy needed to melt the material – depends on the melting temperature. It is known that this dependence is linear.

Results. On Fig. 1, we show the results of our analysis.

It turns out that the bounds on y coming from our method are an order of magnitude smaller that the 2σ-bounds coming from the traditional statistical analysis; see [13] for details. The paper [13] also contains the description of the set of all possible pairs (a, b), i.e., all pairs which are consistent with all the measurement results.

A similar improvement was observed in other applications as well. A similar – albeit not so drastic – improvement was observed in other applications ranging from catalysis and to mechanics; see, e.g., [1], [2], [5], [6], [7], [11].

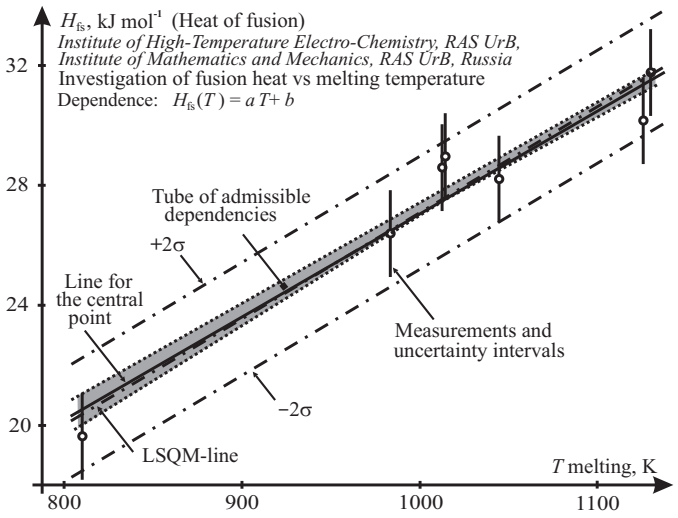

Fig. 1.

V. Conclusions

Traditional engineering techniques for estimating uncertainty of the results of data processing are based on the assumption that the measurement errors are normally distributed. In practice, the distribution of measurement errors is indeed often close to normal, so, in principle, we can use the traditional techniques to gauge the corresponding uncertainty.

In many practical situations, however, we also have an additional information about measurement uncertainty: namely, we also know the upper bounds Δ on the corresponding measurement errors. As a result, once we know the measurement result $\widetilde{x}$, we can compute the interval $[\widetilde{x} - \Delta, \widetilde{x} + \Delta]$ which is guaranteed to contain the actual (unknown) value of the measured quantity. Once we know these intervals, we can use interval computations techniques to estimate the accuracy of the result of data processing. For example, for linear models, we can use linear programming techniques to compute the corresponding bounds.

Which of the two approaches lead to more accurate estimate:

- the traditional approach, in which we ignore the upper bounds and only consider the probability distributions, or
- the interval approach, in which we only take into account the upper bounds and ignore the probabilistic information?

Previous theoretical analysis (see, e.g., [15]) shows that, in general, asymptotically, when the number of measurements n increases, the interval estimates become more accurate than the probabilistic ones.

In this paper, we show, on the example of system identification, that for several reasonable practical situations, interval techniques indeed lead to much more accurate estimates than the statistical ones – even when we have only 7 measurement results. Thus, our recommendation is that in situations when

we also know upper bounds, and we have a reasonable number of measurement results, it is beneficial to use interval techniques – since they lead to more accurate estimates.

For linear models with two parameters, we also provide a new interval-based algorithm for finding the ranges of these parameters, an algorithm which is much faster than the general linear programming techniques.

ACKNOWLEDGMENTS

This work was supported in part by the Russian Foundation for Basic Research grant 15-01-07909, National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721, and by an award "UTEP and Prudential Actuarial Science Academy and Pipeline Initiative" from Prudential Foundation.

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

REFERENCES

[1] P. A. Arkhipov, S. I. Kumkov et al., "Estimation of Pb activity in double systems Ph-Sb and Pb-Bi", *Rasplavy*, 2012, No. 5, pp. 43–52 (in Russian).

[2] S. V. Glakovsky and S. I Kumkov, "Application of approximation methods ot anlaysis of pecularities of breaking-up and to forecasting break-resitstibility of high-strength steell", In: *Mathematical Modeling of Systems and Processes*, Proceedings of the Perm State Technical University, 1997, No. 5, pp. 26–34 (in Russian).

[3] L. Jaulin, M. Kiefer, O. Dicrit, and E. Walter, *Applied Interval Analysis*, Springer, London, 2001.

[4] G. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic*, Prentice Hall, Upper Saddle River, New Jersey, 1995.

[5] S. I. Kumkov, "Processing the experimental data on ion conductivity of molten electrolyte by the interval analysis method", *Rasplavy*, 2010, No. 3, pp. 86–96 (in Russian).

[6] S. I. Kumkov, "Estimation of spring stiffness under conditions of uncertainty: interval approach", *Proceedings of the 1st International Workshop on Radio Electronics and Information Technologies REIT'2017*, Ekaterinburg, Russia, March 15, 2017 (in Russian).

[7] S. I. Kumkov and Yu. V. Mikushina, "Interval approach to identification of catalytical process parameters", *Reliable Computing*, 2013, Vol. 19, pp. 197–214.

[8] D. G. Luenberger and Y. Ye, *Linear and Nonlinear Programming*, Springer, Cham, Switzerland, 2016.

[9] R. E. Moore, R. B. Kearfott, and M. J. Cloud, *Introduction to Interval Analysis*, SIAM, Philadelphia, 2009.

[10] H. T. Nguyen and E. A. Walker, *A First Course in Fuzzy Logic*, Chapman and Hall/CRC, Boca Raton, Florida, 2006.

[11] A. M. Potapov, S. I Kumkov, and Y. Sato, "Processing of experimental data on viscosity under one-sided character of measuring errors", *rasplavy*, 2010, No. 3, pp. 55–70 (in Russian).

[12] S. G. Rabinovich, *Measurement Errors and Uncertainty: Theory and Practice*, Springer Verlag, Berlin, 2005.

[13] A. A. Redkin, Yu. P. Zalkov, I. V. Korzun, O. G. Reznitskih, T. V. Yaroslavela, and S. I Kumkov, "Heat capacity of molten halides", *Journal of Physical Chemistry*, 2015, Vol. 119, pp. 509–512.

[14] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.

[15] G. W. Walster and V. Kreinovich, "For unknown-but-bounded errors, interval estimates are often better than averaging", *ACM SIGNUM Newsletter*, 1996, Vol. 31, No. 2, pp. 6–19.

[16] L. A. Zadeh, "Fuzzy sets", *Information and Control*, 1965, Vol. 8, pp. 338–353.