For example, to find the resistance R, we measure current I and voltage V, and then use the known relation $R = V/I$ to estimate resistance as $\widetilde{R} = \widetilde{V}/\widetilde{I}$.

Computing an estimate for y based on the results of direct measurements is called *data processing*; data processing is the main reason why computers were invented in the first place, and data processing is still one of the main uses of computers as number crunching devices.

Comment. In this paper, for simplicity, we consider the case when the relation between x_i and y is known exactly; in some practical situations, we only known an approximate relation between x_i and y.

Why interval computations? From computing to probabilities to intervals. Measurement are never 100% accurate, so in reality, the actual value x_i of i-th measured quantity can differ from the measurement result $\widetilde{x}_i$. Because of these *measurement errors* $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$, the result $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ of data processing is, in general, different from the actual value $y = f(x_1, \ldots, x_n)$ of the desired quantity y [12].

It is desirable to describe the error $\Delta y \stackrel{\text{def}}{=} \widetilde{y} - y$ of the result of data processing. To do that, we must have some information about the errors of direct measurements.

What do we know about the errors Δx_i of direct measurements? First, the manufacturer of the measuring instrument must supply us with an upper bound Δ_i on the measurement error. If no such upper bound is supplied, this means that no accuracy is guaranteed, and the corresponding "measuring instrument" is practically useless. In this case, once we performed a measurement and got a measurement result $\widetilde{x}_i$, we know that the actual (unknown) value x_i of the measured quantity belongs to the interval $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$, where $\underline{x}_i = \widetilde{x}_i - \Delta_i$ and $\overline{x}_i = \widetilde{x}_i + \Delta_i$.

In many practical situations, we not only know the interval $[-\Delta_i, \Delta_i]$ of possible values of the measurement error; we also know the probability of different values Δx_i within this interval. This knowledge underlies the traditional engineering approach to estimating the error of indirect measurement, in which we assume that we know the probability distributions for measurement errors Δx_i.

In practice, we can determine the desired probabilities of different values of Δx_i by comparing the results of measuring with this instrument with the results of measuring the same quantity by a standard (much more accurate) measuring instrument. Since the standard measuring instrument is much more accurate than the one use, the difference between these two measurement results is practically equal to the measurement error; thus, the empirical distribution of this difference is close to the desired probability distribution for measurement error. There are two cases, however, when this determination is not done:

- First is the case of cutting-edge measurements, e.g., measurements in fundamental science. When a Hubble telescope detects the light from a distant

galaxy, there is no "standard" (much more accurate) telescope floating nearby that we can use to calibrate the Hubble: the Hubble telescope is the best we have.

- The second case is the case of measurements on the shop floor. In this case, in principle, every sensor can be thoroughly calibrated, but sensor calibration is so costly – usually costing ten times more than the sensor itself – that manufacturers rarely do it.

In both cases, we have no information about the probabilities of Δx_i; the only information we have is the upper bound on the measurement error.

In this case, after we performed a measurement and got a measurement result $\widetilde{x}_i$, the only information that we have about the actual value x_i of the measured quantity is that it belongs to the interval $\mathbf{x}_i = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$. In such situations, the only information that we have about the (unknown) actual value of $y = f(x_1, \ldots, x_n)$ is that y belongs to the range $\mathbf{y} = [\underline{y}, \overline{y}]$ of the function f over the box $\mathbf{x}_1 \times \ldots \times \mathbf{x}_n$:

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

For continuous functions $f(x_1, \ldots, x_n)$, this range is an interval. The process of computing this interval range based on the input intervals $\mathbf{x}_i$ is called *interval computations*; see, e.g., [4].

Interval computations techniques: brief reminder. Historically the first method for computing the enclosure for the range is the method which is sometimes called "straightforward" interval computations. This method is based on the fact that inside the computer, every algorithm consists of elementary operations (arithmetic operations, min, max, etc.). For each elementary operation $f(a, b)$, if we know the intervals $\mathbf{a}$ and $\mathbf{b}$ for a and b, we can compute the exact range $f(\mathbf{a}, \mathbf{b})$. The corresponding formulas form the so-called *interval arithmetic*. For example,

$$[\underline{a}, \overline{a}] + [\underline{b}, \overline{b}] = [\underline{a} + \underline{b}, \overline{a} + \overline{b}]; \quad [\underline{a}, \overline{a}] - [\underline{b}, \overline{b}] = [\underline{a} - \overline{b}, \overline{a} - \underline{b}];$$

$$[\underline{a}, \overline{a}] \cdot [\underline{b}, \overline{b}] = [\min(\underline{a} \cdot \underline{b}, \underline{a} \cdot \overline{b}, \overline{a} \cdot \underline{b}, \overline{a} \cdot \overline{b}), \max(\underline{a} \cdot \underline{b}, \underline{a} \cdot \overline{b}, \overline{a} \cdot \underline{b}, \overline{a} \cdot \overline{b})].$$

In straightforward interval computations, we repeat the computations forming the program f step-by-step, replacing each operation with real numbers by the corresponding operation of interval arithmetic. It is known that, as a result, we get an enclosure $\mathbf{Y} \supseteq \mathbf{y}$ for the desired range.

In some cases, this enclosure is exact. In more complex cases (see examples below), the enclosure has excess width.

There exist more sophisticated techniques for producing a narrower enclosure, e.g., a centered form method. However, for each of these techniques, there are cases when we get an excess width. Reason: as shown in [5], the problem of computing the exact range is known to be NP-hard even for polynomial functions $f(x_1, \ldots, x_n)$ (actually, even for quadratic functions f).

Comment. NP-hard means, crudely speaking, that no feasible algorithm can compute the exact range of $f(x_1, \ldots, x_n)$ for all possible polynomials $f(x_1, \ldots, x_n)$ and for all possible intervals $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

What we are planning to do? In this paper, we analyze a specific interval computations problem – when we use traditional statistical data processing algorithms $f(x_1, \ldots, x_n)$ to process the results of direct measurements.

Error Estimation for Traditional Statistical Data Processing Algorithms under Interval Uncertainty: Known Results

Formulation of the problem. When we have n results $x_1, \ldots, x_n$ of repeated measurement of the same quantity (at different points, or at different moments of time), traditional statistical approach usually starts with computing their sample average $E_x = (x_1 + \ldots + x_n)/n$ and their (sample) variance

$$V_x = \frac{(x_1 - E_x)^2 + \ldots + (x_n - E_x)^2}{n} \tag{1}$$

(or, equivalently, the sample standard deviation $\sigma = \sqrt{V}$). If, during each measurement i, we measure the values x_i and y_i of two different quantities x and y, then we also compute their (sample) covariance

$$C_{x,y} = \frac{(x_1 - E_x) \cdot (y_i - E_y) + \ldots + (x_n - E_x) \cdot (y_n - E_y)}{n}; \tag{2}$$

see, e.g., [12].

As we have mentioned, in real life, we often do not know the exact values of the quantities x_i and y_i, we only know the intervals $\mathbf{x}_i$ of possible values of x_i and the intervals $\mathbf{y}_i$ of possible values of y_i. In such situations, for different possible values $x_i \in \mathbf{x}_i$ and $y_i \in \mathbf{y}_i$, we get different values of E_x, E_y, V_x, and $C_{x,y}$. The question is: what are the intervals $\mathbf{E}_x$, $\mathbf{V}_x$, and $\mathbf{C}_{x,y}$ of possible values of E_x, V_x, and $C_{x,y}$?

The practical importance of this question was emphasized, e.g., in [9, 10] on the example of processing geophysical data.

Comment: the problem reformulated in terms of set-valued random variables. Traditional statistical data processing means that we assume that the measured values x_i and y_i are samples of random variables, and based on these samples, we are estimating the actual average, variance, and covariance.

Similarly, in case of interval uncertainty, we can say that the intervals $\mathbf{x}_i$ and $\mathbf{y}_i$ coming from measurements are samples of interval-valued random variables, and we are interested in estimating the actual (properly defined) average, variance, and covariance of these interval-valued random variables.

Bounds on E. For E_x, the straightforward interval computations leads to the exact range:

$$\mathbf{E}_x = \frac{\mathbf{x}_1 + \ldots + \mathbf{x}_n}{n}, \text{ i.e., } \underline{E}_x = \frac{\underline{x}_1 + \ldots + \underline{x}_n}{n}, \text{ and } \overline{E}_x = \frac{\overline{x}_1 + \ldots + \overline{x}_n}{n}.$$