Data Fusion Is More Complex Than Data Processing: A Proof

Robert Alvarez, Salvador Ruiz, Martine Ceberio, and Vladik Kreinovich
Department of Computer Science, University of Texas at El Paso
500 W. University, El Paso, TX 79968, USA
rjalvarez@miners.utep.edu, sruiz13@miners.utep.edu, mceberio@utep.edu, vladik@utep.edu

What is data processing: a brief reminder. In many practical situations, we are interested in the value of a quantity y that is difficult – or even impossible – to measure directly. For example, we may be interested in tomorrow's temperature. Since we cannot measure this quantity directly, we can measure it indirectly. Namely, we find easier-to-measure quantities $x_1, \ldots, x_n$ that are related to y by a known dependence $y = f(x_1, \ldots, x_n)$. Then, we measure the values x_i and apply the algorithm $f(x_1, \ldots, x_n)$ to the measurement results $\widetilde{x}_i$, producing an estimate $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ for y. This is known as *data processing.*

Measurements are never absolutely accurate. In many cases, all we know is the upper bound Δ_i on the absolute value of the measurement error $\widetilde{x}_i - x_i$. In such cases, after the measurement, all we know is that $x_i \in [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$. In this case, it is desirable to find the range of all possible values of $y = f(x_1, \ldots, x_n)$. In general, computing this range is NP-hard, but there are cases when computable is feasible: e.g., if $f(x_1, \ldots, x_n)$ is a *Single Use Expression* (SUE), in which each variable occurs only one, (e.g., $x_1 + x_2^{x_3}$) [2].

What is data fusion: a brief reminder. To describe the state of an object, we need to know the values of the physical quantities $x_1, \ldots, x_m$ that characterize this object. To determine this state, we can measure all these quantities. Usually, the quantities are not completely independent: there are constraints that relate them, and these constraints can help to decrease inaccuracy. For example, if we know that $x_1 \in [0.9, 1.1]$, $x_2 \in [0.8, 1.0]$ and $|x_1 - x_2| \le 0.01$, then we can conclude that $x_1 \in [0.9, 1.01]$. This decreasing-of-inaccuracy combination of several measurement results is known as *data fusion.*

Problem and what we do. Empirical evidence shows that, in general, data fusion is more time-consuming than data processing. In this talk, we *prove* that data fusion is indeed more complex than data processing. Specifically, we prove that even if all the constraints are described by SUE expressions, data fusion is still, in general, NP-hard. Since for SUE, data processing is feasible, this means that data fusion is indeed more complex.

Proof. Let us consider the variables $x_1, \ldots, x_n, y_1, \ldots, y_n$, and y, and let us assume that we only measure x_i, and that the variables are related by SUE constraints $x_i = y_i$ and $y = \dfrac{1}{n} \cdot \displaystyle\sum_{i=1}^n x_i^2 - \left(\dfrac{1}{n}\displaystyle\sum_{i=1}^n y_i\right)^2$. Under these constraints, the range of y is equal to the range of the sample variance $\dfrac{1}{n} \cdot \displaystyle\sum_{i=1}^n x_i^2 - \left(\dfrac{1}{n}\displaystyle\sum_{i=1}^n x_i\right)^2$ under interval uncertainty, and it is known that the problem of computing this range is NP-hard; see, e.g., [3].

[1] D. Dubois, H. Prade, and R. R. Yager, "Computation of intelligent fusion operations based on constraint fuzzy arithmetic", *Proceedings of the 1998 IEEE International Conference on Fuzzy Systems FUZZ-IEEE'98*, Anchorage, Alaska, May 4–9, 1998, Vol. 1, pp. 767–772.

[2] V. Kreinovich, A. Lakeyev, J. Rohn, and P. Kahl, *Computational Complexity and Feasibility of Data Processing And Interval Computations*, Kluwer, Dordrecht, 1998.

[3] H. T. Nguyen, V. Kreinovich, B. Wu, and G. Xiang, *Computing Statistics under Interval and Fuzzy Uncertainty*, Springer Verlag, Berlin, Heidelberg, 2012.