Detecting Outliers under Interval Uncertainty: A New Algorithm Based on Constraint Satisfaction

Evgeny Dantsin
Alexander Wolpert
Department of Computer Science
Roosevelt University
Chicago, IL 60605, USA
{edantsin,awolpert}@roosevelt.edu

Martine Ceberio
Gang Xiang
Vladik Kreinovich
Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968, USA
{mceberio,vladik}@cs.utep.edu

Abstract

In many application areas, it is important to detect outliers. The traditional engineering approach to outlier detection is that we start with some "normal" values $x_1, \ldots, x_n$, compute the sample average E, the sample standard deviation σ, and then mark a value x as an outlier if x is outside the k_0-sigma interval $[E - k_0 \cdot \sigma, E + k_0 \cdot \sigma]$ (for some pre-selected parameter k_0). In real life, we often have only interval ranges $[\underline{x}_i, \overline{x}_i]$ for the normal values $x_1, \ldots, x_n$. In this case, we only have intervals of possible values for the bounds $L \stackrel{\text{def}}{=} E - k_0 \cdot \sigma$ and $U \stackrel{\text{def}}{=} E + k_0 \cdot \sigma$. We can therefore identify outliers as values that are outside all k_0-sigma intervals, i.e., values which are outside the interval $[\underline{L}, \overline{U}]$. In general, the problem of computing $\underline{L}$ and $\overline{U}$ is NP-hard; a polynomial-time algorithm is known for the case when the measurements are sufficiently accurate, i.e., when "narrowed" intervals $\left[\widetilde{x}_i - \dfrac{1+\alpha^2}{n} \cdot \Delta_i, \widetilde{x}_i + \dfrac{1+\alpha^2}{n} \cdot \Delta_i\right]$ – where $\alpha = 1/k_0$ and $\Delta_i \stackrel{\text{def}}{=} (\underline{x}_i - \overline{x}_i)/2$ is the interval's half-width – do not intersect with each other. In this paper, we use constraint satisfaction to show that we can efficiently compute $\underline{L}$ and $\overline{U}$ under a weaker (and more general) condition that neither of the narrowed intervals is a proper subinterval of another narrowed interval.

Keywords: Outliers, Interval Uncertainty, Constraint Satisfaction.

1 Formulation of the problem

1.1 Outlier detection is important

In many application areas, it is important to detect *outliers*, i.e., unusual, abnormal values; see, e.g., [3]. In medicine, unusual values may indicate disease; in geophysics, abnormal values may indicate a mineral deposit or an erroneous measurement result; in structural integrity testing, abnormal values may indicate faults in a structure, etc.

The traditional engineering approach to outlier detection (see, e.g., [5]) is as follows:

- first, we collect measurement results $x_1, \ldots, x_n$ corresponding to normal situations;
- then, we compute the sample average $E \stackrel{\text{def}}{=} \dfrac{1}{n} \cdot \sum\limits_{i=1}^{n} x_i$ of these normal values and the (sample) standard deviation $\sigma = \sqrt{V}$, where $V \stackrel{\text{def}}{=} M - E^2$ and $M \stackrel{\text{def}}{=} \dfrac{1}{n} \cdot \sum\limits_{i=1}^{n} x_i^2$;
- finally, a new measurement result x is classified as an outlier if it is outside

the interval $[L, U]$ (i.e., if either $x < L$ or $x > U$), where $L \stackrel{\text{def}}{=} E - k_0 \cdot \sigma$, $U \stackrel{\text{def}}{=} E + k_0 \cdot \sigma$, and $k_0 > 1$ is some preselected value (most frequently, $k_0 = 2$, 3, or 6).

1.2 Outlier detection under interval uncertainty

In some practical situations, we only have intervals $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$ of possible values of x_i. This happens, for example, if instead of observing the actual value x_i of the random variable, we observe the value $\widetilde{x}_i$ measured by an instrument with a known upper bound Δ_i on the measurement error; then, the actual (unknown) value is within the interval $\mathbf{x}_i = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$. For different values $x_i \in \mathbf{x}_i$, we get different bounds L and U. Possible values of L form an interval – we will denote it by $\mathbf{L} \stackrel{\text{def}}{=} [\underline{L}, \overline{L}]$; possible values of U form an interval $\mathbf{U} \stackrel{\text{def}}{=} [\underline{U}, \overline{U}]$. In other words, we arrive at the following computation problem:

GIVEN:

- an integer $n \geq 1$;
- n intervals $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$;
- a real number $k_0 > 1$.

COMPUTE the intervals

$$\mathbf{L} \stackrel{\text{def}}{=} \{L(x_1, \ldots, x_n) : x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\};$$

$$\mathbf{U} \stackrel{\text{def}}{=} \{U(x_1, \ldots, x_n) : x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\};$$

where:

$$L \stackrel{\text{def}}{=} E - k_0 \cdot \sigma, \quad U \stackrel{\text{def}}{=} E + k_0 \cdot \sigma,$$

$$E \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} x_i, \quad \sigma \stackrel{\text{def}}{=} \sqrt{M - E^2}, \text{ and}$$

$$M \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2.$$

How do we now detect outliers? There are two possible approaches to this question: we can detect *possible* outliers and we can detect *guaranteed* outliers:

- a value x is a possible outlier if it is located outside one of the possible k_0-sigma intervals $[L, U]$ (but is may be inside some other possible interval $[L, U]$);
- a value x is a guaranteed outlier if it is located outside all possible k_0-sigma intervals $[L, U]$.

Which approach is more reasonable depends on a possible situation:

- if our main objective is not to miss an outlier, e.g., in structural integrity tests, when we do not want to risk launching a spaceship with a faulty part, it is reasonable to look for possible outliers;
- if we want to make sure that the value x is an outlier, e.g., if we are planning a surgery and we want to make sure that there is a micro-calcification before we start cutting the patient, then we would rather look for guaranteed outliers.

The two approaches can be described in terms of the endpoints of the intervals $\mathbf{L}$ and $\mathbf{U}$:

- A value x is guaranteed to be normal – i.e., it is not a possible outlier – if x belongs to the *intersection* of all possible intervals $[L, U]$, i.e., to the interval $[\overline{L}, \underline{U}]$.
- A value x is possibly normal – i.e., it is not a guaranteed outlier – if x belongs to the *union* of all possible intervals $[L, U]$, i.e., to the interval $[\underline{L}, \overline{U}]$.

So, to detect outliers under interval uncertainty, we must compute the bounds $\underline{L}$, $\overline{U}$, $\overline{L}$, and $\underline{U}$.

1.3 Detecting outliers under interval uncertainty: what is known

In [3, 4], it was shown that there exist efficient algorithms for computing the bounds $\overline{L}$ and $\underline{U}$ corresponding to possible outliers, but the computation of bounds $\underline{L}$ and $\overline{U}$ corresponding to guaranteed outliers is, in general, NP-hard. It was also shown that if $1+(1/k_0)^2 < n$

(which is true, e.g., if $k_0 > 1$ and $n \geq 2$), then the maximum of U (correspondingly, the minimum of L) is always attained at some combination of endpoints of the intervals $\mathbf{x}_i$; thus, in principle, to determine the values $\overline{U}$ and $\underline{L}$, it is sufficient to try all 2^n combinations of values $\underline{x}_i$ and $\overline{x}_i$.

Efficient algorithms are known for the case when all the interval midpoints ("measured values") $\widetilde{x}_i \stackrel{\text{def}}{=} (\underline{x}_i + \overline{x}_i)/2$ are definitely different from each other, in the sense that the "narrowed" intervals

$$\left[\widetilde{x}_i - \frac{1+\alpha^2}{n} \cdot \Delta_i, \widetilde{x}_i + \frac{1+\alpha^2}{n} \cdot \Delta_i\right]$$

– where $\alpha = 1/k_0$ and $\Delta_i \stackrel{\text{def}}{=} (\underline{x}_i - \overline{x}_i)/2$ is the interval's half-width – do not intersect with each other.

1.4 What we plan to do

In this paper, we use constraint satisfaction techniques to extend known efficient algorithms to a more general case when no two narrowed intervals are proper subsets of one another.

This is a more general case because if they do not intersect, them, of course, they cannot be proper subsets of one another – in the sense that one of them is a subset of the interior of the second one.

2 First Idea: Reduction to $\overline{U}$

When we replace each x_i with $x'_i = -x_i$, we thus replace E with $E' = -E$ while σ remains unchanged. Thus, we replace L with $L' = -U$ and U with $U' = -L$. So, if we know how to compute $\overline{U}$, we can compute $\underline{L}$ as follows:

- first, we apply the algorithm for computing $\overline{U}$ to the intervals

$$\mathbf{x}'_1 = -\mathbf{x}_1, \ldots, \mathbf{x}'_n = -\mathbf{x}_n;$$

- then, we invert the sign of the resulting value $\overline{U}'$: $\underline{L} = -\overline{U}'$.

In view of this reduction, in the following text, we only need to describe how to compute $\overline{U}$.

3 Main Idea: Reduction to Constraint Satisfaction

To find the values x_i which maximize U, we reduce the interval computation problem to the constraint satisfaction problem with the following constraints:

- for every i, if in the maximizing assignment we have $x_i = \underline{x}_i$, then replacing this value with $x_i = \overline{x}_i$ will either decrease U or leave U unchanged;
- similarly, for every i, if in the maximizing assignment we have $x_i = \overline{x}_i$, then replacing this value with $x_i = \underline{x}_i$ will either decrease U or leave U unchanged;
- finally, for every i and j, replacing both values x_i and x_j with the opposite ends of the corresponding intervals $\mathbf{x}_i$ and $\mathbf{x}_j$ will either decrease U or leave U unchanged.

We will show that the solution to the resulting constraint satisfaction problem indeed leads to an efficient algorithm for computing $\overline{U}$.

4 Algorithm

Let us first describe the algorithm itself; in the next section, we provide the justification for this algorithm.

- First, we sort of the values $\widetilde{x}_i$ into an increasing sequence. Without losing generality, we can assume that

$$\widetilde{x}_1 \leq \widetilde{x}_2 \leq \ldots \leq \widetilde{x}_n.$$

- Then, for every k from 0 to n, we compute the value $V^{(k)} = M^{(k)} - (E^{(k)})^2$ of the population variance V for the vector $x^{(k)} = (\underline{x}_1, \ldots, \underline{x}_k, \overline{x}_{k+1}, \ldots, \overline{x}_n)$, and we compute $U^{(k)} = E^{(k)} + k_0 \cdot \sqrt{V^{(k)}}$.
- Finally, we compute $\overline{U}$ as the largest of $n+1$ values $U^{(0)}, \ldots, U^{(n)}$.

To compute the values $V^{(k)}$, first, we explicitly compute $M^{(0)}$, $E^{(0)}$, and $V^{(0)} = M^{(0)} -$

$(E^{(0)})^2$. Once we know the values $M^{(k)}$ and $E^{(k)}$, we can compute

$$M^{(k+1)} = M^{(k)} + \frac{1}{n} \cdot (\underline{x}_{k+1})^2 - \frac{1}{n} \cdot (\overline{x}_{k+1})^2$$

and $E^{(k+1)} = E^{(k)} + \frac{1}{n} \cdot \underline{x}_{k+1} - \frac{1}{n} \cdot \overline{x}_{k+1}$.

5 Number of computation steps

It is well known that sorting requires $O(n \cdot \log(n))$ steps (see, e.g., a textbook [1]). Computing the initial values $M^{(0)}$, $E^{(0)}$, and $V^{(0)}$ requires linear time $O(n)$. For each k from 0 to $n-1$, we need a constant number of steps to compute the next values $M^{(k+1)}$, $E^{(k+1)}$, and $V^{(k+1)}$. Computing $U^{(k+1)}$ also requires a constant number of steps. Finally, finding the largest of $n+1$ values $U^{(k)}$ also requires $O(n)$ steps. Thus, overall, we need

$$O(n \cdot \log(n)) + O(n) + O(n) + O(n) =$$
$$O(n \cdot \log(n)) \text{ steps.}$$

It is worth mentioning that if the measurement results $\widetilde{x}_i$ are already sorted, then we only need linear time to compute $\overline{U}$.

6 Justification of the algorithm

We have already mentioned that the maximum $\overline{U}$ of the function U is attained at a vector $x = (x_1, \ldots, x_n)$ in which each value x_i is equal either to $\underline{x}_i$ or to $\overline{x}_i$.

To justify our algorithm, we need to prove that this maximum is attained at one of the vectors $x^{(k)}$ in which all the lower bounds $\underline{x}_i$ precede all the upper bounds $\overline{x}_i$. We will prove this by reduction to a contradiction. Indeed, let us assume that the maximum is attained at a vector x in which one of the lower bounds follows one of the upper bounds. In each such vector, let i be the largest upper bound index followed by the lower bound; then, in the optimal vector x, we have $x_i = \overline{x}_i$ and $x_{i+1} = \underline{x}_{i+1}$.

Since the maximum is attained for $x_i = \overline{x}_i$, replacing it with $\underline{x}_i = \overline{x}_i - 2 \cdot \Delta_i$ will either decrease the value of U or keep it unchanged. Let us describe how U changes under this replacement. Since U is defined in terms of E, M, and V, let us first describe how E, M, and V change under this replacement. In the sum for M, we replace $(\overline{x}_i)^2$ with

$$(\underline{x}_i)^2 = (\overline{x}_i - 2 \cdot \Delta_i)^2 = (\overline{x}_i)^2 - 4 \cdot \Delta_i \cdot \overline{x}_i + 4 \cdot \Delta_i^2.$$

Thus, the value M changes into $M + \Delta M_i$, where

$$\Delta M_i = -\frac{4}{n} \cdot \Delta_i \cdot \overline{x}_i + \frac{4}{n} \cdot \Delta_i^2. \quad (1)$$

The population mean E changes into $E + \Delta E_i$, where

$$\Delta E_i = -\frac{2 \cdot \Delta_i}{n}. \quad (2)$$

Thus, the value E^2 changes into $(E + \Delta E_i)^2 = E^2 + \Delta(E^2)_i$, where

$$\Delta(E^2)_i = 2 \cdot E \cdot \Delta E_i + \Delta E_i^2 =$$
$$-\frac{4}{n} \cdot E \cdot \Delta_i + \frac{4}{n^2} \cdot \Delta_i^2. \quad (3)$$

So, the variance V changes into $V + \Delta V_i$, where

$$\Delta V_i = \Delta M_i - \Delta(E^2)_i =$$
$$-\frac{4}{n} \cdot \Delta_i \cdot \overline{x}_i + \frac{4}{n} \cdot \Delta_i^2 + \frac{4}{n} \cdot E \cdot \Delta_i - \frac{4}{n^2} \cdot \Delta_i^2 =$$
$$\frac{4}{n} \cdot \Delta_i \cdot \left(-\overline{x}_i + \Delta_i + E - \frac{\Delta_i}{n} \right).$$

By definition, $\overline{x}_i = \widetilde{x}_i + \Delta_i$, hence $-\overline{x}_i + \Delta_i = -\widetilde{x}_i$. Thus, we conclude that

$$\Delta V_i = \frac{4}{n} \cdot \Delta_i \cdot \left(-\widetilde{x}_i + E - \frac{\Delta_i}{n} \right). \quad (4)$$

The function $U = E + k_0 \cdot \sigma$ attains its maximum if and only if the function $u \stackrel{\text{def}}{=} \alpha \cdot U = \alpha \cdot E + \sigma$ attains its maximum. After the change, the value u changes into

$$u + \Delta u_i = \alpha \cdot (E + \Delta E_i) + \sqrt{V + \Delta V_i},$$

so the condition $u + \Delta u_i \le u$ leads to

$$\alpha \cdot (E + \Delta E_i) + \sqrt{V + \Delta V_i} \le \alpha \cdot E + \sigma.$$

By moving the term proportional to α to the right-hand side, we conclude that $\sqrt{V + \Delta V_i} \le \sigma - \alpha \cdot \Delta E_i$. In the new inequality, the left-hand side is the new value of the standard deviation, so it is a non-negative number, hence the right-hand side is also non-negative, so we can square both sides of the inequality and conclude that

$$V + \Delta V_i \le \sigma^2 - 2 \cdot \alpha \cdot \sigma \cdot \Delta E_i + \alpha^2 \cdot (\Delta E_i)^2.$$

Moving all the terms to the left-hand side and using the fact that $V = \sigma^2$, we conclude that

$$z_i \stackrel{\text{def}}{=} \Delta V_i + 2 \cdot \alpha \cdot \sigma \cdot \Delta E_i - \alpha^2 \cdot (\Delta E_i)^2 \le 0. \quad (5)$$

Substituting the known values of ΔV_i and ΔE_i, we get:

$$z_i = \frac{4}{n} \cdot \Delta_i \cdot e_i, \quad (6a)$$

where

$$e_i = -\widetilde{x}_i + E - \frac{\Delta_i}{n} - \alpha \cdot \sigma - \alpha^2 \cdot \frac{\Delta_i}{n},$$

i.e.,

$$e_i = (E - \alpha \cdot \sigma) - \left(\widetilde{x}_i + \frac{1 + \alpha^2}{n} \cdot \Delta_i\right). \quad (6b)$$

Thus, from $z_i \le 0$, we conclude that

$$E - \alpha \cdot \sigma \le \widetilde{x}_i + \frac{1 + \alpha^2}{n} \cdot \Delta_i. \quad (7)$$

Similarly, since the maximum of u is attained for $x_{i+1} = \underline{x}_{i+1}$, replacing it with $\overline{x}_{i+1} = \underline{x}_{i+1} + 2 \cdot \Delta_{i+1}$ will either decrease the value of u or keep it unchanged. Let us describe how variance changes under this replacement. In the sum for M, we replace $(\underline{x}_{i+1})^2$ with

$$(\overline{x}_{i+1})^2 = (\underline{x}_{i+1} + 2 \cdot \Delta_{i+1})^2 =$$

$$(\underline{x}_{i+1})^2 + 4 \cdot \Delta_{i+1} \cdot \underline{x}_{i+1} + 4 \cdot \Delta_{i+1}^2.$$

Thus, the value M changes into $M + \Delta M_{i+1}$, where

$$\Delta M_{i+1} = \frac{4}{n} \cdot \Delta_{i+1} \cdot \underline{x}_{i+1} + \frac{4}{n} \cdot \Delta_{i+1}^2. \quad (8)$$

The population mean E changes into $E + \Delta E_{i+1}$, where

$$\Delta E_{i+1} = \frac{2 \cdot \Delta_{i+1}}{n}. \quad (9)$$

Thus, the value E^2 changes into

$$(E + \Delta E_{i+1})^2 = E^2 + \Delta(E^2)_{i+1},$$

where

$$\Delta(E^2)_{i+1} = 2 \cdot E \cdot \Delta E_{i+1} + \Delta E_{i+1}^2 =$$

$$\frac{4}{n} \cdot E \cdot \Delta_{i+1} + \frac{4}{n^2} \cdot \Delta_{i+1}^2. \quad (10)$$

So, the variance V changes into $V + \Delta V_{i+1}$, where

$$\Delta V_{i+1} = \Delta M_{i+1} - \Delta(E^2)_{i+1} =$$

$$\frac{4}{n} \cdot \Delta_{i+1} \cdot \underline{x}_{i+1} + \frac{4}{n} \cdot \Delta_{i+1}^2 -$$

$$\frac{4}{n} \cdot E \cdot \Delta_{i+1} - \frac{4}{n^2} \cdot \Delta_{i+1}^2 =$$

$$\frac{4}{n} \cdot \Delta_{i+1} \cdot \left(\underline{x}_{i+1} + \Delta_{i+1} - E - \frac{\Delta_{i+1}}{n}\right).$$

By definition, $\underline{x}_{i+1} = \widetilde{x}_{i+1} - \Delta_{i+1}$, hence $\underline{x}_{i+1} + \Delta_{i+1} = \widetilde{x}_{i+1}$. Thus, we conclude that

$$\Delta V_{i+1} = \frac{4}{n} \cdot \Delta_{i+1} \cdot \left(\widetilde{x}_{i+1} - E - \frac{\Delta_{i+1}}{n}\right). \quad (11)$$

Since u attains maximum at x, we have $\Delta u_{i+1} \le 0$, i.e., $z_{i+1} \le 0$, where

$$z_{i+1} \stackrel{\text{def}}{=}$$

$$\Delta V_{i+1} + 2 \cdot \alpha \cdot \sigma \cdot \Delta E_{i+1} - \alpha^2 \cdot (\Delta E_{i+1})^2. \quad (12)$$

Substituting the expressions (11) for ΔV_{i+1} and (9) for ΔE_{i+1} into this formula, we conclude that

$$z_{i+1} = \frac{4}{n} \cdot \Delta_{i+1} \cdot e_{i+1}, \quad (13a)$$

where

$$e_{i+1} \stackrel{\text{def}}{=}$$

$$-(E - \alpha \cdot \sigma) + \left(\widetilde{x}_{i+1} - \frac{1 + \alpha^2}{n} \cdot \Delta_{i+1}\right) \quad (13b)$$

and

$$E - \alpha \cdot \sigma \ge \widetilde{x}_{i+1} - \frac{1 + \alpha^2}{n} \cdot \Delta_{i+1}. \quad (14)$$

We can also change *both* x_i and x_{i+1} at the same time. In this case, from the fact that u attains the maximum at x, we conclude that $u + \Delta u \le u$, i.e., that

$$z \stackrel{\text{def}}{=} \Delta V + 2 \cdot \alpha \cdot \sigma \cdot \Delta E - \alpha^2 \cdot (\Delta E)^2. \quad (15)$$

Here, the change ΔM in M is simply the sum of the changes coming from x_i and x_{i+1}:

$$\Delta M = \Delta M_i + \Delta M_{i+1}, \qquad (16)$$

and the change ΔE in E is also the sum of the corresponding changes:

$$\Delta E = \Delta E_i + \Delta E_{i+1}. \qquad (17)$$

So, for

$$\Delta V = \Delta M - \Delta(E^2) = \Delta M - 2\cdot E\cdot \Delta E - \Delta E^2,$$

we get

$$\Delta V = \Delta M_i + \Delta M_{i+1} -$$
$$2\cdot E\cdot \Delta E_i - 2\cdot E\cdot \Delta E_{i+1} -$$
$$(\Delta E_i)^2 - (\Delta E_{i+1})^2 - 2\cdot \Delta E_i\cdot \Delta E_{i+1}.$$

Hence,

$$\Delta V = (\Delta M_i - 2\cdot E\cdot \Delta E_i - (\Delta E_i)^2) +$$
$$(\Delta M_{i+1} - 2\cdot E\cdot \Delta E_{i+1} - (\Delta E_{i+1})^2) -$$
$$2\cdot \Delta E_i\cdot \Delta E_{i+1},$$

i.e.,

$$\Delta V = \Delta V_i + \Delta V_{i+1} - 2\cdot \Delta E_i\cdot \Delta E_{i+1}. \qquad (18)$$

Substituting expressions (16), (17), and (18) into the formula (15) for z, we conclude that

$$z = \Delta V + 2\cdot\alpha\cdot\sigma\cdot\Delta E - \alpha^2\cdot(\Delta E)^2 =$$
$$\Delta V_i + \Delta V_{i+1} - 2\cdot \Delta E_i\cdot \Delta E_{i+1} +$$
$$2\alpha\cdot\sigma\cdot\Delta E_i + 2\alpha\cdot\sigma\cdot\Delta E_{i+1} -$$
$$\alpha^2\cdot(\Delta E_i)^2 - \alpha^2\cdot(\Delta E_{i+1})^2 -$$
$$2\cdot\alpha^2\cdot\Delta E_i\cdot\Delta E_{i+1}.$$

Hence,

$$z = (\Delta V_i + 2\cdot\alpha\cdot\sigma\cdot\Delta E_i - \alpha^2\cdot(\Delta E_i)^2) +$$
$$(\Delta V_{i+1} + 2\cdot\alpha\cdot\sigma\cdot\Delta E_{i+1} - \alpha^2\cdot(\Delta E_{i+1})^2) -$$
$$2\cdot(1+\alpha^2)\cdot\Delta E_i\cdot\Delta E_{i+1}.$$

From the formulas (5) and (12), we know that the first expression is z_i and that the second expression is z_{i+1}, so

$$z = z_i + z_{i+1} - 2\cdot(1+\alpha^2)\cdot\Delta E_i\cdot\Delta E_{i+1}.$$

We already have the expressions (6), (13), (2), and (9) for, correspondingly, z_i, z_{i+1}, ΔE_i, and ΔE_{i+1}, so we conclude that $z = \frac{4}{n}\cdot D(E')$, where $E' \stackrel{\text{def}}{=} E - \alpha\cdot\sigma$ and

$$D(E') \stackrel{\text{def}}{=} \Delta_i\cdot\left(E' - \left(\widetilde{x}_i + \frac{1+\alpha^2}{n}\cdot\Delta_i\right)\right) +$$
$$\Delta_{i+1}\cdot\left(-E' + \left(\widetilde{x}_{i+1} - \frac{1+\alpha^2}{n}\cdot\Delta_{i+1}\right)\right) +$$
$$2\cdot(1+\alpha^2)\cdot\frac{\Delta_i\cdot\Delta_{i+1}}{n}. \qquad (19)$$

Since $z \le 0$, we have $D(E') \le 0$ (for the value $E' = E - \alpha\cdot\sigma$ corresponding to the optimizing vector x).

The expression $D(E')$ is a linear function of E'. From (7) and (14), we know that

$$\widetilde{x}_{i+1} - \frac{1+\alpha^2}{n}\cdot\Delta_{i+1} \le E' \le \widetilde{x}_i + \frac{1+\alpha^2}{n}\cdot\Delta_i.$$

For $E' = E^- \stackrel{\text{def}}{=} \widetilde{x}_{i+1} - \frac{1+\alpha^2}{n}\cdot\Delta_{i+1}$, we have

$$D(E^-) = \Delta_i\cdot f_i + \frac{2\cdot(1+\alpha^2)}{n}\cdot\Delta_i\cdot\Delta_{i+1},$$

where

$$f_i \stackrel{\text{def}}{=} -\widetilde{x}_i + \widetilde{x}_{i+1} - \frac{1+\alpha^2}{n}\cdot\Delta_{i+1} - \frac{1+\alpha^2}{n}\cdot\Delta_i,$$

hence $D(E^-) = \Delta_i\cdot g_i$, where

$$g_i \stackrel{\text{def}}{=}$$
$$-\widetilde{x}_i + \widetilde{x}_{i+1} + \frac{1+\alpha^2}{n}\cdot\Delta_{i+1} - \frac{1+\alpha^2}{n}\cdot\Delta_i.$$

We assumed that no narrowed interval is a proper subset of any other. How can we describe this condition in algebraic terms? Let us denote $\delta_i \stackrel{\text{def}}{=} \frac{1+\alpha^2}{n}\cdot\Delta_i$; then, the i-th narrowed interval has the form $[\widetilde{x}_i - \delta_i, \widetilde{x}_i + \delta_i]$. If $[\widetilde{x}_i - \delta_i, \widetilde{x}_i + \delta_i]$ is a proper subinterval of $[\widetilde{x}_j - \delta_j, \widetilde{x}_j + \delta_j]$, this means that $\widetilde{x}_i - \delta_i > \widetilde{x}_j - \delta_j$ and $\widetilde{x}_i + \delta_i < \widetilde{x}_j + \delta_j$, i.e., equivalently, that

$$\delta_i - \delta_j < \widetilde{x}_i - \widetilde{x}_j < \delta_j - \delta_i.$$

This inequality is equivalent to $\delta_j > \delta_i$ and $|\widetilde{x}_i - \widetilde{x}_j| < \delta_j - \delta_i$. Similarly, the condition that the j-th narrowed interval is a proper subinterval of the i-th is equivalent to $\delta_j < \delta_i$ and $|\widetilde{x}_i - \widetilde{x}_j| < \delta_i - \delta_j$. Both cases can be described by a single inequality $|\widetilde{x}_i - \widetilde{x}_j| < |\delta_i - \delta_j|$. Thus, the condition that no narrowed interval can be a proper subinterval of any other narrowed interval can be described as

$$|\widetilde{x}_i - \widetilde{x}_j| \geq |\delta_i - \delta_j|. \quad (20)$$

In particular, we have $|\widetilde{x}_i - \widetilde{x}_{i+1}| \geq |\delta_i - \delta_{i+1}|$.

Let us first consider the case when

$$|\widetilde{x}_{i+1} - x_i| > |\delta_i - \delta_{i+1}|.$$

Since the values $\widetilde{x}_i$ are sorted in increasing order, we have $\widetilde{x}_{i+1} \geq \widetilde{x}_i$, hence

$$\widetilde{x}_{i+1} - \widetilde{x}_i = |\widetilde{x}_{i+1} - \widetilde{x}_i| > |\delta_i - \delta_{i+1}| \geq \delta_i - \delta_{i+1}.$$

So, we conclude that $D(E^-) > 0$.

For $E = E^+ \stackrel{\text{def}}{=} \widetilde{x}_i + \frac{1+\alpha^2}{n} \cdot \Delta_i$, we have

$$D(E^+) = \Delta_{i+1} \cdot f_{i+1} + \frac{2 \cdot (1+\alpha^2)}{n} \cdot \Delta_i \cdot \Delta_{i+1},$$

where

$$f_{i+1} \stackrel{\text{def}}{=}$$
$$-\widetilde{x}_i + \widetilde{x}_{i+1} - \frac{1+\alpha^2}{n} \cdot \Delta_{i+1} - \frac{1+\alpha^2}{n} \cdot \Delta_i,$$

hence $D(E^+) = \Delta_{i+1} \cdot g_{i+1}$, where

$$g_{i+1} \stackrel{\text{def}}{=}$$
$$-\widetilde{x}_i + \widetilde{x}_{i+1} + \frac{1+\alpha^2}{n} \cdot \Delta_i - \frac{1+\alpha^2}{n} \cdot \Delta_{i+1}.$$

Here, from $|\widetilde{x}_{i+1} - \widetilde{x}_i| > |\delta_i - \delta_{i+1}|$, we also conclude that $D(E^+) > 0$.

Since the linear function $D(E')$ is positive on both endpoints of the interval $[E^-, E^+]$, it must be positive for every value E' from this interval, which contradicts to our conclusion that $D(E') \leq 0$ for the actual value $E' = E - \alpha \cdot \sigma \in [E^-, E^+]$. This contradiction shows that the maximum of U is indeed attained at one of the values $x^{(k)}$, hence the algorithm is justified.

The general case when $|\widetilde{x}_i - \widetilde{x}_j| \geq |\delta_i - \delta_j|$ can be obtained as a limit of cases when we have strict inequality. Since the function U is continuous, the value $\overline{U}$ continuously depends on the input bounds, so by tending to a limit, we can conclude that our algorithm works in the general case as well.

Comment. It is worth mentioning that there is another polynomial-time algorithm for computing $\overline{U}$ [4] – an algorithm which computes $\overline{U}$ for the case when no *intervals* are proper subintervals of each other. That condition can be similarly described as $|\widetilde{x}_i - \widetilde{x}_j| \geq |\Delta_i - \Delta_j|$, hence that condition implies our condition (20). So, our algorithm generalizes that algorithm as well.

Acknowledgements

This work was supported in part by NASA under cooperative agreement NCC5-209, NSF grant EAR-0225670, NIH grant 3T34GM008048-20S1, and Army Research Lab grant DATM-05-02-C-0046. The authors are thankful to the anonymous referees for valuable suggestions.

References

[1] Th. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein (2001). *Introduction to Algorithms*, MIT Press, Cambridge, MA.

[2] L. Jaulin, M. Kieffer, O. Didrit, and E. Walter (2001). *Applied interval analysis: with examples in parameter and state estimation, robust control and robotics*, Springer Verlag, London.

[3] V. Kreinovich, L. Longpré, P. Patangay, S. Ferson, and L. Ginzburg (2005). Outlier Detection Under Interval Uncertainty: Algorithmic Solvability and Computational Complexity, *Reliable Computing*, 2005, Vol. 11, No. 1, pp. 59–76.

[4] V. Kreinovich, G. Xiang, S. A. Starks, L. Longpré, M. Ceberio, R. Araiza, J. Beck, R. Kandathi, A. Nayak, R. Torres, and J. Hajagos (2006). Towards combining probabilistic and interval uncertainty in engineering calculations: algorithms for computing statistics under

interval uncertainty, and their computational complexity, *Reliable Computing* (to appear).

[5] S. Rabinovich (1993). *Measurement Errors: Theory and Practice*, American Institute of Physics, New York.