Exact Bounds for Interval and Fuzzy Functions Under Monotonicity Constraints, with Potential Applications to Biostratigraphy

Emil Platon
Energy & Geoscience Institute, University of Utah

Kavitha Tupelly, Vladik Kreinovich, Scott A. Starks
Pan-American Center for Earth & Environ. Stud.
University of Texas, El Paso, TX 79968, USA

Karen Villaverde
New Mexico State U., Las Cruces, NM, 88003, USA

1. Biostratigraphy is important

- *Biostratigraphy* is concerned with the stratigraphic analysis of rocks based on their paleontologic content.
- Generally speaking, stratigraphy analyses the rock strata and is concerned with their succession and age relationship.
- All aspects of rocks as strata are, however, of concern for stratigraphy.
- The analysis of fossil can also provide useful information regarding the environment in which rocks have accumulated.
- *Example:* a coral is an unambiguous indication of a warm ocean.

2. The notion of a stratigraphic map

- *Problem:* how to determine the age of the fossil?
- *Fact:* in a normal sequence, the age increases with the depth in the well that penetrates that sequence.
- *Solution:* if the rock accumulation rate is known, the depth x at which the fossil species was found can be used to determine its age y.
- *Stratigraphic map:* the dependence between the depth x and the age y.
- Once we know the depth x and the stratigraphic map $y = f(x)$, we can determine the age y of the fossil.
- *Complication:* a stratigraphic map is different for different locations, because it depends on the geological history (of accumulation rates) at this location.

3. Main ideas behind constructing a stratigraphic map

- In every area, we have several fossils whose age y has been determined.
- For the selected fossil, we know the depth x_i at which it was found, and we know the estimated age y_i.
- Based on the points (x_i, y_i), we must find the desired dependence $y = f(x)$.
- Since deeper layers are older, we should have a monotonic (increasing) dependence $y = f(x)$ for which $y_i = f(x_i)$.
- So, ideally, we should have a monotonic function that passes through all the points.

4. The practical construction of a stratigraphic map is not that easy

- The conclusion about monotonicity is based on the *idealized assumption*:
- y_i is the age of the oldest (for wells, youngest) of many fossils of this type.
- For some types, we do have many fossils, so the oldest of these fossils represents a reasonable size sample.
- Corresponing values x_i and y_i are highly reliable.
- For other types of fossils, however, we may have only a few sample fossils of this type in a given area.
- So, x_i and y_i are not very accurate.
- As a result of this inaccuracy, in practice, it is usually impossible to have a monotonic dependence that passes exactly through all the points (x_i, y_i).

5. Traditional approach and its drawbacks

- *Problem:* few-sample data points do not fit to a monotonic curve.
- *Idea:* we select a threshold n_0 and only consider points (x_i, y_i) which came from samples of size $\geq n_0$.
- *Remaining problems:* we
  - ignore all the points (x_i, y_i) with lower accuracy, and
  - consider all the points with higher accuracy as exact, ignoring the fact that these points are not absolutely accurate.
- *Objective:* it is desirable to use the ignored information, to get a more accurate stratigraphic map.

6. Interval uncertainty

- For few-sample fossil types, the actual oldest age y_i is different from the estimated oldest age $\widetilde{y}_i$.
- Due to chaotic rock movements, the ideal depth x_i differs from the depth $\widetilde{x}_i$ at which the fossil was found.
- *Problem:* we have too few fossils to determine the probability of different values $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$ and $\Delta y_i \stackrel{\text{def}}{=} \widetilde{y}_i - y_i$.
- *What we do have:* expert estimates for the upper bound Δ_i on Δx_i.
- *Interval uncertainty:* for each fossil type i, we know the intervals $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i] = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$ and, similarly, $\mathbf{y}_i = [\underline{y}_i, \overline{y}_i]$ that contain the actual (unknown) values of x_i and y_i.

7. Fuzzy uncertainty

- *Interval information* comes from the guaranteed bound on Δx_i and Δy_i.
- *Additional information:* often, an expert can also provide bounds that contain Δy_i with a certain degree of confidence.
- Usually, we know several such bounding intervals corresponding to different degrees of confidence.
- Such a *nested family* of intervals is also called a *fuzzy set*, because it turns out to be equivalent to a more traditional definition of fuzzy set:
- If a traditional fuzzy set is given, then:
  - different *intervals* from the nested family
  - can be viewed as α-*cuts* corresponding to different levels of uncertainty α.

8. Towards the precise formulation of the problem

- *Interval uncertainty:*
  - We know the n boxes $\mathbf{x}_i \times \mathbf{y}_i$ corresponding to different types of fossils.
  - We know that the monotonic dependence $y = f(x)$ is such that $y_i = f(x_i)$ for some $(x_i, y_i) \in \mathbf{x}_i \times \mathbf{y}_i$.
  - *Objective:* to find, for every depth x, the bounds of the possible values of age $y = f(x)$ for all the dependencies that are consistent with the given data.
- *Fuzzy uncertainty:*
  - For each degree of confidence α, we must solve the problem corresponding to the α-cut intervals.
  - Thus, for each x, we want to have a fuzzy set of possible values of $f(x)$.

9. Other practical applications of the resulting mathematical problem

- *Spectral analysis*: chemical species are identified by locating local maxima of the spectra.
- *Radioastronomy:* sources of celestial radio emission and their subcomponents.
- *Elementary particles* are local maxima in the dependence of scattering intensity y on the energy x.
- *Landscape analysis:* mountain slopes.
- *Financial analysis:* growth or decline periods.
- *Clustering:* 1-D clusters are separated by local minima of the probability density.
- *Comment:* once we know how to check monotonicity, we can also find the local extrema as borders between monotonicity intervals.

10. Additional complexity

- Algorithms for solving the subproblem of checking motonoticity have been previously described.
- *Additional complexity:* it is possible to have several different ages $y_i < y_j$ for the same depth $x_i = x_j$.
- *In mathematical terms:* this means that the dependence $y = f(x)$ is not necessarily a monotonic *function*.
- It may be a *limit* of the graphs of monotonic functions in the sense of Hausdorff metric.

11. First Problem: Checking Monotonicity

- By a *monotonic dependence* f, we mean the graph of a continuous mapping $m(s) = (m_1(s), m_2(s))$ from the real line $\mathbb{R}$ to the plane $\mathbb{R}^2$ for which $t < s$ implies that $m_1(t) \le m_1(s)$ and $m_2(t) \le m_2(s)$.
- We say that a monotonic dependence f is *consistent* with a box $\mathbf{x} \times \mathbf{y}$ if $f \cap (\mathbf{x} \times \mathbf{y}) \neq \emptyset$.
- By *data* d, we mean a finite collection of boxes.
- We say that the data is *consistent* if there exists a monotonic dependence that is consistent with all its boxes.
- *Theorem.* The data d is consistent $\leftrightarrow$ for every i and j, $\overline{x}_i < \underline{x}_j$ implies $\underline{y}_i \le \overline{y}_j$.

12. Second Problem: Computing the Range of $f(x)$ for Consistent Data

- *Given:*
  - the data $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ $(1 \le i \le n)$ and
  - a real number x.
- *Objective:* to find the exact lower and upper bounds of the corresponding y over all the monotonic dependences that are consistent with this data:

$$\underline{f}(x) \stackrel{\text{def}}{=} \inf\{y : \exists f\,((x,y) \in f \,\&\, Con(f,d))\}.$$

$$\overline{f}(x) \stackrel{\text{def}}{=} \sup\{y : \exists f\,((x,y) \in f \,\&\, Con(f,d))\}.$$

- *Solution:*

$$\underline{f}(x) = \max_{i:\overline{x}_i < x} \underline{y}_i; \quad \overline{f}(x) = \min_{j:x < \underline{x}_j} \overline{y}_j.$$

13. Algorithms for checking consistency

- *Straightforward algorithm*: checking, for every i and j, whether $\overline{x}_i < \underline{x}_j$ implies $\underline{y}_i \le \overline{y}_j$.
- *Problem:* we need $O(n^2)$ comparisons – too long.
- *Equivalent condition:* $\forall i$, $\underline{y}_i \le \min_{j:\underline{x}_j \ge \overline{x}_i} \overline{y}_j$.
- *Resulting fast algorithm:*
  - Sort the values $\underline{x}_i$ ($O(n \cdot \log(n))$ steps).
  - For $i = n, \ldots, 1$, compute $M_i \stackrel{\text{def}}{=} \min(\overline{y}_n, \overline{y}_{n-1}, \ldots, \overline{y}_i)$:
  $$M_n = \overline{y}_n; \quad M_{i-1} = \min(M_i, \overline{y}_{i-1}).$$
  - For each i from 1 to n, use binary search to find $m(i)$ s.t. $\underline{x}_{m(i)-1} < \overline{x}_i \le \underline{x}_{m(i)}$.
  - For every i from 1 to n, check $\underline{y}_i \le M_{m(i)}$.
- The data is consistent $\leftrightarrow$ all these inequalities hold.
- This algorithm requires $O(n \cdot \log(n)) \ll O(n^2)$ steps.

14. Possibility of parallelization

- For a potentially unlimited number of processors, we:
  - sort the values $\underline{x}_i$ in time $O(\log(n))$;
  - compute the values M_i (solve the prefix-sum problem) in time $O(\log(n))$;
  - find all n values $m(i)$ in parallel $(O(\log(n)))$;
  - check all n inequalities in parallel (time $O(1)$).
- If we have $p < n$ processors, then we:
  - sort n values in time $O((n \cdot \log(n))/p + \log(n))$;
  - compute M_i in time $O(n/p + \log(p))$;
  - have each processor compute n/p values $m(i)$; time $O((n \cdot \log(n))/p)$;
  - have each processor checks n/p inequalities; time $O(n/p)$.
- Overall time $O\left(\dfrac{n \cdot \log(n)}{p} + \log(p)\right)$.

15. Algorithms for constructing lower and upper bounds

- The function $\underline{f}(x)$ is piecewise constant.
- When x increases, the value of $\underline{f}(x)$ changes only if when $x = \overline{x}_i$ for some i.
- So, to compute $\underline{f}(x)$, we:
  - sort $\overline{x}_i$ into an increasing sequence:

$$\overline{x}_1 \leq \overline{x}_2 \leq \ldots \leq \overline{x}_n;$$

  - compute $m_i \stackrel{\text{def}}{=} \max(\underline{y}_1, \ldots, \underline{y}_i)$.
- Overall, we need $O(n \cdot \log(n))$ steps to compute $\underline{f}(x)$.
- Similarly, we need $O(n \cdot \log(n))$ steps to compute $\overline{f}(x)$.
- In parallel, we need time $O(\log(n))$ for $p > n$ processors and $O\left(\frac{n \cdot \log(n)}{p} + \log(p)\right)$ for $p < n$.

16. From Monotonicity to More Complex Constraints

- *Situation:* in some practical problems, we also know that the rate of increase cannot be smaller than a certain value $c > 0$.
- *Question:* what are the possible values of dy/dx?
- *Mathematical formulation:* for a given interval $[a, b]$, for each of such functions f, we take a connected interval hull $co(f'([a, b]))$ of the range of the derivative.
- Then, we consider the intersection $F'([a, b])$ of these ranges over all such f.
- *Result:* $F'([a, b]) = \{x : p \le x \le q\}$, where

$$p \stackrel{\text{def}}{=} \min_{i,j: a \le \overline{x}_i \le \underline{x}_j \le b} \frac{\underline{y}_j - \overline{y}_i}{\underline{x}_j - \overline{x}_i}, \quad q \stackrel{\text{def}}{=} \min_{i,j: a \le \overline{x}_i \le \underline{x}_j \le b} \frac{\overline{y}_j - \underline{y}_i}{\underline{x}_j - \overline{x}_i}.$$

- These formulas provides a $O(n^2)$ time algorithm for computing the range.

17. Future Work

- *Future work: algorithms.*
  - *Situation:* sometimes, we also know the probabilities of different values (x_i, y_i) from the data boxes.
  - *Objective:* find the probabilities of different stratigraphic maps.
- *Future work: applications.*
  - *Task 1:* finalize actual applications of our algorithms to biostratigraphy.
  - *Task 2:* apply to other areas, including areas where fuzzy knowledge is available.

18. Acknowledgments

- This work was supported:
  - by NASA grant NCC5-209,
  - by USAF grant F49620-00-1-0365,
  - by NSF grants EAR-0112968, EAR-0225670, and EIA-0321328,
  - by Army Research Laboratories grant DATM-05-02-C-0046, and
  - by the NIH grant 3T34GM008048-20S1.
- The authors are thankful:
  - to Luc Longpré (El Paso, Texas),
  - to all the participants of the Geoinformatics meeting at the San Diego Supercomputer Center (August 13–15, 2004), and
  - to the anonymous referees.