Why Convex Combinations of Interval Endpoints: Related Explanations for Cases of Data Processing and Decision Making

Christian Servin[1], Olga Kosheleva[2], and Vladik Kreinovich[3]
[1]Computer Science and Information Technology Systems Department,
El Paso Community College, 919 Hunter, El Paso, Texas 79915, USA,
cservin1@epcc.edu
[2,3]Departments of [2]Teacher Education and [3]Computer Science
University of Texas at El Paso,
500 W. University, El Paso, Texas 79968, USA
olgak@utep.edu, vladik@utep.edu

1. Formulation of the problem

- In two phenomena, empirically, a convex combination $\gamma\cdot\overline{x}+(1-\gamma)\cdot\underline{x}$ of the endpoints of an interval $[\underline{x},\overline{x}]$ leads to the best results.
- In this talk, we explain this.
- The first phenomenon deals with data processing, namely, with one of its simplest cases – linear regression.
- There, based on the approximately known values x_i and y, we need to find the coefficients a_i for which:

$$y\approx a_0+a_1\cdot x_1+\ldots+a_n\cdot x_n.$$

- Specifically, in several situations j, we know:
  - bounds $[\underline{x}_{ij},\overline{x}_{ij}]$ for the corresponding values x_{ij} of the quantity x_i, and
  - bounds $[\underline{y}_j,\overline{y}_j]$ for the corresponding value y_j of the quantity y.
- When we know the exact values of x_{ij} and y_j, a natural way to estimate the coefficients a_i is by using the Least Squares method.

2. Formulation of the problem (cont-d)

- There, we find the values a_j for which the following expression attains the smallest possible value:

$$\sum_j (y_j - (a_0 + a_1 \cdot x_{1j} + \ldots + a_n \cdot x_{nj}))^2.$$

- It turns out that in the interval-valued case, the best results are obtained if:
  - for an appropriate $\gamma \in [0, 1]$,
  - we apply the Least Squares method to the values

$$x_{ij} = \gamma \cdot \overline{x}_{ij} + (1 - \gamma) \cdot \underline{x}_{ij} \text{ and } y_j = \gamma \cdot \overline{y}_j + (1 - \gamma) \cdot \underline{y}_j.$$

- Processing fuzzy data is, in effect, equivalent to processing α-cut intervals for each γ.
- Thus, the same technique can be thus naturally extended to the case when we know x_{ij} and y_j with fuzzy uncertainty.

3. Second phenomenon

- The second phenomenon deals with decision making in situations in which we know both:
  - approximate probabilities $\widetilde{p}_i$ of different outcomes i, and
  - possibility of different outcomes,
  - i.e., in effect, the largest possible probability $\overline{p}_i$ of these outcomes.
- It turns out that empirically:
  - the best results are obtained
  - if we based our decisions on the probabilities p_i which are equal to a convex combination of approximate probability and possibility:

$$p_i = \gamma \cdot \overline{p}_i + (1 - \gamma) \cdot \widetilde{p}_i.$$

4. First phenomenon: analysis of the problem

- We want a technique that, given an interval $[\underline{x}, \overline{x}]$, selects a value from this interval.
- Let us denote the selected value by $s(\underline{x}, \overline{x})$, where s comes from "select".
- The empirical values x_i and y are, usually, values of physical quantities.
- A numerical value of a physical quantity depends on the choice of the measuring unit and of the starting point.
- If we replace the measuring unit with another unit which is $a > 0$ times smaller, then all numerical values will multiply by a: $x \mapsto a \cdot x$.
- For example, if we replace meters with centimeters, then 1.7 m becomes $100 \cdot 1.7 = 170$ cm.

5. First phenomenon: analysis of the problem (cont-d)

- If we replace the original starting point with a new one which is b units earlier, then this value b will be added to all numerical values:

$$x \mapsto x + b.$$

- For example, we can replace the 0 point of the Celsius temperature scale with the 0 point of the Klevin scale.
- That point is approximately 273 degree earlier, so 20 C becomes $10 + 273 = 293$ K.
- In general, if we replace both the measuring unit and the starting point, we get a linear transformation $x \mapsto a \cdot x + b$.
- Both changes – of the measuring unit and of the starting point:
  - change the numerical value
  - but do not change the physical quantity itself.
- E.g., a person who is 1.7 m tall is exactly 170 cm tall.

6. First phenomenon: analysis of the problem (cont-d)

- Thus, it is reasonable to require that the selection function should not be affected by these changes.
- For example:
  - the selection corresponding to 1.7 and 1.8 m, when described in centimeters,
  - should be exactly the same as the selection corresponding to 170 and 180 cm.
- In precise terms, for general a and b and for all $\underline{x} < \overline{x}$, this natural property takes the following form:
  - if $a = s(\underline{x}, \overline{x})$,
  - then $X = s(\underline{X}, \overline{X})$, where $X = a \cdot x + b$, $\underline{X} = a \cdot \underline{x} + b$, and $\overline{X} = a \cdot \overline{x} + b$.
- In other words, we should always have:

$$s(a \cdot \underline{x} + b, a \cdot \overline{x} + b) = a \cdot s(\underline{x}, \overline{x}) + b.$$

7. Main result of this section

- Let us prove that the selection function that satisfies the above property has the desired form – of the convex combination.
- Let us denote the value $s(0,1)$ by γ.
- By definition of the selection function, it must return the value from the input interval.
- Thus, we have $\gamma \in [0,1]$, i.e., $0 \le \gamma \le 1$.
- Now, for any $\underline{x}_i < \overline{x}_i$, let us take $a = \overline{x}_i - \underline{x}_i$, $b = \underline{x}_i$, $\underline{x} = 0$ and $\overline{x} = 1$.
- Then, $a \cdot \underline{x} + b = (\overline{x}_i - \underline{x}_i) \cdot 0 + \underline{x}_i = \underline{x}_i$, and $a \cdot \overline{x} + b = (\overline{x}_i - \underline{x}_i) \cdot 1 + \underline{x}_i = \overline{x}_i$.
- Thus, the above formula takes the following form:

$$s(\underline{x}_i, \overline{x}_i) = (\overline{x}_i - \underline{x}_i) \cdot \gamma + \underline{x}_i = \gamma \cdot \overline{x}_i + (1 - \gamma) \cdot \underline{x}_i.$$

- This is exactly what we wanted to explain.

8. Second phenomenon: why do we need a different explanation?

- The above explanation only applies to physical quantities.
- It uses the fact that their numerical values depends on the choice of the measuring unit and the starting point.
- However, in the second phenomenon, we deal with probabilities – and the numerical value of probability is absolute.
- Since we cannot directly deal with probabilities, let us take into account where these probabilities are used.
- As we have mentioned earlier, these probabilities are used to make decisions.

9. Second phenomenon: why do we need a different explanation (cont-d)

- Let us therefore briefly recall:
  - how decisions are made
  - or, to be more precise, how decisions *should be* made by rational decision makers.
- Such recommended decisions are dealt with by *decision theory*.

10. Decision theory: a brief reminder

- To make appropriate decisions, it is important to properly describe people's preferences.
- For this purpose, decision theory has the notion of *utility* – that enables us to describe preferences in numerical form.
- To introduce this notion, we need to select two alternatives:
  - a very good alternative A_+ that is better than anything that can actually happen, and
  - a very bad alternative A_- that is worse than anything that can actually happen.
- Now, to define the utility of each alternative A, we need to compare this alternative, for different $p \in [0, 1]$, with *lotteries* $L(p)$ in which:
  - we get A_+ with probability p and
  - we get A_- with the remaining probability $1 - p$.

11. Decision theory: a brief reminder (cont-d)

- When $p \approx 0$, the lottery $L(p)$ is close to the very bad alterative A_-.
- We selected A_- to be worse than anything that we will actually encounter.
- So we conclude that $L(p)$ is worse than A; we will denote this by $L(p) < A$.
- Similarly, when $p \approx 1$, the lottery $L(p)$ is close to the very good alterative A_+.
- We selected A_+ to be better than anything that we will actually encounter.
- So we conclude that A is worse than $L(p)$: $A < L(p)$.
- Clearly, the larger the probability p of the getting the very good alternative, the better the lottery: if $p < q$, then $L(p) < L(q)$.

12. Decision theory: a brief reminder (cont-d)

- Thus:
  - if $L(q) < A$ and $p < q$, then we have $L(q) < A$, and
  - if $A < L(p)$ and $p < q$, then we have $A < L(q)$.
- So, the set $\{p : L(p) < A\}$ is closed under adding smaller numbers, and the set $\{p : A < L(p)\}$ is closed under adding larger numbers.
- And there can be no more than one value p for which A and $L(p)$ are equivalent:
  - when $A \sim L(p)$ and $p < q$, then $L(p) < L(q)$ implies that $A < L(q)$;
  - thus, $A \not\sim L(q)$.
- So, there exists a threshold value

$$u(A) \stackrel{\text{def}}{=} \sup\{p : L(p) < A\} = \inf\{p : A < L(p)\}.$$

13. Decision theory: a brief reminder (cont-d)

- For this value:
  - if $p < u(A)$, then $L(p) < A$, and
  - if $p > u(A)$, the $A < L(p)$.
- This threshold value is called the *utility* of the alternative A.
- Due to the above property, for every positive number $\varepsilon > 0$, no matter how small it is, we have $L(u(A) - \varepsilon) < A < L(u(A) + \varepsilon)$.
- When ε is sufficiently small, it is not possible to feel the difference between the probabilities $u(A) - \varepsilon$, $u(A)$, and $u(A) + \varepsilon$.
- In this sense, we can say that the alternative A is *equivalent* to the lottery $L(u(A))$ in which:
  - we get A_+ with probability $u(A)$ and
  - we get A_- with the remaining probability $1 - u(A)$.
- We will denote this equivalence by $A \equiv L(u(A))$.

14. Decision theory: a brief reminder (cont-d)

- Each alternative A is equivalent to the lottery $L(u(A))$, and the best lottery $L(p)$ is the lottery with the largest probability p.
- We can therefore conclude that the alternative A is better than the alternative B if and only if $u(A) > u(B)$.
- The numerical value of utility depends on our selection of A_- and A_+.
- It can be shown that if we select a different pair (A'_-, A'_+), then:
  - instead of the original utilities $u(A)$,
  - we will get $u'(A) = a \cdot u(A) + b$ for some constants $a > 0$ and b that depend only on the two pairs (A_-, A_+) and (A'_-, A'_+).
- Thus, similarly to physical quantities, utility is defined modulo a linear transformation.
- How can we use this notion to make a decision?
- Ideally, for each possible decision, we know what are possible outcomes A_i, and what is probability p_i of each outcome.

15. Decision theory: a brief reminder (cont-d)

- By using the above description, we can determine the utility u_i of each outcome.
- Thus, the result of this decision is equivalent to a lottery in which we get the outcome A_i with probability p_i.
- Each outcome is, as we have mentioned, equivalent to a lottery in which:
  - we get A_+ with probability u_i and
  - we get A_- with the remaining probability $1 - u_i$.
- So, the result of each decision is equivalent to a two-stage lottery, in which:
  - first, we select i so that each i has probability p_i, and then
  - depending on what i we selected, we select A_+ with probability u_i and A_- with the probability $1 - u_i$.
- As a result of this two-stage lottery, we get either A_+ or A_-.

16. Decision theory: a brief reminder (cont-d)

- The probability u of getting A_+ can be determined by using the law of total probability $u = p_1 \cdot u_1 + \ldots + p_n \cdot u_n$.
- In mathematical terms, this formula described the expected value of utility.
- Thus, each decision is equivalent to a lottery in which we get A_+ with probability u and A_- with the remaining probability.
- By definition of utility, this means that the utility of this possible decision is equal to u.
- Thus, we need to select a decision for which the expected utility u is the largest possible.

17. Resulting explanation

- We are interesting in recommendations to decision making.
- So, instead of the probabilities $\widetilde{p}$ and $\overline{p}$, let us consider the corresponding utilities

$$\widetilde{u} = \widetilde{p} \cdot u_+ + (1 - \widetilde{p}) \cdot u_- = \widetilde{p} \cdot (u_+ - u_-) + u_-;$$

$$\overline{u} = \overline{p} \cdot u_+ + (1 - \overline{p}) \cdot u_- = \overline{p} \cdot (u_+ - u_-) + u_-.$$

- Here:
  - u_+ is the utility of the situation when the given event occurs, and
  - u_- is the utility of the situation in which this event does not occur.

18. Resulting explanation (cont-d)

- Since $\overline{p}$ is the upper bound on the possible probabilities, we must have $\widetilde{p} \le \overline{p}$; so:
  - if the event is favorable for us, i.e., if $u_- < u_+$, then we have $\widetilde{u} \le \overline{u}$;
  - vice versa, if the event is not favorable for us, i.e., if $u_+ < u_-$, then we have $\overline{u} \le \widetilde{u}$.
- In these terms, what we want to have is the utility that we shall actually use for decision making:

$$u = p \cdot u_+ + (1 - p) \cdot u_- = p \cdot (u_+ - u_-) + u_-.$$

- If $\widetilde{u} = \overline{u}$, then it makes sense to use this value as the desired utility, i.e., to take $u = \widetilde{u} = \overline{u}$.
- In the case of $\widetilde{u} \ne \overline{u}$, we need to come up with a mapping $u = s(\widetilde{u}, \overline{u})$ that transforms the two utility values into a single utility value.

19. Resulting explanation (cont-d)

- As we have mentioned, utility is defined modulo a linear transformation.
- So, it makes sense to require that this mapping should not change if we apply some linear transformation, i.e., if we use a different pair

$$(A_-, A_+).$$

- In precise terms, this means that the function $s(\widetilde{u}, \overline{u})$ should satisfy the following requirement:

$$s(a \cdot \widetilde{u} + b, a \cdot \overline{u} + b) = a \cdot s(\widetilde{u}, \overline{u}) + b.$$

- This is exactly the same requirement as for the first phenomenon.
- We have already shown that when $u_- < u_+$ and thus, $\widetilde{u} < \overline{u}$, this requirement leads to $u = s(\widetilde{u}, \overline{u}) = \gamma \cdot \overline{u} + (1 - \gamma) \cdot \overline{u}$.
- When $u_+ < u_-$ and $\overline{u} < \widetilde{u}$, the similar derivation leads, for some γ', to $u = s(\widetilde{u}, \overline{u}) = \gamma' \cdot \widetilde{u} + (1 - \gamma') \cdot \overline{u}$.

20. Resulting explanation (cont-d)

- This leads to the same formula as for $u_- < u_+$, with $\gamma = 1 - \gamma'$.
- So, in both case, we get that formula.
- Substituting the expressions for u, $\overline{u}$, and $\widetilde{u}$ into this formula, we get

$$p \cdot (u_+ - u_-) + u_- =$$
$$\gamma \cdot (\overline{p} \cdot (u_+ - u_-) + u_-) + (1 - \gamma) \cdot (\widetilde{p} \cdot (u_+ - u_-) + u_-).$$

- If we open parentheses, we get:

$$p \cdot (u_+ - u_-) + u_- = \gamma \cdot \overline{p} \cdot (u_+ - u_-) + \gamma \cdot u_- + (1 - \gamma) \cdot \widetilde{p} \cdot (u_+ - u_-) + (1 - \gamma) \cdot u_-.$$

- One can easily see that terms proportional to u_- cancel each other, so we have

$$p \cdot (u_+ - u_-) = \gamma \cdot \overline{p} \cdot (u_+ - u_-) + (1 - \gamma) \cdot \widetilde{p} \cdot (u_+ - u_-).$$

- If we divide both sides of this equality by $u_+ - u_-$, we get:

$$p = \gamma \cdot \overline{p} + (1 - \gamma) \cdot \widetilde{p}.$$

21. Resulting explanation (cont-d)

- This is exactly the desired formula.
- Thus, we get an explanation for the second phenomenon as well.

22. Summary

- This paper explains why using a convex combination of interval endpoints gives best results in two cases
- First, in interval-based linear regression:
  - choosing a value inside each interval as a weighted average of its endpoints
  - leads to better model accuracy.
- Second, in decision making with approximate probabilities and possibilities:
  - forming probability as a weighted average of both
  - leads to better decisions.

23. Remaining open questions

- Our analysis explains why convex combinations works well.
- However, it does not explain which value γ we should chose.
- Empirically, for different applications, different values γ lead to better results.
- So how can we select the value γ which is the best for a given application?
- Sometimes, common sense can help with the selection of the parameter γ.
- For example, sometimes, we have a large amount of data and thus, our probability estimates are reasonably accurate.
- Then, it makes sense to put more weight on probabilities, i.e., to use $\gamma \approx 1$.

24. Remaining open questions (cont-d)

- On the other hand:
  - when we have a small amount of data (and thus, our probability estimates are very crude),
  - while the possibility values are provided by a highly skilled expert,
  - it makes sense to put more weight on the expert's opinion, i.e., to use $\gamma \approx 0$.
- In other applications, we do not have any such intuition.
- In such cases, time-consuming trial-and-error methods of selecting γ are, at present, the only available choice.
- It is therefore desirable to come up with a general method of determining:
  - based on the application,
  - which values γ would lead to better results.

25. Acknowledgments

This work was supported in part:

- by the US National Science Foundation grants:
  - 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science),
  - HRD-1834620 and HRD-2034030 (CAHSI Includes),
  - EAR-2225395 (Center for Collective Impact in Earthquake Science C-CIES),
- by the AT&T Fellowship in Information Technology, and
- by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Focus Program SPP 100+ 2388, Grant Nr. 501624329,