Which Distributions (or Families of Distributions) Best Represent Interval Uncertainty: Case of Permutation-Invariant Criteria

Michael Beer\textsuperscript{1}, Julio Urenda\textsuperscript{2}
Olga Kosheleva\textsuperscript{2}, and Vladik Kreinovich\textsuperscript{2}

\textsuperscript{1}Leibniz University Hannover
30167 Hannover, Germany
beer@irz.uni-hannover.de

\textsuperscript{2}University of Texas at El Paso
500 W. University
El Paso, Texas 79968, USA
jcurenda@utep.edu, olgak@utep.edu, vladik@utep.edu
1. **Interval Uncertainty Is Ubiquitous**

- An engineering design comes with numerical values of the corresponding quantities, be it:
  - the height of ceiling in civil engineering or
  - the resistance of a certain resistor in electrical engineering.

- Of course, in practice, it is not realistic to maintain the exact values of all these quantities.

- We can only maintain them with some tolerance.

- As a result, the engineers:
  - not only produce the desired (“nominal”) value $x$ of the corresponding quantity,
  - they also provide positive and negative tolerances $\varepsilon_+ > 0$ and $\varepsilon_- > 0$. 
2. Interval Uncertainty Is Ubiquitous (cont-d)

- The actual value must be in the interval $\mathbf{x} = [\underline{x}, \overline{x}]$, where $\underline{x} \equiv x - \varepsilon_-$ and $\overline{x} \equiv x + \varepsilon_+$.
- All the manufacturers need to do is to follow these interval recommendations.
- There is no special restriction on probabilities of different values within these intervals.
- These probabilities depends on the manufacturer.
- Even for the same manufacturer, they may change when the manufacturing process changes.
3. Data Processing Under Interval Uncertainty Is Often Difficult

- Interval uncertainty is ubiquitous.
- So, many researchers have considered different data processing problems under this uncertainty.
- This research area is known as *interval computations*.
- The problem is that the corresponding computational problems are often very complex.
- They are much more complex than solving similar problems under *probabilistic* uncertainty:
  - when we know the probabilities of different values within the corresponding intervals,
  - we can use Monte-Carlo simulations to gauge the uncertainty of data processing results.
4. Interval Data Processing Is Difficult (cont-d)

- A similar problem for interval uncertainty:
  - is NP-hard already for the simplest nonlinear case
  - when the whole data processing means computing the value of a quadratic function.

- It is even NP-hard to find the range of variance when inputs are known with interval uncertainty.

- This complexity is easy to understand.

- Interval uncertainty means that we may have different probability distributions on the given interval.

- So, to get guaranteed estimates, we need, in effect, to consider all possible distributions.

- And this leads to very time-consuming computations.

- For some problems, this time can be sped up, but in general, the problems remain difficult.
5. It Is Desirable to Have a Family of Distributions Representing Interval Uncertainty

- Interval computation problems are NP-hard.
- In practical terms, this means that the corresponding computations will take forever.
- So, we cannot consider all possible distributions on the interval.
- A natural idea is to consider some typical distributions.
- This can be a finite-dimensional family of distributions.
- This can be even a finite set of distributions – or even a single distribution.
- For example, in measurements, practitioners often use uniform distributions on the corresponding interval.
- This selection is even incorporated in some international standards for processing measurement results.
6. Family of Distributions (cont-d)

• Of course, we need to be very careful which family we choose.

• By limiting the class of possible distributions, we introduce an artificial “knowledge”.

• Thus, we modify the data processing results.

• So, we should select the family depending on what characteristic we want to estimate.

• We need to beware that:
  – a family that works perfectly well for one characteristic
  – may produce a completely misleading result when applied to some other desired characteristic.

• Examples of such misleading results are well known.
7. Continuous Vs. Discrete Distributions

- Usually, in statistics and in measurement theory:
  - when we say that the actual value $x$ belongs to the interval $[a, b]$,
  - we assume that $x$ can take any real value between $a$ and $b$.

- However, in practice:
  - even with the best possible measuring instruments,
  - we can only measure the value of the physical quantity $x$ with some uncertainty $h$.

- Thus, from the practical viewpoint, it does not make any sense to distinguish between $a$ and $a + h$.

- Even with the best measuring instruments, we will not be able to detect this difference.
8. Continuous Vs. Discrete (cont-d)

- From the practical viewpoint, it makes sense to divide the interval $[a, b]$ into small subintervals
  
  
  $[a, a + h], [a + h, a + 2h], \ldots$

- Within each of them the values of $x$ are practically indistinguishable.

- It is sufficient to find the probabilities $p_1, p_2, \ldots, p_n$ that the actual value $x$ is in one of the subintervals:
  
  - the probability $p_1$ that $x$ is in the first small subinterval $[a, a + h]$;
  
  - the probability $p_2$ that $x$ is in the first small subinterval $[a + h, a + 2h]$; etc.

- These probabilities should, of course, add up to 1:
  
  $$\sum_{i=1}^{n} p_i = 1.$$
9. Continuous Vs. Discrete (cont-d)

• In the ideal case, we get more and more accurate measuring instruments – i.e., $h \to 0$.

• Then, the corresponding discrete probability distributions will tend to continuous ones.

• So, from this viewpoint:
  
  – selecting a probability distribution means selecting a tuple of values $p = (p_1, \ldots, p_n)$, and
  
  – selecting a family of probability distributions means selecting a family of such tuples.
10. Example: Estimating Maximum Entropy

• Whenever we have uncertainty, a natural idea is to provide a numerical estimate for this uncertainty.

• It is known that one of the natural measures of uncertainty is Shannon’s entropy $- \sum_{i=1}^{n} p_i \cdot \log_2(p_i)$.

• In the case of interval uncertainty, we can have several different tuples.

• In general, for different tuples, entropy is different.

• As a measure of uncertainty of the situation, it is reasonable to take the largest possible value.

• Indeed, Shannon’s entropy can be defined as:
  
  – the average number of binary (“yes”-“no”) questions
  
  – that are needed to uniquely determine the situation.
11. Maximum Entropy (cont-d)

- The larger this number, the larger the initial uncertainty.
- Thus, it is natural to take the largest number of such questions as a characteristic of interval uncertainty.
- For this characteristic, we want to select a distribution:
  - whose entropy is equal to
  - the largest possible entropy of all possible probability distributions on the interval.
- Selecting such a “most uncertain” distribution is known as the Maximum Entropy approach.
- This approach has been successfully used in many practical applications.
12. Maximum Entropy (cont-d)

- It is well known that:
  - out of all possible tuples with $\sum_{i=1}^{n} p_i = 1$,
  - the entropy is the largest possible when all the probabilities are equal to each other, i.e., when $p_1 = \ldots = p_n = 1/n$.

- In the limit $h \to 0$, such distributions tend to the uniform distribution on the interval $[a, b]$.

- This is one of the reasons why uniform distributions are recommended in some measurement standards.
13. Modification of This Example

- In addition to Shannon’s entropy, there are other measures of uncertainty.
- They are usually called generalized entropy.
- For example, in many applications, practitioners use the quantity $-\sum_{i=1}^{n} p_i^\alpha$ for some $\alpha \in (0, 1)$.
- It is known that when $\alpha \to 0$, this quantity, in some reasonable sense, tends to Shannon’s entropy.
- To be more precise:
  - the tuple at which the generalized entropy attains its maximum under different condition
  - tends to the tuple at which Shannon’s entropy attains its maximum.
- The maximum of this characteristic is also attained when all the probabilities $p_i$ are equal to each other.
14. Other Examples and Idea

- A recent paper analyzed how to estimate sensitivity of Bayesian networks under interval uncertainty.
- It also turned out that:
  - if we limit ourselves to a single distribution,
  - then the most adequate result also appears if we select a uniform distribution.
- The same uniform distribution appears in many different situations, under different optimality criteria.
- This makes us think that there must be a general reason for this distribution.
- In this talk, we indeed show that there is such a reason.
15. Beyond the Uniform Distribution

- For other characteristics, other possible distributions provide a better estimate. For example:
  - if we want to estimate the smallest possible value of the entropy,
  - then the corresponding optimal value 0 is attained for several different distributions.

- Specifically, there are \( n \) such distributions corresponding to different values \( i_0 = 1, \ldots, n \).

- In each of these distributions, we have \( p_{i_0} = 1 \) and \( p_i = 0 \) for all \( i \neq i_0 \).

- In the continuous case \( h \to 0 \):
  - these probability distributions correspond to point-wise probability distributions
  - in which a certain value \( x_0 \) appears with probability 1.
16. Beyond the Uniform Distribution (cont-d)

- Similar distributions appear for several other optimality criteria.
- For example, when we minimize generalized entropy.
- How can we explain that these distributions appear as solutions to different optimization problems?
- Similar to the uniform case, there should also be a general explanation.
- A simple general explanation will indeed be provided in this talk.
17. Let Us Use Symmetries

• In general, our knowledge is based on *symmetries*, i.e., on the fact that some situations are similar.

• Indeed, if all the world’s situations were completely different, we would not be able to make any predictions.

• Luckily, real-life situations have many features in common.

• So we can use the experience of previous situations to predict future ones.

• For example, when a person drops a pen, it starts falling down with the acceleration of 9.81 m/sec^2.

• If this person moves to a different location, he or she will get the exact same result.

• This means that the corresponding physics is invariant with respect to shifts in space.
18. Let Us Use Symmetries (cont-d)

- Similarly, if the person repeats this experiment in a year, the result will be the same.
- This means that the corresponding physics is invariant with respect to shifts in time.
- Alternatively, if the person turns around a little bit, the result will still be the same.
- This means that the underlying physics is also invariant with respect to rotations, etc.
- This is a very simple example, but such symmetries are actively used in modern physics.
19. Let Us Use Symmetries (cont-d)

- Moreover, many previously proposed fundamental physical theories can be derived from symmetries:
  - Maxwell’s equations that describe electrodynamics,
  - Schroedinger’s equations that describe quantum phenomena,
  - Einstein’s General Relativity equation that describe gravity.
- Symmetries also help to explain many empirical phenomena in computing.
- From this viewpoint:
  - a natural way to look for what the two examples have in common
  - is to look for invariances that they have in common.
20. Permutations – Natural Symmetries in the Entropy Example

• We have $n$ probabilities $p_1, \ldots, p_n$.

• What can we do with them that would preserve the entropy?

• The easiest possible transformations is when we do not change the values themselves, just swap them.

• Bingo! Under such swap, the value of the entropy does not change.

• Interestingly, the above-described generalized entropy is also permutation-invariant.

• Thus, we are ready to present our general results.
21. Definitions and Results

- We say that a function \( f(p_1, \ldots, p_n) \) is permutation-invariant if for every permutation, we have
  \[
  f(p_1, \ldots, p_n) = f(p_{\pi(1)}, \ldots, p_{\pi(n)}).
  \]

- By a permutation-invariant optimization problem, we mean a problem of optimizing:
  - a permutation-invariant function \( f(p_1, \ldots, p_n) \)
  - under constraints of the type \( g_i(p_1, \ldots, p_n) = a_i \) or \( h_j(p_1, \ldots, p_n) \geq b_j \)
  - for permutation-invariant functions \( g_i \) and \( h_j \).

- **Proposition.** *If a permutation-invariant optimization problem has only one solution, then for this solution:*
  \[
  p_1 = \ldots = p_n.
  \]

- This explains why we get the uniform distribution in several cases (maximum entropy etc.)
22. Proof

- We will prove this result by contradiction.
- Suppose that the values $p_i$ are not all equal.
- This means that there exist $i$ and $j$ for which $p_i \neq p_j$.
- Let us swap $p_i$ and $p_j$, and denote the corresponding values by $p'_i$, i.e.:
  - we have $p'_i = p_j$,
  - we have $p'_j = p_i$, and
  - we have $p'_k = p_k$ for all other $k$.
- The values $p_i$ satisfy all the constraints.
- All the constraints are permutation-invariant.
- So, the new values $p'_i$ also satisfy all the constraints.
- Since the objective function is permutation-invariant, we have $f(p_1, \ldots, p_n) = f(p'_1, \ldots, p'_n)$. 
23. Proof (cont-d)

- Since the values \((p_1, \ldots, p_n)\) were optimal, the values \((p'_1, \ldots, p'_n) \neq (p_1, \ldots, p_n)\) are thus also optimal.

- This contradicts the assumption that the original problem has only one solution.

- This contradiction proves for the optimal tuple \((p_1, \ldots, p_n)\) that all the values \(p_i\) are indeed equal to each other.

- The proposition is proven.
24. Discussion

• What if the optimal solution is not unique?

• We can have a case when we have a small finite number of solutions.

• We can also have a case when we have a 1-parametric family of solutions – depending on one parameter.

• In our discretized formulation, each parameter has $n$ values, so this means that we have $n$ possible solutions.

• Similarly, a 2-parametric family means that we have $n^2$ possible solutions, etc.

• We say that a problem has a small finite number of solutions if it has $< n$ solutions.

• We say that a problem has a $d$-parametric family of solutions if it has $\leq n^d$ solutions.
25. Second Result

- Proposition.
  - If a permutation-invariant optimization problem has a small finite number of solutions,
  - then it has only one solution.

- Due to Proposition 1, in this case, the only solution is the uniform distribution $p_1 = \ldots = p_n$. 
26. Proof

- Since $\sum p_i = 1$:
  - there is only one possible solution for which $p_1 = \ldots = p_n$:
  - the solution for which $p_1 = \ldots = p_n = 1/n$.

- Thus, if the problem has more than one solution, some values $p_i$ are different from others.

- In particular, some values are different from $p_1$.

- Let $S$ denote the set of all $j$ for which $p_j = p_1$.

- Let $m$ denote the number of elements in this set.

- Since some values $p_i$ are different from $p_1$, we have $1 \leq m \leq n - 1$. 
27. Proof (cont-d)

- Due to permutation-invariance, each permutation of this solution is also a solution.
- For each \( m \)-size subset of \( \{1, \ldots, n\} \), we can have a permutation that transforms \( S \) into this set.
- Thus, it produces a new solution to the original problem.
- There are \( \binom{n}{m} \) such subsets.
- For \( 0 < m < n \), the smallest value \( n \) of \( \binom{n}{m} \) is attained when \( m = 1 \) or \( m = n - 1 \).
- Thus, if there is more than one solution, we have at least \( n \) different solutions.
- Since we assumed that we have fewer than \( n \) solutions, this means that we have only one. Q.E.D.
28. One More Result

• **Proposition.** *If a permutation-invariant optimization problem has a 1-parametric family of solutions, then:*
  
  – *this family of solutions is characterized by a real number* $c \leq 1/(n - 1)$, *for which*
  
  – *all these solutions have the following form: $p_i = c$ for* $i \neq i_0$ *and* $p_{i_0} = 1 - (n - 1) \cdot c$.

• In particular, for $c = 0$:
  
  – *we get the above-mentioned 1-parametric family of distributions for which*
  
  – *Shannon’s entropy (or generalized entropy) attain the smallest possible value.*
29. Proof

- We have shown that:
  - if in one of the solutions, for some value $p_i$ we have $m$ different indices $j$ with this value,
  - then we will have at least $\binom{n}{m}$ different solutions.

- For all $m$ from 2 to $n - 2$, this number is at least as large as $\binom{n}{2} = \frac{n \cdot (n - 1)}{2}$ and is, thus, larger than $n$.

- Since overall, we only have $n$ solutions, this means that it is not possible to have $2 \leq m \leq n - 2$.

- So, the only possible values of $m$ are 1 and $n - 1$. 
30. Proof (cont-d)

• If there was no group with \( n - 1 \) values:
  – this would mean that all the groups must have \( m = 1 \),
  – i.e., consist of only one value.

• In other words, in this case, all \( n \) values \( p_i \) would be different.

• In this case, each of \( n! \) permutations would lead to a different solution.

• So we would have \( n! > n \) solutions, but there are only \( n \) solutions.

• Thus, this case is also impossible.

• So, we do have a group of \( n - 1 \) values with the same \( p_i \).

• Then we get exactly one of the solutions described in the formulation.
31. Conclusions

- Traditionally, in engineering, uncertainty is described by a probability distribution.
- In practice, we rarely know the exact distribution.
- In many practical situations:
  - the only information we know about a quantity
  - is the interval of possible values of this quantity.
- And we have no information about the probability of different values within this interval.
- Under such interval uncertainty, we cannot exclude any mathematically possible probability distribution; so:
  - to estimate the range of possible values of the desired uncertainty characteristic,
  - we must, in effect, consider all possible distributions.
32. Conclusions (cont-d)

• Not surprisingly, for many characteristics, the corresponding computational problem becomes NP-hard.

• For some characteristics, we can provide a reasonable estimate for their desired range if:
  
  – instead of all possible distributions,
  
  – we consider only distributions from some finite-dimensional family.

• For example:
  
  – to estimate the largest possible value of Shannon’s entropy (or of its generalizations),
  
  – it is sufficient to consider only the uniform distribution.
33. Conclusions (cont-d)

• Similarly:
  – to estimate the smallest possible value of Shannon’s entropy or of its generalizations,
  – it is sufficient to consider point-wise distributions.

• Different optimality criteria lead to the same distribution – or to the same family of distributions.

• This made us think that there should be a general reason for the appearance of these families.

• In this talk, we show that indeed:
  – the appearance of these distributions and these families can be explained
  – by the fact that all the corresponding optimization problems are permutation-invariant.
34. Conclusions (cont-d)

• Thus, in the future, if a reader encounters a permutation-invariant optimization problem:
  – for which it is known that there is a unique solution
  – or that there is only a 1-parametric family of solutions,
  – then there is no need to actually solve the corresponding problem.

• In such situations, it is possible to simply use our general symmetry-based results.

• Thus, we can find a distribution (or a family of distributions) that:
  – for the corresponding characteristic,
  – best represents interval uncertainty.
35. Acknowledgments

This work was supported in part by the National Science Foundation grants:

- 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science),
- HRD-1242122 (Cyber-ShARE Center of Excellence).