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

Michael Beer[1], Julio Urenda[2], Olga Kosheleva[2][0000−0003−2587−4209], and Vladik Kreinovich[2][0000−0002−1244−1650]

[1] Leibniz University Hannover, 30167 Hannover, Germany
beer@irz.uni-hannover.de
[2] University of Texas at El Paso, El Paso TX 79968, USA
{jcurenda,olgak,vladik}@utep.edu

Abstract. In many practical situations, we only know the interval containing the quantity of interest, we have no information about the probabilities of different values within this interval. In contrast to the cases when we know the distributions and can thus use Monte-Carlo simulations, processing such interval uncertainty is difficult – crudely speaking, because we need to try all possible distributions on this interval. Sometimes, the problem can be simplified: namely, for estimating the range of values of some characteristics of the distribution, it is possible to select a single distribution (or a small family of distributions) whose analysis provides a good understanding of the situation. The most known case is when we are estimating the largest possible of Shannon's entropy: in this case, it is sufficient to consider the uniform distribution on the interval. Interesting, estimating other characteristics leads to the selection of the same uniform distribution: e.g., estimating the largest possible values of generalized entropy or of some sensitivity-related characteristics. In this paper, we provide a general explanation of why uniform distribution appears in different situations – namely, it appears every time we have a permutation-invariant optimization problem with the unique optimum. We also discuss what happens if we have an optimization problem that attains its optimum at several different distributions – this happens, e.g., when we are estimating the smallest possible value of Shannon's entropy (or of its generalizations).

Keywords: Interval uncertainty · Maximum Entropy approach · Uniform distribution · Sensitivity analysis

* 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) and HRD-1242122 (Cyber-ShARE Center of Excellence).

1 Formulation of the Problem

Interval uncertainty is ubiquitous. When an engineer designs an object, the original design comes with exact 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 $\varepsilon_- > 0$ with which we need to maintain the value of this quantity. The actual value must be in the interval $\mathbf{x} = [\underline{x}, \overline{x}]$, where $\underline{x} \stackrel{\text{def}}{=} x - \varepsilon_-$ and $\overline{x} \stackrel{\text{def}}{=} x + \varepsilon_+$.

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, and even for the same manufacturer, they may change every time the manufacturer makes some adjustments to the manufacturing process.

Data processing under interval uncertainty is often difficult. Because of the ubiquity of interval uncertainty, many researchers have considered different data processing problems under this uncertainty; this research area is known as *interval computations*; see, e.g., [5, 10, 11, 14].

The problem is that the corresponding computational problems are often very complex, much more complex than solving similar problems under *probabilistic* uncertainty – when we know the probabilities of different values within the corresponding intervals. For example, while for the probabilistic uncertainty, we can, in principle, always use Monte-Carlo simulations to understand how the input uncertainty affects the result of data processing, 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 – actually, it is even NP-hard if we want to find the range of possible values of variance in a situation when inputs are only known with interval uncertainty [8, 13].

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 – which leads to very time-consuming computations. For some problems, this time can be sped up, but in general, the problems remain difficult.

It is desirable to have a reasonably small family of distributions representing interval uncertainty. In the ideal world, we should always take into account interval uncertainty – i.e., take into account that, in principle, all mathematically possible probability distributions on the given interval are actually possible.

However, as we have just mentioned, many of the corresponding interval computation problems are NP-hard. In practical terms, this means that the corresponding computations will take forever.

Since in such situations, it is not possible to exactly take interval uncertainty into account – i.e., 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; see, e.g., [14].

Of course, we need to be very careful which family we choose: by limiting the class of possible distributions, we introduce an artificial "knowledge", and thus, modify the data processing results. So, we should select the family depending on what characteristic we want to estimate – and 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 – and we will present some such results later.

Continuous vs. discrete distributions: idealized mathematical description vs. practical description. 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, e.g., the values a and $a + h$ – even with the best measuring instruments, we will not be able to detect this difference.

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 which the values of x are practically indistinguishable.

Correspondingly, to describe the probabilities of different values x, it is sufficient to find the probabilities $p_1, p_2, \ldots, p_n$ that the actual value x is in one of these small 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: $\sum_{i=1}^{n} p_i = 1$.

In the ideal case, when we get more and more accurate measuring instruments – i.e., when $h \to 0$ – the corresponding discrete probability distributions will tend to the corresponding continuous distribution. 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.

First example of selecting a family of distributions: 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 $-\sum_{i=1}^{n} p_i \cdot \log_2(p_i)$; see, e.g., [6, 13]. When we know the probability distribution, i.e., when we know all the values p_i, then the above formula enables us to uniquely determine the corresponding entropy.

However, in the case of interval uncertainty, we can have several different tuples, and, 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: 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 – or, if needed, a family of distributions – 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; see, e.g., [6].

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

$$p_1 = \ldots = p_n = 1.$$

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, as we have mentioned, uniform distributions are recommended in some measurement standards.

Modification of this example. In addition to Shannon's entropy, there are other measures of uncertainty – which 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.

Other examples. The authors of [4] analyzed how to estimate sensitivity of Bayesian networks under interval uncertainty. It also turned out that if, for the purpose of this estimation, we limit ourselves to a single distribution, then the most adequate result also appears if we select a uniform distribution, i.e., in effect, the values $p_1 = \ldots = p_n$; see [4] for technical details.

Idea. The fact that the same uniform distribution appears in many different situations, under different optimality criteria, make us think that there must be a general reason for this distribution. In this paper, we indeed show that there is such a reason.

Beyond the uniform distribution. For other characteristics, other possible distributions provide a better estimate. For example, if instead of estimating the *largest* possible value of the entropy, 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.

Similar distributions appear for several other optimality criteria: e.g., when we minimize generalized entropy instead of minimizing Shannon's entropy. A natural question is: 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 – and a simple general explanation will indeed be provided in this paper.

2 Analysis of the Problem

What do entropy, generalized entropy, etc. have in common? We would like to come up with a general result that generalizes both the maximum entropy, the maximum generalized entropy, and other cases. To come up with such a generalization, it is reasonable to analyze what these results have in common.

Let us use symmetries. In general, our knowledge is based on *symmetries*, i.e., on the fact that some situations are similar to each other. 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.

The idea of using symmetries is well-known to many readers. However, since not everyone is very familiar with this idea, we added a brief explanation in this subsection. Readers who are well familiar with the idea of symmetry are welcome to skip the rest of this subsection, and go straight to the subsection about permutations.

So here is our brief explanation. For example, when a person drops a pen, it starts falling down to Earth with the acceleration of 9.81 m/sec^2. If this person moves to a different location and repeats the same experiment, he or she will get the exact same result. This means that the corresponding physics is invariant with respect to shifts in space.

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 invariances are actively used in modern physics (see, e.g., [1, 15]) – and moreover, many previously proposed fundamental physical theories such as:

- Maxwell's equations that describe electrodynamics,
- Schroedinger's equations that describe quantum phenomena,
- Einstein's General Relativity equation that describe gravity,

can be derived from the corresponding invariance assumptions; see, e.g., [2, 3, 7, 9].

Symmetries also help to explain many empirical phenomena in computing; see, e.g., [12]. 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.

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? In principle, we can transform the values into something else, but 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. In precise terms, both the objective function $S = -\sum_{i=1}^{n} p_i \cdot \ln(p_i)$ and the constraint $\sum_{i=1}^{n} p_i = 1$ do not change is we perform any permutation

$$\pi : \{1, \ldots, n\} \to \{1, \ldots, n\},$$

i.e., replace the values $p_1, \ldots, p_n$ with the permuted values $p_{\pi(1)}, \ldots, p_{\pi(n)}$.

Interestingly, the above-described generalized entropy is also permutation-invariant. Thus, we are ready to present our general results.

3 Our Results

Definition 1.

- *We say that a function $f(p_1, \ldots, p_n)$ is* permutation-invariant *if for every permutation $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$, 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.*

Comment. In other words, we consider the following problem:

- *given* permutation-invariant functions $f(p_1, \ldots, p_n)$, $g_1(p_1, \ldots, p_n)$, $g_2(p_1, \ldots, p_2)$, $\ldots$, $h_1(p_1, \ldots, p_n)$, $h_2(p_1, \ldots, p_2)$, $\ldots$, and values a_1, a_2, $\ldots$, b_1, b_2, $\ldots$;

– *find:* among all tuples $p = (p_1, \ldots, p_n)$ that satisfy the conditions $\sum_{i=1}^{n} p_1 = 1$,

$$g_1(p_1, \ldots, p_n) = a_1, \quad g_2(p_1, \ldots, p_n) = a_2, \quad \ldots,$$

and

$$h_1(p_1, \ldots, p_n) \geq b_1, \quad h_2(p_1, \ldots, p_n) \geq b_2, \quad \ldots$$

find the tuple with the largest (or smallest) possible value of the objective function $f(p_1, \ldots, p_n)$.

Proposition 1. *If a permutation-invariant optimization problem has only one solution, then for this solution, we have* $p_1 = \ldots = p_n$.

Discussion. This explains why we get the uniform distribution in several cases: in the maximum entropy case, in the maximum generalized entropy case, etc.

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.

Since the values p_i satisfy all the constraints, and all the constraints are permutation-invariant, the new values p'_i also satisfy all the constraints. Since the objective function is permutation-invariant, we have $f(p_1, \ldots, p_n) = f(p'_1, \ldots, p'_n)$. 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 – which contradicts to 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.

Discussion. What is 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 – i.e., a family 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.

Here are precise definitions and related results.

Definition 2.

– *We say that a permutation-invariant optimization problem with n unknowns $p_1, \ldots, p_n$ has a* small finite number of solutions *if it has fewer than n solutions.*
– *We say that a permutation-invariant optimization problem with n unknowns $p_1, \ldots, p_n$ has a d-*parametric family of solutions *if it has no more than n^d solutions.*

Proposition 2. *If a permutation-invariant optimization problem has a small finite number of solutions, then it has only one solution.*

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

Proof. Since $\sum p_i = 1$, there is only one possible solution for which $p_1 = \ldots = p_n$: the solution for which all the values p_i are equal to $1/n$.

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 the indices j for which $p_j = p_1$, and let m denote the number of elements in this set. Since some values p_i are different from p_1, we have $1 \le m \le n-1$.

Due to permutation-invariance, each permutation of this solution is also a solution. For each m-size subset of the set of n-element set of indices $\{1, \ldots, n\}$, we can have a permutation that transforms S into this set and thus, produces a new solution to the original problem. There are $\binom{n}{m}$ such subsets. For all m from 1 to $n-1$, the smallest value of the binomial coefficient $\binom{n}{m}$ is attained when $m = 1$ or $m = n-1$, and this smallest value is equal to n. Thus, if there is more than one solution, we have at least n different solutions – and since we assumed that we have fewer than n solutions, this means that we have only one. The proposition is proven.

Proposition 3. *If a permutation-invariant optimization problem has a 1-parametric family of solutions, then this family of solutions is characterized by a real number $c \le 1/(n-1)$, for which all these solutions have the following form: $p_i = c$ for all i but one and $p_{i_0} = 1 - (n-1) \cdot c$ for the remaining value i_0.*

Discussion. 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.

Proof. As we have shown in the proof of Proposition 2, 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 $\binom{n}{2} = \dfrac{n \cdot (n-1)}{2}$ and is, thus, larger than n.

Since overall, we only have n solutions, this means that it is not possible to have $2 \le m \le n-2$. So, the only possible values of m are 1 and $n-1$.

If there was no group with $n-1$ values, this would means 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 to the original problem – but we assumed that overall, 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 of the proposal, plus solutions obtained from it by permutations – which is exactly the described family.

The proposition is proven.

4 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. Thus, to estimate the range of possible values of the desired uncertainty characteristic, we must, in effect, consider all possible distributions. 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. Similarly, to estimate the smallest possible value of Shannon's entropy or of its generalizations, it is sufficient to consider point-wise distributions, in which a single value from the interval appears with probability 1. The fact that different optimality criteria lead to the same distribution – or to the same family of distributions – made us think that there should be a general reason for the appearance of these families. In this paper, 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.

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 (which may be complex to directly solve). In such situations, it is possible to simply use our general symmetry-based results for finding the corresponding solution – and thus, for finding a distribution (or a family of distributions) that, for the corresponding characteristic, best represent interval uncertainty.

Acknowledgments

The authors are greatly thankful to all the participants of the 2019 IEEE Symposium on Computational Intelligence in Engineering Solutions CIES'2019 (Xiamen, China, December 6–9, 2019) for useful thought-provoking discussions and to the anonymous referees for valuable suggestions.

References

1. R. Feynman, R. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Addison Wesley, Boston, Massachusetts, 2005.
2. A. M. Finkelstein and V. Kreinovich, "Derivation of Einstein's, Brans-Dicke and other equations from group considerations," In: Y. Choque-Bruhat and T. M. Karade (eds), *On Relativity Theory. Proceedings of the Sir Arthur Eddington Centenary Symposium, Nagpur, India 1984*, Vol. 2, World Scientific, Singapore, 1985, pp. 138–146.
3. A. M. Finkelstein, V. Kreinovich, and R. R. Zapatrin. "Fundamental physical equations uniquely determined by their symmetry groups," Lecture Notes in Mathematics, Springer-Verlag, Berlin-Heidelberg-N.Y., Vol. 1214, 1986, pp. 159–170.
4. L. He, M. Beer, M. Broggi, P. Wei, and A. T. Gomes, "Sensitivity analysis of prior beliefs in advanced Bayesian networks", *Proceedings of the 2019 IEEE Symposium Series on Computational Intelligence SSCI'2019*, Xiamen, China, December 6–9, 2019, pp. 775–782.
5. L. Jaulin, M. Kiefer, O. Didrit, and E. Walter, *Applied Interval Analysis, with Examples in Parameter and State Estimation, Robust Control, and Robotics*, Springer, London, 2001.
6. E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.
7. V. Kreinovich, "Derivation of the Schroedinger equations from scale invariance," *Theoretical and Mathematical Physics*, 1976, Vol. 8, No. 3, pp. 282–285.
8. V. Kreinovich, A. Lakeyev, J. Rohn, and P. Kahl, *Computational Complexity and Feasibility of Data Processing and Interval Computations*, Kluwer, Dordrecht, 1998.
9. V. Kreinovich and G. Liu, "We live in the best of possible worlds: Leibniz's insight helps to derive equations of modern physics", In: R. Pisano, M. Fichant, P. Bussotti, and A. R. E. Oliveira (eds.), *The Dialogue between Sciences, Philosophy and Engineering. New Historical and Epistemological Insights, Homage to Gottfried W. Leibnitz 1646–1716*, College Publications, London, 2017, pp. 207–226.
10. G. Mayer, *Interval Analysis and Automatic Result Verification*, de Gruyter, Berlin, 2017.
11. R. E. Moore, R. B. Kearfott, and M. J. Cloud, *Introduction to Interval Analysis*, SIAM, Philadelphia, 2009.
12. H. T. Nguyen and V. Kreinovich, *Applications of Continuous Mathematics to Computer Science*, Kluwer, Dordrecht, 1997.
13. H. T. Nguyen, V. Kreinovich, B. Wu, and G. Xiang, *Computing Statistics under Interval and Fuzzy Uncertainty*, Springer Verlag, Berlin, Heidelberg, 2012.
14. S. G. Rabinovich, *Measurement Errors and Uncertainties: Theory and Practice*, Springer, New York, 2005.
15. K. S. Thorne and R. D. Blandford, *Modern Classical Physics: Optics, Fluids, Plasmas, Elasticity, Relativity, and Statistical Physics*, Princeton University Press, Princeton, New Jersey, 2017.