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 probability 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, 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 use the Maximum Entropy approach and get the uniform distribution on the interval. Interesting, sensitivity analysis – which has completely different objectives – leads to selection of the same uniform distribution. 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 objective functions with the unique optimum. We also discuss what happens if there are several optima.

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

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

* 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).

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 a tolerance $\varepsilon > 0$ 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 aready 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 to consider all of them – 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. Considering all possible distributions on an interval will take forever. It is therefore desirable to look for cases when interval uncertainty can be represented by a single distribution – or at least by a reasonably small family of distributions, e.g., by finitely many distributions or by a finite-dimensional family.

Maximum entropy idea. In general, the problem of selecting a single distribution from the family of all distributions which are consistent with our knowledge (i.e., with measurement results and known general principles) is well known in data processing. A usual solution to this problem is to select the distribution which best represents the corresponding uncertainty. Some possible distributions have little uncertainty – e.g., we can have a distribution which is located at some point with probability 1. Selecting such a distribution would mislead the data

processing algorithm into thinking that we have no uncertainty at all. Similarly, selecting a distribution which is located on a proper subinterval of the original interval would be misleading – since it will lead to a decrease in perceived uncertainty. From this viewpoint, a proper selection should select a representative distribution with the largest possible uncertainty.

A reasonable measure of uncertainty is *entropy*

$$S = -\int \rho(x) \cdot \ln(\rho(x))\, dx, \tag{1}$$

where $\rho(x)$ denotes the probability density function; see, e.g., [6, 13]. So, a reasonable idea is to select, among all possible distributions, a distribution with the largest possible entropy. This idea is known as the *Maximum Entropy* approach, and it has indeed very successful in many applications; see, e.g., [6].

For interval uncertainty, maximum entropy leads to the uniform distribution. What do we get when we apply the maximum entropy approach to the case of interval uncertainty, when all we know is that the probability distribution is located in some interval $[a, b]$? In this case, we want to find a distribution $\rho(x)$ for which the entropy (1) is maximized under the condition that the overall probability is 1, i.e., that $\int_a^b \rho(x)\, dx = 1$.

The usual way to solve such constraint optimization problems is to use the Lagrange multiplier method, where the problem of optimizing a function $f(A)$ under the constraint $g(A) = 0$ is reduced to an unconstrained problem of optimizing the auxiliary function $f(A) + \lambda \cdot g(A)$, with the parameter λ (known as the *Lagrange multiplier*) to be determined from the condition that the resulting optimizing alternative A satisfy the original constraint.

In our case, this means that we maximize the auxiliary function

$$-\int \rho(x) \cdot \ln(\rho(x))\, dx + \lambda \cdot \left(\int_a^b \rho(x)\, dx - 1 \right).$$

Strictly speaking, this expression has infinitely many unknowns – namely, the values $\rho(x)$ corresponding to all possible values x. However, in practice, we can always take into account that 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 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 indistinguishable, and instead of the function $\rho(x)$, consider the probabilities $p_1, p_2, \ldots$ of the value x being in each of these intervals. In these terms, the entropy takes the form $S = -\sum_{i=1}^{n} p_i \cdot \ln(p_i)$, the requirement that probabilities add up to 1 take the form $\sum_{i=1}^{n} p_i = 1$, and the resulting equivalent unconstrained optimization problem takes the form of maximizing the expression

$$-\sum_{i=1}^{n} p_i \cdot \ln(p_i) + \lambda \cdot \left(\sum_{i=1}^{n} p_i - 1 \right).$$

To find the maximum value of this expression, the usual idea is to differentiate this expression with respect to each unknown p_i and equate the resulting derivative to 0. As a result, we get the formula $-\ln(p_i)-1+\lambda = 0$, hence $\ln(p_i) = 1+\lambda$, and $p_i = \exp(1+\lambda)$.

This value is the same for all i. Thus, the probability to be in each of the small subintervals is the same – i.e., we select a uniform distribution on the original interval $[0, 1]$.

Need to go beyond the maximum entropy approach: first argument. In different practical problems, we can have different objective functions. For example, in many cases, it is important to know how sensitive is the system to different perturbations. In this case, it is also desirable to select one distribution (or at least a reasonably small family of distributions).

The corresponding objective function is very different from the entropy. However, interestingly, it turned out that the best choice of a representative distribution is still a uniform distribution on the given interval; see, e.g., [4]. Why?

The fact that two different optimization problems lead to the exact same selection makes us think that there must be a fundamental reason behind these two results – and in this paper, we indeed describe such a reason.

Need to go beyond the maximum entropy approach: second argument. When we select a single distribution, we can find the *largest* possible entropy of the corresponding distributions. But what if, in addition to the largest possible value of the entropy, we want to know the whole *range* of values of entropy – i.e., we also want to know the *smallest* possible value?

This smallest possible value is not attained on a single probability distribution – in the discrete case, it is attained for all distributions for which $p_{i_0} = 1$ for some i_0 and for which $p_i = 0$ for all $i \neq i_0$. In this case, we cannot select a single distribution – the maximum entropy approach does not help, but we can still select a small representative *family* of distributions.

Again, a natural question is: can we generalize this result so that it would cover other practically useful situations?

2 Analysis of the Problem

What do entropy and sensitivity measure have in common? We would like to come up with a general result that generalizes both the maximum entropy and the sensitivity results. To come up with such a generalization, it is reasonable to analyze what these two 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, it 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 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 $p_{\pi(1)}, \ldots, p_{\pi(n)}$.

Permutations also work for the sensitivity example. Interestingly, a more complex criterion used in the sensitivity example is also permutation-invariant: as well as many other generalization of entropy.

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* $g(p_1, \ldots, p_n) = a$ *or* $g(p_1, \ldots, p_n) \geq a$ *for permutation-invariant functions* g*.*

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 both in the maximum entropy case and in the sensitivity case.

Proof. We will prove 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. *Let the number* n *of variable* p_i *be fixed.*

– *We say that a problem has a* small finite number of solutions *if its number of solutions is smaller than* n*.*
– *We say that a problem has a* d-parametric family of solutions *if it has* n^d *solutions.*

Proposition 2. *If a permutation-invariant optimization problem has a small finite number solutions with* $\sum p_i = 1$*, 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. Let us pick one such value p_k. Let S denote the set of all the indices j for which $p_j = p_k$, and let m denote the number of elements in this set. Since some values p_i are different, 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 $p_i \ge 0$ with $\sum p_i = 1$, then this family 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 entropy is the smallest possible.

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.

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.