How to Reconcile Maximum Entropy Approach with Intuition: e.g., Should Interval Uncertainty Be Represented by a Uniform Distribution

Vladik Kreinovich, Olga Kosheleva, and Songsak Sriboonchitta

Abstract In many practical situations, we only have partial information about the probabilities; this means that there are several different probability distributions which are consistent with our knowledge. In such cases, if we want to select one of these distributions, it makes sense not to pretend that we have a small amount of uncertainty – and thus, it makes sense to select a distribution with the largest possible value of uncertainty. A natural measure of uncertainty of a probability distribution is its entropy. So, this means that out of all probability distributions consistent with our knowledge, we select the one whose entropy is the largest. In many cases, this works well, but in some cases, this Maximum Entropy approach leads to counterintuitive results. For example, if all we know is that the variable is located on a given interval, then the Maximum Entropy approach selects the uniform distribution on this interval. In this distribution, the probability density $\rho(x)$ abruptly changes at the interval's endpoints, while intuitively, we expect that it should change smoothly with x. To reconcile the Maximum Entropy approach with our intuition, we propose to limit distributions to those for which the probability density's rate of change is bounded by some a priori value – and to limit the search for the distribution with the largest entropy only to such distributions. We show that this natural restriction indeed reconciles the Maximum Entropy approach with our intuition.

Vladik Kreinovich and Olga Kosheleva
University of Texas at El Paso, El Paso, Texas 79968, USA
e-mail: olgak@utep.edu, vladik@utep.edu

Songsak Sriboonchitta
Faculty of Economics, Chiang Mai University, Thailand
e-mail: songsakecon@gmail.com

1 Formulation of the Problem

Interval uncertainty is ubiquitous. Most of the information about the physical world comes from measurements. Measurements are never 100% accurate. Because of this, the measurement result $\widetilde{x}$ is, in general, different from the actual (unknown) value x of the desired quantity; see, e.g., [5].

In many practical situations, the only information that we have about the measurement error $\Delta x \stackrel{\text{def}}{=} \widetilde{x} - x$ is the upper bound Δ on its absolute value: $|\Delta x| \le \Delta$. In such situations, once we have performed the measurement, the only information that we have about the actual value x is that this value belongs to the interval $[\widetilde{x} - \Delta, \widetilde{x} + \Delta]$; we do not know the probabilities of different values from this interval. In principle, many different distributions are possible – namely, all distribution which are located on this interval with probability 1.

There exist many techniques for dealing with such uncertainty; they are known as *interval computations*; see, e.g., [1, 3, 4].

Sometimes, we need to select a single distribution. While there exist many techniques for dealing with interval uncertainty, interval computations is a reasonably new field, its founding papers appeared only in mid-1950s. In contrast, traditional probabilistic methods have been in existence for several centuries. Not surprisingly, there are much more techniques for processing probabilistic uncertainty – when we know the distribution of measurement errors – than for processing interval uncertainty.

In many data processing situations, interval methods are still being developed, while there exist well-tested efficient probabilistic techniques. In such situations, we have no choice but to apply these techniques. However, to apply them, we need to select a single probability distribution from all the distributions located on the given interval.

A similar problem appears in more general situations, when we have partial information about the probabilities – and thus, to apply the existing probabilistic techniques, we need to select one of the possible distributions. Which distribution should we select?

Laplace Indeterminacy Principle and Maximum Entropy Approach. In the interval case, we do not have any reason to believe that some values from the given interval are more probable than others. Thus, it is reasonable to assume that all the points from the interval are equally probable, i.e., that we have a uniform distribution on this interval, with the constant probability density function $\rho(x) = \text{const}$. This idea is known as *Laplace Indeterminacy Principle*. It is a particular case of a more general *Maximum Entropy Principle* (see, e.g., [2]), according to which, from all possible probability distributions, we should select the one with the largest possible value of the *entropy*

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

For interval uncertainty, Maximum Entropy approach selects a uniform distribution. Let us apply the Maximum Entropy approach to the situation of interval uncertainty, when all we know is that $\rho(x) = 0$ for all x outside the given interval $[\underline{x}, \overline{x}]$. In this situation, among all the functions $\rho(x)$ located on this interval and that satisfy the condition that

$$\int \rho(x)\, dx = 1, \tag{2}$$

we must select the one for which the entropy S is the largest possible.

By applying the Lagrange multiplier technique to the this constraint optimization problem of maximizing S, we get unconstrained optimization problem of maximizing the expression

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

for some parameter λ. Explicitly differentiating this expression with respect to $\rho(x)$ and equating the derivative to 0, we conclude that $-\ln(\rho(x)) - 1 + \lambda = 0$, hence $\ln(\rho(x)) = \text{const} = \exp(\lambda - 1)$ and therefore, $\rho(x)$ is also a constant. This constant can be found from the condition that $\int_{\underline{x}}^{\overline{x}} \rho(x)\, dx = 1$, i.e., that $(\overline{x} - \underline{x}) \cdot \rho(x) = 1$. Thus,

$$\rho(x) = \frac{1}{\overline{x} - \underline{x}}.$$

As a result, we get a uniform distribution on the interval $[\underline{x}, \overline{x}]$.

Why this conclusion is counter-intuitive. Intuitively, we expect that when the two events are close, their probabilities should also be close. In particular, we expect that when the values x and x' are close to each other, then the corresponding values $\rho(x)$ and $\rho(x')$ should also be close to each other.

Since we know that the probability distribution is located on the interval $[\underline{x}, \overline{x}]$ and thus, $\rho(\underline{x} - \varepsilon) = 0$ for all $\varepsilon > 0$, we thus expect to have $\rho(\underline{x}) = 0$, and similarly $\rho(\overline{x}) = 0$ – and we expect the probability density function to be continuously rising from 0 to some value and then decreasing again to 0 as we reach the right endpoint $\overline{x}$ of the given interval.

In contrast to this natural intuition, for the uniform distribution – coming from using the Maximum Entropy approach – the value of the probability density function:

- jumps abruptly from 0 to $\dfrac{1}{\overline{x} - \underline{x}}$ as we cross into the interval,
- remains the same throughout this interval, and then
- abruptly drops back to 0 as we leave this interval.

This is not the only case when Maximum Entropy approach leads to counterintuitive results. Similar counterintuitive results happen in more complex situations as well.

For example, suppose that, in addition to the bounds $\underline{x} \le x \le \overline{x}$ on the random variable x, we also know its first moment

$$E = \int x \cdot \rho(x)\,dx. \tag{3}$$

Then, the maximum entropy approach means that we maximize entropy S under constraints (2) and (3). For this constraint optimization problem, the Lagrange multiplier methods leads to the unconstrained problem of maximizing the expression

$$-\int \rho(x)\cdot \ln(\rho(x))\,dx + \lambda \cdot \left(\int \rho(x)\,dx - 1\right) + \lambda_1 \cdot \left(\int x \cdot \rho(x)\,dx - E\right)$$

for some values λ and λ_1. Differentiating this expression with respect to $\rho(x)$ and equating the derivative to 0, we conclude that $-\ln(\rho(x)) - 1 + \lambda + \lambda_1 \cdot x = 0$. So, $\ln(\rho(x)) = a + \lambda_1 \cdot x$, where we denoted $a \overset{\text{def}}{=} \lambda - 1$ and thus, $\rho(x) = \exp(a + \lambda_1 \cdot x)$. This expression is positive on both endpoints $x = \underline{x}$ and $x = \overline{x}$ and thus, on each of the endpoints, there is a discontinuous jump to 0.

Similarly, if, in addition to the first moment E, we also know the second moment

$$M_2 = \int x^2 \cdot \rho(x)\,dx, \tag{4}$$

the Lagrange multiplier method leads to the unconstrained problem of maximizing the expression

$$-\int \rho(x)\cdot \ln(\rho(x))\,dx + \lambda \cdot \left(\int \rho(x)\,dx - 1\right) + \lambda_1 \cdot \left(\int x \cdot \rho(x)\,dx - E\right) +$$
$$\lambda_2 \cdot \left(\int x^2 \cdot \rho(x)\,dx - M_2\right)$$

for which we get $\rho(x) = \exp(a + \lambda_1 \cdot x + \lambda_2 \cdot x^2)$, i.e., a truncated normal distribution – which also has jumps from 0 to positive values on each of the two endpoints $\underline{x}$ and $\overline{x}$ of the original interval.

How can we reconcile maximum entropy approach with intuition? It is desirable to modify the Maximum Entropy approach so that it will be reconciled with our intuition. In this paper, we propose a natural way to do it.

2 Reconciling Maximum Entropy and Intuition: Idea and Consequences

How to formalize our intuition. Let us first try to describe our intuition in precise terms. Intuitively, we do not expect the values of the probability density functions to jump. Moreover, we do not expect them to change too abruptly – this would be similar to jumping. Thus, our intuition means that there is an upper bound B of the rate with which the probability density function can change, i.e., that:

$$|\rho'(x)| \le B \text{ for all } x. \tag{5}$$

This formalization naturally leads us to the following idea.

How to reconcile the Maximum Entropy approach and our intuition: main idea. In view of the above discussion, a reasonable idea is to add the inequality (5) as an additional constraint when maximizing the entropy. In other words, we select the distribution with the largest possible value of the entropy among all distributions which are consistent with our knowledge *and* which satisfy the additional constraint (5).

How to actually implement this idea. In general, if we do not have any inequality constraints, if we simply want to maximize an objective function $F(y)$, then, from the fact that at the optimizing point y_{opt}, small changes of y do not increase the value of the objective functions, we conclude that all the partial derivatives of the objective function should be equal to 0 at this point. Indeed:

- if one of partial derivatives $\dfrac{\partial F}{\partial y_i}(y_{\text{opt}})$ was positive, then a small increase in the component y_i will increase the value of the objective function beyond the largest possible value $F(y_{\text{opt}})$, and
- if one of partial derivatives $\dfrac{\partial F}{\partial y_i}(y_{\text{opt}})$ was negative, then a small decrease in the component y_i will increase the value of the objective function beyond the largest possible value $F(y_{\text{opt}})$.

The only remaining option is that all the derivatives are zeros – this is exactly the usual calculus-based criterion that we used in the previous section to find the corresponding maxima.

When maximize a function $F(y)$ under an inequality constraint $G(y) \ge 0$, then for the optimizing point y_{opt}, we either have $G(y_{\text{opt}}) > 0$ or $G(y_{\text{opt}}) = 0$. In the first case, when we make small changes to y, the condition $G(y) > 0$ will still be satisfied. Thus, all small changes are allowed – and since these small changes cannot lead to an increase in the value of the objective function, we make the conclusion that all the partial derivatives will be 0.

In our examples, this means that for each x:

- we either have a strict inequality $|\rho'(x)| < B$, in which case all derivatives are 0s and we can derive the same formulas as before,
- or we have the equality $|\rho'(x)| = B$.

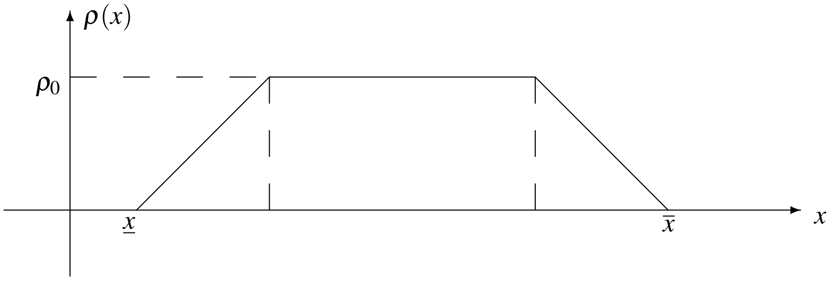
Let us describe what this means for each of the above three problems:

- when we only know that the random variable is located in an interval $[\underline{x}, \overline{x}]$,
- when we also know the first moment E, and
- when, in addition to the interval, we also know the first moment E and the second moment M_2.

Case when we only know that the random variable is located on the interval $[\underline{x}, \overline{x}]$. In this case, the above argument leads to the following optimal function $\rho(x)$:

- near the left endpoint $\underline{x}$, where $\rho(\underline{x}-\varepsilon)=0$, we cannot have a constant function $\rho(x)$ (which comes from equating the partial derivative with respect to $\rho(x)$ to 0), and we cannot have a function which is decreasing ($\rho'(x)=-B$); so, the only remaining choice is an increasing function with $\rho'(x)=B$ and thus,
$$\rho(x)=B\cdot(x-\underline{x});$$
- after the increasing function reaches a certain height, we get
$$\rho(x)=\rho_0$$
for some constant ρ_0;
- after that, when we are close to the right endpoint $\overline{x}$, the only remaining option is $\rho'(x)=-B$, for which
$$\rho(x)=B\cdot(\overline{x}-x).$$

So, in this case, we have a trapezoid probability density function, as shown in the following picture.



The value ρ can be determined from the condition (1), i.e., from the condition that the area under this graph be equal to 1. With the rate of rising equal to B, the rise from 0 to ρ_0 requires an x-length $\frac{\rho_0}{B}$. The joint areas of the ρ-increasing and the ρ-decreasing triangles is thus equal to $\rho_0\cdot\frac{\rho_0}{B}$. At the remaining part $L-2\cdot\frac{\rho_0}{B}$ of the interval, where we denoted $L\stackrel{\text{def}}{=}\overline{x}-\underline{x}$, the pdf $\rho(x)$ is equal to ρ_0, so its integral on this part of the original interval is equal to $\rho_0\cdot\left(L-2\cdot\frac{\rho_0}{B}\right)$. Thus, the overall integral of $\rho(x)$ over the whole interval $[\underline{x},\overline{x}]$ is equal to
$$\rho_0\cdot\frac{\rho_0}{B}-\rho_0\cdot\left(L-2\cdot\frac{\rho_0}{B}\right)=1.$$

If we open the parentheses, combine similar terms, multiply all the terms by B, and move all the terms to the right-hand side, we get the following quadratic equation for ρ_0:
$$\rho^2-L\cdot B\cdot\rho_0+B=0,$$
from which we can determine ρ_0 as

$$\rho_0 = \frac{L\cdot B - \sqrt{L^2\cdot B^2 - 4B}}{2}. \quad (6)$$

Comment. The usual uniform distribution corresponds to the case when there is no limitation on the rate of change, i.e., when $B \to \infty$. Let us show that in this limit, the expression (6) indeed leads to the usual formula $\rho_0 = \frac{1}{L}$. Indeed, we have

$$L^2\cdot B^2 - 4B = L^2\cdot B^2\cdot\left(1-\frac{4B}{L^2\cdot B^2}\right) = L^2\cdot B^2\cdot\left(1-\frac{4}{L^2\cdot B}\right),$$

thus

$$\sqrt{L^2\cdot B^2 - 4B} = \sqrt{L^2\cdot B^2\cdot\left(1-\frac{4}{L^2\cdot B}\right)} = L\cdot B\cdot\sqrt{1-\frac{4}{L^2\cdot B}}.$$

Since the ratio $z \stackrel{\text{def}}{=} \frac{4}{L^2\cdot B}$ is, for large B, close to 0, we can use the fact that

$$\sqrt{1-z} = 1-\frac{z}{2}+O(z^2).$$

Then, we get

$$\sqrt{L^2\cdot B^2 - 4B} = L\cdot B\cdot\left(1-\frac{2}{L^2\cdot B}+O\left(\frac{1}{B^2}\right)\right) = L\cdot B - \frac{2}{L} + O\left(\frac{1}{B}\right).$$

Thus,

$$\rho_0 = \frac{L\cdot B - \sqrt{L^2\cdot B^2 - 4B}}{2} = \frac{L\cdot B - \left(L\cdot B - \frac{2}{L} + O\left(\frac{1}{B}\right)\right)}{2} = \frac{1}{L} + O\left(\frac{1}{B}\right).$$

So, for large B, we indeed get $\rho_0 \approx \frac{1}{L}$.

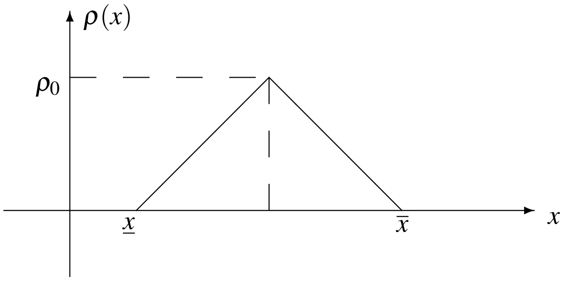
Discussion. In the above analysis, we assumed that we know the value B corresponding to our intuition. What if we do not know this value? In this case, it makes sense to select this bound B to be as small as possible – to guarantee the smallest possible rate of change of the corresponding probability density function. As we decrease B, the periods of increase and decrease become longer – until we reach the point when these two periods fill the whole interval and further decrease in B is not possible. For the resulting smallest possible value B, the probability density function becomes triangular.

Since the increase of $\rho(x)$ from 0 to the largest value ρ_0 is happening at the same rate as the decrease back to 0, the maximum value ρ_0 of $\rho(x)$ is attained at the midpoint $\widetilde{x} = \frac{\underline{x}+\overline{x}}{2}$ of the original intervals. So, the resulting rate of increase B can be obtained by dividing the maximum value ρ_0 by the half-width $\Delta = \dfrac{\overline{x}-\underline{x}}{2} = \dfrac{L}{2}$ of the original interval: $B = \dfrac{\rho_0}{\Delta}$. In this case:

- for $x \le \widetilde{x}$, we have $\rho(x) = B \cdot (x - \underline{x})$, and
- for $x \ge \widetilde{x}$, we have $\rho(x) = B \cdot (\overline{x} - x)$.

The area under this triangular curve is equal to $\dfrac{1}{2} \cdot L \cdot \rho_0 = 1$, thus $\rho_0 = 2L$ and hence,

$$B = \frac{\rho_0}{L/2} = \frac{2/L}{L/2} = \frac{4}{L^2}.$$



What if we also know the first moment. In this case, similar to the previous case:

- we first have a linear increase $\rho(x) = B \cdot (x - \underline{x})$ until some value x_-;
- then we have an exponential distribution $\rho(x) = \exp(a + \lambda_1 \cdot x)$ until we reach some other value x_+;
- after that, we have a linear decrease $\rho(x) = B \cdot (\overline{x} - x)$.

Once we know a and λ_1, we can determine the transition values x_- and x_+ from the condition that the probability density function be continuous, i.e., from the conditions that $B \cdot (x - \underline{x}_-) = \exp(a + \lambda_1 \cdot x_-)$ and that $\exp(a + \lambda_1 \cdot x_+) = B \cdot (\overline{x} - x_+)$. The values a and λ_1 must be determined from the conditions (2) and (3).

What if we also know the first and the second moment. In this case, similar to the previous two cases:

- we first have a linear increase $\rho(x) = B \cdot (x - \underline{x})$ until some value x_-;
- then we have a (truncated) Gaussian distribution $\rho(x) = \exp(a + \lambda_1 \cdot x + \lambda_2 \cdot x^2)$ until we reach some other value x_+;
- after that, we have a linear decrease $\rho(x) = B \cdot (\overline{x} - x)$.

Once we know a, λ_1, and λ_2, we can determine the transition values x_- and x_+ from the condition that the probability density function be continuous, i.e., from the conditions that

$$B \cdot (x - \underline{x}_-) = \exp(a + \lambda_1 \cdot x_- + \lambda_2 \cdot x_-^2)$$

and that

$$\exp(a+\lambda_1 \cdot x_+ + \lambda_2 \cdot x_+^2) = B \cdot (\overline{x} - x_+).$$

The values a, λ_1, and λ_2 must be determined from the conditions (2), (3), and (4).

Multi-D case. In the multi-D case, if all we know is that each variable $x_1,\ldots,x_n$ is located on the corresponding interval $[\underline{x}_i,\overline{x}_i]$, then the Maximum Entropy approach leads to a uniform distribution on the corresponding box

$$\mathscr{B} = [\underline{x}_1,\overline{x}_1] \times \ldots \times [\underline{x}_n,\overline{x}_n],$$

a distribution with the constant probability density

$$\rho(x) = \rho_0 = \frac{1}{(\overline{x}_1 - \underline{x}_1) \cdot \ldots \cdot (\overline{x}_n - \underline{x}_n)}.$$

This conclusion is also counterintuitive since on the border of this box, the probability density function changes abruptly from 0 to ρ_0, while intuitively, it should be continuous – and moreover, it should not change too fast.

In this case, it is reasonable to require a similar limitation on the rate of change. In the 1-D case, the limitation $|\rho'(x)| \le B$ means that if, for some $\varepsilon > 0$, the two values x and y are ε-close (in the sense that $|x-y| \le \varepsilon$), then the corresponding values of the probability density should not differ by more than $B \cdot \varepsilon$, i.e., that we should have $|\rho(x)-\rho(y)| \le B \cdot \varepsilon$. Similarly, in the multi-D case, it is reasonable to require that if, for some $\varepsilon > 0$, the points $x = (x_1,\ldots,x_n)$ and $y = (y_1,\ldots,y_n)$ are ε-close – in the sense that all their components are ε-close, i.e., that $|x_i - y_i| \le \varepsilon$ for all i – then we should have $|\rho(x)-\rho(y)| \le B \cdot \varepsilon$.

The ε-closeness of two points can be equivalently described as $d_{\max}(x,y) \le \varepsilon$, where we denoted

$$d_{\max}((x_1,\ldots,x_n),(y_1,\ldots,y_n)) \stackrel{\text{def}}{=} \max(|x_1-y_1|,\ldots,|x_n-y_n|).$$

In these terms, the above requirement takes the form

$$|\rho(x)-\rho(y)| \le B \cdot d_{\max}(x,y). \tag{7}$$

Similarly to the 1-D case, we propose, when looking for a distribution with the largest entropy, to limit ourselves only to probability distributions that satisfy this condition (7) for all x and y.

Under this restriction, in the situation when all we know is that the distribution is located in a box, then we only have $\rho(x)$ equal to some constant ρ_0 for all the points $x = (x_1,\ldots,x)$ whose max-distance $d_{\max}(x,\partial\mathscr{B}) \stackrel{\text{def}}{=} \min\limits_{y\in\partial\mathscr{B}} d_{\max}(x,y)$ to the box's border $\partial\mathscr{B}$ does not exceed the ratio $\dfrac{\rho_0}{B}$. For points x which are closer to the border, we have

$$\rho(x) = B \cdot d_{\max}(x,\partial\mathscr{B}).$$

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

References

1. 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.
2. E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.
3. G. Mayer, *Interval Analysis and Automatic Result Verification*, de Gruyter, Berlin, 2017.
4. R. E. Moore, R. B. Kearfott, and M. J. Cloud, *Introduction to Interval Analysis*, SIAM, Philadelphia, 2009.
5. S. G. Rabinovich, *Measurement Errors and Uncertainties: Theory and Practice*, Springer, New York, 2005.