How to Get Beyond Uniform When Applying MaxEnt to Interval Uncertainty

Songsak Sriboonchitta[†] and Vladik Kreinovich[‡,1]

[†]Faculty of Economics, Chiang Mai University
Chiang Mai 50200, Thailand
e-mail : `songsakecon@gmail.com`

[‡]Department of Computer Science, University of Texas at El Paso
El Paso, TX 79968, USA
e-mail : `vladik@utep.edu`

Abstract : In many practical situations, the Maximum Entropy (MaxEnt) approach leads to reasonable distributions. However, in an important case when all we know is that the value of a random variable is somewhere within the interval, this approach leads to a uniform distribution on this interval – while our intuition says that we should have a distribution whose probability density tends to 0 when we approach the interval's endpoints. In this paper, we show that in most cases of interval uncertainty, we have additional information, and if we account for this additional information when applying MaxEnt, we get distributions which are in perfect accordance with our intuition.

Keywords : maximum entropy; interval uncertainty.
2010 Mathematics Subject Classification : 94A17; 62B10; 65G99.

1   Formulation of the Problem

General problem: we often have partial information about the probabilities. Most practically used statistical methods assume that we know the probability distribution – or at least that we know a finite-dimensional class of distributions that contains the actual (unknown) distribution; see, e.g., [1]. For example, we may know that the distribution is normal or that the distribution is uniform, but we do not necessarily know the parameters describing this particular distribution.

In practice, however, we often encounter situations when we only have partial information about the probabilities – and not enough information to specify a finite-parametric family containing the distribution. For example, we may know only the first two moments of the distribution, and/or we may know only that the distribution is located somewhere on an interval $[\underline{x}, \overline{x}]$, but we do not have any additional information about the distribution.

A natural idea: select the most reasonable distribution. In such situations, when we have many possible probability distributions which are consistent with our knowledge, a reasonable idea is:

- to select the most "reasonable" of these distributions, and
- to use statistical methods corresponding to this selected distribution.

This natural idea leads to the maximum entropy approach. How can we select this "most reasonable" distribution? A reasonable requirement for such a selection is to make sure that this distribution properly represents the original class of possible distributions as much as possible.

A natural way to formalize this requirement is to require that narrowing down from the class of distributions to a single distribution minimally changes our uncertainty.

For example, if we started with the situation in which we know only that the random variable has two values v_1 and v_2, but we do not know the probabilities p_1 and p_2 of these values, it would be not right to select a distribution in which v_1 appears with probability 1 – or even with probability 0.99, since this would drastically decrease our original uncertainty.

How can we gauge uncertainty? In situations when we know only that our alternative is one of n possible alternatives, a reasonable measure of uncertainty is the smallest possible number of binary ("yes"-"no") questions needed to determine the actual alternative; this number is equal to $\lceil \log_2(n) \rceil$ [2, 3].

In situations in which we know the probabilities $p_1, \ldots, p_n$ of different uncertainties, it is reasonable to minimize the *average* number of binary questions needed to select the actual alternative. One can show (see, e.g., [3]) that this average number of binary question is equal to the entropy $S = -\sum_{i=1}^{n} p_i \cdot \log_2(p_i)$.

For a continuous random variable with probability density $\rho(x)$, we can, for each ε, estimate the mean number $N(\varepsilon)$ of binary questions which are needed to

determine the actual value with accuracy ε. It turns out that for small $\varepsilon > 0$, this number of questions has the form $S - \log_2(2\varepsilon)$, where $S(\rho) = -\int \rho(x) \cdot \log_2(\rho(x))\, dx$ is the entropy of the corresponding distribution [3].

When we only know that the distribution belongs to a certain class $\mathcal{C}$, then for each way of asking questions, the average number of question may depend on the actual distribution. It is therefore reasonable to look for the worst-case average number of questions. It turns out that the smallest value of this worst-case average number of questions is asymptotically equal to $\overline{S}(\mathcal{C}) - \log_2(2\varepsilon)$, where $\overline{S}(\mathcal{C}) \overset{\text{def}}{=} \max_{\rho \in \mathcal{C}} S(\rho)$ is the largest entropy among all distributions from the class $\mathcal{C}$; see, e.g., [2, 3].

From this viewpoint, selecting a distribution that maximally represents uncertainty means selecting a distribution $S(\rho)$ is maximally close to $\overline{S}(\mathcal{C}) = \max_{\rho} S(\rho)$. In other words, out of all possible distributions, we need to select the distribution with the maximum possible value of the entropy. This is the main idea behind the Maximum Entropy (MaxEnt) approach to uncertainty [2].

MaxExt approach: successes. The MaxEnt approach has many successful applications; see, e.g., [2]. Suffice is to say that for the case when we know only the first two moments of the probability distribution, the MaxEnt approach leads to Gaussian (normal) distribution – the distribution which is indeed ubiquitous in practice.

Similarly, in situations when we only know the marginal distributions $\rho_1(x_1)$ and $\rho_2(x_2)$, but we have no information about the relation between different random variables, the MaxEnt approach selects the joint distribution $\rho(x_1, x_2) = \rho_1(x_1) \cdot \rho_2(x_2)$ according to which these variables are independent – which also makes perfect sense.

MaxEnt approach: challenges. While in many situations, the MaxEnt approach indeed select a reasonable distribution, a distribution agreeing with our intuition, in many other situations, the MaxEnt selection is not in full agreement with our intuition.

Let us give a very simple example of such a situation. Let us assume that all we know is that a quantity x is always located within the known bounds $\underline{x}$ and $\overline{x}$: $\underline{x} \leq x \leq \overline{x}$.

This is a very widely spread situations. Indeed, the information about the quantity x usually comes from measurements. Measurement are never absolutely accurate, the measurement result $\widetilde{x}$ is, in general, somewhat different from the actual (unknown) value x, and thus, the corresponding *measurement error* $\Delta x \overset{\text{def}}{=} \widetilde{x} - x$ is, in general, different from 0. In many practical situations, the only information that we have about the measurement error is the upper bound Δ on its absolute value: $|\Delta x| \leq \Delta$; see, e.g., [4]. This upper bound is usually provided by the manufacturer of the measuring instrument. In such situations, after we get the measurement result $\widetilde{x}$, the only information that we have about the actual value x is that x is between $\underline{x} = \widetilde{x} - \Delta$ and $\overline{x} = \widetilde{x} + \Delta$.

From the common sense viewpoint, what would be the reasonable probability distribution on the interval $[\underline{x}, \overline{x}]$? Intuitively, since we know that the values below $\underline{x}$ are strictly prohibited, we expect the values equal to $\underline{x}$ and slightly larger than $\underline{x}$ to have a low probability, with probability increasing as we increase x – until we get closer to the upper bound, in which we also expect the probability to decrease. In other words, we expect a continuous probability density function $\rho(x)$ that it equal to 0 for $x = \underline{x}$ and for $x = \overline{x}$ and that increases as x increases from $x = \underline{x}$ and decreases as we approach $x = \overline{x}$.

However, when we apply the MaxEnt approach to this situation, we get a uniform distribution, with a discontinuous probability density $\rho(x)$ which is equal to a constant $\rho(x) = \dfrac{1}{\overline{x} - \underline{x}}$ on the interval $[\underline{x}, \overline{x}]$ and is equal to 0 outside this interval.

Resulting problem. On the one hand, the MaxEnt approach sounds reasonable, in good accordance with our intuition on how to select a probability distribution. On the other hand, when we apply this seemingly reasonable approach to a simple case of interval uncertainty, we get counter-intuitive results. How can we reconcile these two intuitions? What do we need to modify to make sure that for the case of interval uncertainty, the MaxEnt approach leads to a more intuitive result?

In this paper, we show that such a reconciliation is indeed possible, and that a slightly deeper analysis of the corresponding practical problems leads to an intuitively acceptable results of applying MaxEnt to interval uncertainty.

2 Let Us Recall How Other Intuition-Friendly Uncertainty Techniques Treat Interval Uncertainty

Let us look at other techniques. Since for interval uncertainty, the result of MaxEnt-based probabilistic analysis is not in perfect accordance with our intuition, it may be a good idea to see how other intuition-motivated approaches to uncertainty deal with the interval-uncertainty case.

Fuzzy technique – a technique specifically designed to describe human intuition. A natural idea is look at fuzzy techniques, techniques specifically designed to deal with human statements – especially statements that use imprecise ("fuzzy") words from natural language, such as "small", "approximately 5", etc.; see, e.g., [5, 6, 7]. In fuzzy logic, to translate the expert's imprecise statement into precise terms, for every real number x, we describe the expert's degree of confidence $\mu(x)$ that this particular value x is in agreement with the expert's statement (e.g., the degree of confidence that x is small).

This degree of confidence $\mu(x)$ can be determined, e.g., by asking the expert to mark his or her degree by a mark on a scale, e.g., form 0 to 10. If the expert selects 7 on a scale from 0 to 10, we take $\mu(x) = 7/10$.

The resulting function function $\mu(x)$ is known as a *membership function*. For two close values x and x', the corresponding degrees $\mu(x)$ and $\mu(x')$ should be

close; thus, usually, we consider continuous membership functions.

How fuzzy techniques treat interval uncertainty. In this formalism, an expert's interval-related statement – that x is always in between $\underline{x}$ and $\overline{x}$ – is usually interpreted by a continuous function $\mu(x)$ which is equal to 0 on both ends of the corresponding interval.

A typical choice is a *triangular* membership function $\mu(x)$ which:

- linearly increases up to a midpoint of the corresponding interval, and then

- linearly decreases from there.

Another widely used choice is a *trapezoidal* membership function, which:

- linearly increases from 0 to 1,

- then stays at 1, and

- then linearly decreases back to 0.

Fuzzy techniques only keep one of the two intuitions, but we want to keep both. We started this paper by saying that there are two reasonable intuitions:

- an intuition on how to select a distribution, and

- an intuition on what to do in the case of interval uncertainty,

and that the problem is that these two intuitions seem to conflict with each other.

One way to deal with this situation is to ignore one of the intuitions. This is exactly what happens now:

- The current MaxEnt approach keeps the intuition about selecting a distribution, but ignores the intuition about the interval uncertainty.

- The fuzzy approach keeps the intuition about interval uncertainty, but ignores the intuition about selecting a distribution.

In this paper, we show that we can do better: we can keep both intuitions practically intact.

3 How to Reconcile the Two Intuitions: General Idea

The problem arises when we apply MaxEnt to a situation in which the only information that we have about the desired quantity x is that this quantity is between the two known bounds $\underline{x}$ and $\overline{x}$. For example, in case of measurement uncertainty, we have bounds $\underline{x} = \widetilde{x} - \Delta$ and $\overline{x} = \widetilde{x} + \Delta$.

But is this indeed the only information we have? Let us consider the case of measurement uncertainty. We get the bound Δ from the manufacturer of the measuring instrument, but where does the manufacturer gets this value? There are two possible way to determine such a bound (see, e.g., [4]):

- In some cases, the manufacturer can empirically calibrate the measuring instrument. Specifically, in some cases, there exists a much more accurate ("standard") measuring instrument, and so we can estimate Δ experimentally, by comparing the results of using this instrument with the results of using the standard measuring instrument.

- In other cases, e.g., when the corresponding instrument is itself a state-of-the-art one, we do not have any much more accurate measuring instruments that we can use for calibration. In such cases, we need to perform some theoretical analysis to come up with an appropriate bound Δ.

Let us consider these two situations one by one.

4 Cases When the Upper Bound Δ Comes from Calibration

Analysis of the problem. The fact that the manufacturer provides us with an upper bound Δ definitely means that values of Δx which are larger that this bound are not possible (or at least have a very very small probability) – otherwise, the misleading upper bound may lead to a disaster, and the manufacturer of the measuring instrument will be sued into bankruptcy.

However, by the same logic, it does not mean that all the values below Δ are possible. Indeed, how can the manufacturer know this bound?

The manufacturer performs some tests, in which we compare the measurement result with the result of measuring the same quantity by a much more accurate ("standard") measuring instrument – whose measurement results can be therefore taken as actual values of the corresponding quantities. Bases on the results of each test, we can determine the absolute value of the measurement error

$$|\Delta x_i| = |\widetilde{x}_i - x_i|,$$

and we can then calculate the largest of these absolute values $\delta \stackrel{\text{def}}{=} \max_i |\Delta x_i|$. It is well known that, based on finitely many tests, we cannot determine the exact upper bound, so we need to use some statistical methods, e.g.:

- get a confidence interval for $\max |\Delta x|$ corresponding to the allowed very small probability of exceeding this bound, and

- use the upper endpoint of this confidence interval as Δ.

We have an additional information. In this case, we have some additional information about uncertainty. Namely:

- we know that there is the *actual* (unknown) bound B on the absolute value of the measurement error, the bound which can take any value from δ to Δ, and

- we know that the actual value x is somewhere between $\widetilde{x} - B$ and $\widetilde{x} + B$.

How does this help? From the purely mathematical viewpoint, it is the same statement. At first glance, one may ask: how can this additional information help? We use more words than before to describe what we know, but, from the purely mathematical viewpoint, the result is the same:

- values x inside the interval $[\widetilde{x} - \Delta, \widetilde{x} + \Delta]$ are possible, while

- values x outside the interval are not possible.

Indeed, it is possible that $B = \Delta$, in which case all values from the interval $[\widetilde{x} - B, \widetilde{x} + B] = [\widetilde{x} - \Delta, \widetilde{x} + \Delta]$ are possible. On the other hand, we know that $B \leq \Delta$, so from the fact that $|\Delta x| \leq B$ we conclude that $|\Delta x| = |\widetilde{x} - x| \leq \Delta$ and thus, that any value outside the interval $[\widetilde{x} - \Delta, \widetilde{x} + \Delta]$ are not possible.

In spite of this mathematical equivalence, we will show that from the MaxEnt viewpoint, this idea does help.

From the MaxEnt viewpoint, the above more detailed description of uncertainty leads to a different – and more intuitive – result. In the detailed description, we have *two* unknown: B and x (or, equivalently, B and Δx).

For B, the only information that we have is that B is between δ and Δ. So, if we apply MaxEnt to this information to select a probability distribution for the unknown B, we conclude that B is uniformly distributed on the interval $[\delta, \Delta]$, with probability density $\rho_b(B) = \dfrac{1}{\Delta - \delta}$ for $B \in [\delta, \Delta]$ and $\rho_b(B) = 0$ for $B \notin [\delta, \Delta]$.

For each possible bound $B \in (\delta, \Delta)$, we know that Δx is located in the interval $[-B, B]$. If we apply the MaxEnt to the information to select a probability for the unknown Δx, we conclude that that for each B, the unknown Δx is uniformly distributed on the interval $[-B, B]$, with the probability density $\rho(\Delta x \,|\, B) = \dfrac{1}{2B}$ for $\Delta x \in [-B, B]$ and $\rho(\Delta x \,|\, B) = 0$ for $\Delta x \notin [-B, B]$.

To get the final distribution for Δx, we need to combine the probabilities corresponding to different values B. Thus, we end up with the following pdf:

$$\rho(\Delta x) = \int_{\delta}^{\Delta} \rho_b(B) \cdot \rho(\Delta x \,|\, B) \, dB = \int_{\delta}^{\Delta} \frac{1}{\Delta - \delta} \cdot \rho(\Delta x \,|\, B) \, dB.$$

For each Δx, the value $\rho(\Delta x \,|\, B)$ is only different from 0 when $|\Delta x| \leq B$.

When $|\Delta x| \leq \delta$, the inequality $|\Delta x| \leq B$ holds for all possible values $B \in [\delta, \Delta]$, thus we get

$$\rho(\Delta x) = \int_{\delta}^{\Delta} \frac{1}{\Delta - \delta} \cdot \frac{1}{2B} \, dB = \frac{1}{2(\Delta - \delta)} \cdot (\ln(\Delta) - \ln(\delta)).$$

On the other hand, when $|\Delta x|$ is between δ and Δ, the inequality $|\Delta x| \leq B$ is only satisfies from values $B \geq |\Delta x|$; thus, only for these values, $\rho(\Delta x \,|\, B)$ are different from 0, and we get:

$$\rho(\Delta x) = \int_{|\Delta x|}^{\Delta} \frac{1}{\Delta - \delta} \cdot \frac{1}{2B} \, dB = \frac{1}{2(\Delta - \delta)} \cdot (\ln(\Delta) - \ln(|\Delta x|)).$$

Thus, we arrive at the following pdf:

- when $|\Delta x| \leq \delta$, we get $\rho(\Delta x) = \dfrac{1}{2(\Delta - \delta)} \cdot (\ln(\Delta) - \ln(\delta))$, i.e., the distribution is uniform here; however,

- when $\delta \leq |\Delta x| \leq \Delta$, we get $\rho(\Delta x) = \dfrac{1}{2(\Delta - \delta)} \cdot (\ln(\Delta) - \ln(|\Delta x|)$.

One can easily check that when $|\Delta x|$ tends to Δ, the value of this pdf $\rho(x)$ tends to 0. Thus, this MaxEnt-related function is indeed in good accordance with our intuition.

5   Cases The Upper Bound Δ Is Estimated Theoretically

Analysis of the problem. Let us now consider the cases when the upper bound Δ is determined theoretically.

How is it determined? Usually, there are several different sources of the measurement error. As a result, the overall measurement error Δx is a sum of several components corresponding to these different sources:

$$\Delta x = \Delta_1 x + \ldots + \Delta_m x.$$

So, specialists in different types of uncertainty estimate the bounds Δ_i on the (absolute value of the) corresponding component $\Delta_i x$ of measurement error, and then we add these bounds to get a bound $\Delta = \Delta_1 + \ldots + \Delta_m$ for the overall measurement error.

We have an additional information. In this case, we have some additional information about uncertainty. Namely:

- we know that there are actual value $\Delta_i x$ corresponding to different components of measurement error; each of these components can take any value from $-\Delta_i$ to Δ_i; and

- we know that the actual value Δx is equal to the sum of all these components: $\Delta x = \Delta_1 x + \ldots + \Delta_m x$.

How does this help? From the purely mathematical viewpoint, it is the same statement. Similarly to the previous section, at first glance, one may ask: how can this additional information help? We use more words than before to describe what we know, but, from the purely mathematical viewpoint, the result is the same:

- values Δx inside the interval $[-\Delta, \Delta]$ are possible, while

- values Δx outside this interval are not possible.

Indeed, since $|\Delta_i x| \le \Delta_i$, we can conclude that

$$|\Delta x| = |\Delta_1 x + \ldots + \Delta_m x| \le |\Delta_1 x| + \ldots + |\Delta_m x| \le \Delta_1 + \ldots + \Delta_m = \Delta.$$

Vice versa, every value Δx for which $|\Delta x| \le \Delta$ can be described as the sum $\Delta x = \Delta_1 x + \ldots + \Delta_m x$ of value $\Delta_i x$ for which $|\Delta_i x| \le \Delta_i$: indeed, we can take $\Delta_i x = \dfrac{\Delta x}{\Delta} \cdot \Delta_i$.

In spite of this mathematical equivalence, we will show that from the MaxEnt viewpoint, this idea does help.

From the MaxEnt viewpoint, the above more detailed description of uncertainty leads to a different – and more intuitive – result. In the detailed description, m unknown: $\Delta_1 x$, ..., $\Delta_m x$. For each of these components $\Delta_i x$, we know that it can take any value between $-\Delta_i$ and Δ_i. So, if we apply MaxEnt to this information to select a probability distribution for the unknown B, we conclude that $\Delta_i x$ is uniformly distributed on the interval $[-\Delta_i, \Delta_i]$, with probability density $\rho_i(\Delta x_i) = \dfrac{1}{2\Delta_i}$ for $\Delta_i x \in [-\Delta_i, \Delta_i]$ and $\rho_i(\Delta_i x) = 0$ for $\Delta_i x \notin [-\Delta_i, \Delta_i]$.

We have no information about the correlation between different components $\Delta_i x$. Thus, if we apply MaxEnt, we conclude that all these components are independent. So, Δx is the sum of m independent unform distributions.

Such distributions are well known, and for them, $\rho(\Delta x)$ tends to 0 as Δx approaches Δ or $-\Delta$. In particular, for $m = 2$, we have

$$\rho(\Delta x) = \int \rho_1(\Delta_1 x) \cdot \rho_2(\Delta x - \Delta_2 x) \, d(\Delta_1 x).$$

The product is different from 0 – and equal to $\dfrac{1}{4\Delta_1 \cdot \Delta_2}$ – when both pdfs are different from 0, i.e., when $-\Delta_1 \le \Delta_1 x \le \Delta_1$ and $-\Delta_2 \le \Delta x - \Delta_1 x \le \Delta_2$. By combining these inequalities, we conclude that the product is different from 0 when

$$\max(-\Delta_1, \Delta x - \Delta_2) \le \Delta_1 x \le \min(\Delta_1, \Delta x + \Delta_2).$$

Thus, the value $\rho(\Delta x)$ is equal to

$$\rho(x) = \frac{\min(\Delta_1, \Delta x + \Delta_2) - \max(-\Delta_1, \Delta x - \Delta_2)}{4\Delta_1 \cdot \Delta_2}.$$

For $\Delta x = \Delta$, the numerator turns into $\Delta_1 - (\Delta - \Delta_2) = 0$. So, this pdf is continuous on the whole real line – in good accordance with our intuition.

What is the shape of this distribution? We can see that it is a sum of two distributions symmetric relative to changing sign $\Delta_i x \to -\Delta_i x$, so their sum is also symmetric, and it is thus sufficient to consider only values $\Delta x \geq 0$. Without losing generality, we can always assume that $\Delta_1 \leq \Delta_2$. In this case, $\min(\Delta_1, \Delta x + \Delta_2) = \Delta_1$, and $-\Delta_1 \geq \Delta x - \Delta_2$ when $\Delta x \leq \Delta_2 - \Delta_1$. So:

- when $|\Delta x| \leq \Delta_2 - \Delta_1$, we get $\rho(\Delta x) = \dfrac{2\Delta_1}{2\Delta_1 \cdot \Delta_2} = \dfrac{1}{2\Delta_2}$;

- when $|\Delta x| \geq \Delta_2 - \Delta_1$, we get $\rho(\Delta x) = \dfrac{\Delta_1 - |\Delta x| + \Delta_2}{4\Delta_1 \cdot \Delta_2} = \dfrac{\Delta - |\Delta x|}{4\Delta_1 \cdot \Delta_2}$.

So:

- when $|\Delta x| \leq \Delta_2 - \Delta_1$, we have a uniform distribution, with a constant pdf, and

- when $\Delta_1 - \Delta_1 \leq |\Delta x| \leq \Delta$, the pdf $\rho(x)$ linearly decreases to 0.

So, in general, we get the trapezoidal distribution – exactly what describes interval uncertainty in the fuzzy approach. Since the whole idea of fuzzy uncertainty is to describe our intuitive (imprecise) knowledge, the fact that we get the same result is a very good sign.

Why cannot we apply the same idea to the case of normal distribution? Indeed, why cannot we apply the same idea to situations when we only know the first two moments. In these situations, similarly, the measurement errors usually consists of several independent components, and the overall mean and variance are obtained by adding the means and variances corresponding to different components: $E = E_1 + \ldots + E_m$ and $V = V_1 + \ldots + V_m$.

A short answer to this question is: yes, we can, but for normal distributions, it will not change anything. Indeed, if we apply the above approach, and apply MaxEnt to each of the components, we conclude that each of the error components is normally distributed with mean E_i and variance V_i. The overall error is thus the sum of several independent normally distributed random variables and is, thus, itself, normal with the mean $E = E_1 + \ldots + E_m$ and variance $V = V_1 + \ldots + V_m$.

References

[1] D. J. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, Chapman & Hall/CRC, Boca Raton, Florida, 2011.

[2] E. T. Jaynes, G. L. Bretthorst, Probability Theory: The Logic of Science, Cambridge University Press, Cambridge, UK, 2003.

[3] H. T. Nguyen, V. Kreinovich, B. Wu, G. Xiang, Computing Statistics under Interval and Fuzzy Uncertainty, Springer Verlag, Berlin, Heidelberg, 2012.

[4] S. G. Rabinovich, Measurement Errors and Uncertainty: Theory and Practice, Springer Verlag, Berlin, 2005.

[5] G. Klir, B. Yuan, Fuzzy Sets and Fuzzy Logic, Prentice Hall, Upper Saddle River, New Jersey, 1995.

[6] H. T. Nguyen, E. A. Walker, A First Course in Fuzzy Logic, Chapman and Hall/CRC, Boca Raton, Florida, 2006.

[7] L. A. Zadeh, Fuzzy sets, Information and Control, 1965, Vol. 8, pp. 338–353.