While, In General, Uncertainty Quantification (UQ) Is NP-Hard, Many Practical UQ Problems Can Be Made Feasible

Ander Gray[1], Scott Ferson[1]
Olga Kosheleva[2], and Vladik Kreinovich[2]
[1]Institute for Risk and Uncertainty, University of Liverpool
Liverpool, UK, {Ander.Gray,Scott.Ferson}@liverpool.ac.uk
University of Texas at El Paso, El Paso, TX 79968, USA
{olgak,vladik}@utep.edu

Presentation at SSCI'2021

1. Uncertainty Quantification (UQ) is important

- Most of our knowledge about the world comes ultimately from measurements.
- We also use theoretical models.
- However, these models – even those that are not directly inspired by measurements results:
  - have to be confirmed by measurements, and
  - are only as reliable as the measurements that confirm them.
- No matter how accurate are the measurements, they are never absolutely accurate.
- In general, the result $\widetilde{x}$ of measuring a quantity x is different from the actual (unknown) value x of this quantity.
- In other words, we have a non-zero measurement error $\Delta x \stackrel{\text{def}}{=} \widetilde{x} - x$.
- Because of this uncertainty, the estimates that we make based on these measurements are also uncertain.

2. Uncertainty Quantification (UQ) is important (cont-d)

- To make proper decisions, we need to understand how accurate are these estimates.
- For example, support that we are prospecting for oil in a certain area and the current estimate is that this area contains 100 million tons of oil.
- Then whether it is very good news or just maybe good news depends on how accurate is this estimate.
- If it is 100 ± 20, then we should start exploiting this oil field;
- However, if it is 100 ± 200, then maybe there is no oil there at all.
- So, it is advisable to do some more measurements before investing money into this area.

3. How uncertainty is usually described

- Let us denote all the quantities that we measure by $x_1, \ldots, x_m$.
- In these terms, we want to know which tuples $X = (x_1, \ldots, x_m)$ are possible, i.e.:
  - what is the set $\mathcal{X}$ of possible tuples, and
  - what are the probabilities of different tuples X from this set.
- Ideally, we should therefore find the probability distribution on the set of all possible tuples.
- But where can we get the corresponding probabilities?
- There is only one source of all the information about the world – and, in particular, about the probabilities – measurements.
- During any period of time, we can only perform finitely many measurements.

4. How uncertainty is usually described (cont-d)

- A general probability distribution requires infinitely many parameters to describe.
- Based on a finite set of measurements, we cannot uniquely determine the probability distribution.
- So, in practice, we never know the exact probability distribution.
- We only have partial information about the actual (unknown) probability distribution.
- We may get different types of information about this distribution:
- For example, we can have bounds on the distribution's moments.
- We can have bounds $\underline{F}_i(v_i) \le F_i(v_i) \le \overline{F}_i(v_i)$ on the values of the cumulative distribution functions $F_i(v_i) \stackrel{\text{def}}{=} \text{Prob}(x_i \le v_i)$.
- Such bounds are known as *probability boxes*, or *p-boxes*, for short.

5. In general, UQ problems are NP-hard

- In general, uncertainty quantification problems are NP-hard.
- This means, crudely speaking, that:
  - unless P = NP (which most computer scientists believe to be false),
  - no feasible algorithm is possible that would solve all possible UQ problems.

6. What we do in this talk

- In this talk, we argue that:
  - if we go down to the level of the original measurement results,
  - then most (if not all) *practical* UQ problems become feasible.
- To make this argument, we need to formulate a general UQ problem, namely, we need to specify:
  - what information can we have, and
  - what do we want to estimate.
- Based on this general description, we need to explain:
  - how we can feasibly estimate what we want
  - based on the information that we have.

7. The main claim of this section

- The main claim of this section is that:
  - many types of partial information about the probability distribution
  - consist of linear inequalities, i.e., inequalities of the type $\int a(X) \cdot f(X)\, dX \le b$ for some function $a(X)$.
- Here $f(X)$ is the actual (unknown) probability density function.
- To support this claim, we will first:
  - describe typical types of partial information, and
  - show that they indeed have this form – or can be equivalently reformulated in this form.
- After that, we provide general arguments that *any* reasonable partial information has this form.

8. Upper bounds vs. lower bounds

- The above formula provides an upper bound b on the statistical characteristic $\int a(X) \cdot f(X)\, dX$.
- In some practical situations, we have lower bounds:

$$b \le \int a(X) \cdot f(X)\, dX.$$

- From the physical viewpoint, lower bounds are different.
- However, from the computational viewpoint, each lower bound can be easily reformulated in terms of equivalent upper bound.
- Namely, if we multiply both sides of the inequality by -1, we get the inequality $\int a'(X) \cdot f(X) \le b'$, where $a'(X) \stackrel{\text{def}}{=} -a(X)$ and $b' \stackrel{\text{def}}{=} -b$.
- Because of this possibility, in the following text, we will mostly consider upper bounds.

9. What if we know the exact value

- Sometimes, we know the exact value b of the corresponding statistical characteristics: $\int a(X) \cdot f(X)\, dX = b$.
- This knowledge can be described as two bounds: $b \le \int a(X) \cdot f(X)\, dX$ and $\int a(X) \cdot f(X)\, dX \le b$.
- This is equivalent to two upper bounds: $\int a(X) \cdot f(X)\, dX \le b$ and $\int (-a(X)) \cdot f(X)\, dX \le (-b)$.

10. Continuous vs. discrete

- From the purely mathematical viewpoint, most physical quantities x can take any real-number values.
- There are infinitely many possible real numbers.
- However, values which are too close to each other are practically indistinguishable.
- Also, for each measuring instrument, there are natural bounds within which it can provide meaningful measurements. For example:
  - a ruler cannot be used to measure distances larger than a certain amount,
  - all scales have their limitations after which the scale will be simply crushed by the weight,
  - thermometers melt if the temperature is too high and freeze when it is too low, etc.

11. Continuous vs. discrete (cont-d)

- Because of this, in reality, each quantity x_i has only finitely many practically distinguishable values:
  - the smallest detectable value $x_{i,0}$,
  - the next value $x_{i,1} = x_{i,0} + h$, where h_i is the smallest difference that makes practical sense,
  - the value $x_{i,2} = x_{i,0} + 2h_i$, etc., all the way to
  - the largest possible value $x_{i,n_i} = x_{i,0} + N_i \cdot h_i$ for an appropriate N_i.
- As a result, there are finitely many possible values of the tuple $X = (x_1, \ldots, x_m)$: only the tuples $X_{n_1,\ldots,n_m} \stackrel{\text{def}}{=} (x_{1,n_1}, \ldots, x_{m,n_m})$.
- In these terms:
  - to describe the probability distribution,
  - it is sufficient to describe the probability

$$p_{n_1,\ldots,n_m} \stackrel{\text{def}}{=} \text{Prob}(X = X_{n_1,\ldots,n_m}) \text{ of each possible tuple.}$$

12. Continuous vs. discrete (cont-d)

- In terms of these probabilities, the general inequality takes the form

$$\sum_{n_1,\ldots,n_m} a(X_{n_1,\ldots,n_m}) \cdot p_{n_1,\ldots,n_m} \le b.$$

- The left-hand side of the formula is the integral sum.
- When h is small – and usually, it is small – the integral sum is very close to the actual integral.
- Thus, from the practical viewpoint, we will consider continuous and discrete formulas interchangeable.

13. Moments and bounds on moments

- Let us start providing examples of partial information about probabilities that can be described in this linear form.
- Our first example is *moments*, i.e., expressions of the type

$$M_{k_1,\ldots,k_m} \stackrel{\text{def}}{=} \int x_1^{k_1} \cdot \ldots \cdot x_m^{k_m} \cdot f(x_1,\ldots,x_m)\, dx_1 \ldots dx_m.$$

- In discrete form,

$$M_{k_1,\ldots,k_m} = \sum_{n_1=1}^{N_1} \ldots \sum_{n_m=1}^{N_m} x_{1,n_1}^{k_1} \cdot \ldots \cdot x_{m,n_m}^{k_m} \cdot p_{n_1,\ldots,n_m}.$$

- Both expressions are clearly linear in terms of the corresponding probabilities.

14. Cumulative distribution functions (cdf) and p-boxes

- Another possible information is information about (i.e., bounds on) the cumulative distribution function $F(v)$ of:
  - either the quantities x_i themselves,
  - or, more generally, of a quantity $y = s(X) = s(x_1, \ldots, x_n)$ depending on these quantities.
- The value $F(v) = \text{Prob}(y \le v)$ can be described as

$$F(v) = \int_{X:\ s(X) \le v} f(X)\, dX.$$

- So, it is equivalent to $F(v) = \int_X a(X) \cdot f(X)\, dX$, where:
  - we have $a(X) = 1$ if $s(X) \le v$, and
  - we have $a(X) = 0$ otherwise.

15. Information about the probability density function

- We can also have information about (i.e., bounds on) the values $f(X)$ of the probability density function $f(X)$ itself.
- In this case, of course, the bounds $f(X) \leq b$ are already inequalities which are linear in terms of $f(X)$.

16. Symmetry information

- We may also have information about the invariance of these characteristics with respect to some transformations.
- For example, we may know that a 1-D distribution $f(x_1)$ is symmetric with respect to the transformation $x_1 \mapsto -x_1$.
- This invariance means that $f(-x_1) = f(x_1)$, i.e., that

$$f(-x_1) - f(x_1) = 0.$$

- This is also a linear equality in terms of the function $f(X)$ – and can, thus, be described as two linear inequalities.

17. What about the general case

- In general, how do we estimate a general statistical characteristic?
- Like every other knowledge about the words, we estimates these characteristics based on the measurement results.
- Specifically, we have several tuples $X^{(1)}, \ldots, X^{(K)}$ corresponding to different independent measurements.
- We want to estimate the desired characteristics based on this K-element sample.
- In practice, at any given moment of time, we can have only finitely many measurement results.
- Based on these measurement results, we can only determine the values of finitely many parameters.
- Thus, we restrict ourselves – explicitly or implicitly – to a finite-parametric family of distributions $f(X, c_1, \ldots, c_t)$.

18. What about the general case (cont-d)

- For example, in engineering applications, we often explicitly restrict ourselves to Gaussian distributions.
- There, the known formula describes the pdf in terms of means and the covariance matrix.
- A widely used way to determine a probability distribution based on partial information is the Maximum Entropy approach:
  - among all probability distributions $f(X)$ which are consistent with observations,
  - we select the distribution for which the entropy $-\int f(X) \cdot \ln(f(X))\, dX$ attains the largest possible value.
- We can use machine learning to determine the desired distribution.
- We still get a finite-parametric family of the distributions.
- This time, this family is implicitly described.

19. What about the general case (cont-d)

- There is no simple analytical expression for distributions from this family.
- In terms of the family $f(X, c_1, \ldots, c_t)$, identifying a distribution means estimating the values of all the parameters c_j.
- How can we estimate these parameters based on the sample?
- For each combination of values c_j, the probability of observing each tuple $X^{(k)}$ is equal to $f\left(X^{(k)}, c_1, \ldots, c_t\right)$.
- Since these measurements are independent, the probability that we have observed all tuples is equal to the product of these probabilities:

$$f\left(X^{(1)}, c_1, \ldots, c_t\right) \cdot \ldots \cdot f\left(X^{(K)}, c_1, \ldots, c_t\right).$$

- This value represents, in effect, the probability that the values c_j are the good fit for the observed tuples.
- We need to select a single combination of the parameters c_j.

20. What about the general case (cont-d)

- A reasonable idea is to select:
  - the most probable combination $c = (c_1, \ldots, c_t)$, i.e.,
  - the combination for which the above product attains the largest possible value.
- This is known as the *Maximum Likelihood* approach.
- It is one of the most widely used techniques for estimating the values of different statistical characteristics.
- How does this lead to linear inequalities?
- Maximizing the product is equivalent to maximizing its logarithm, i.e., the sum $\sum_{k=1}^{K} \ln\left(f\left(X^{(k)}, c_1, \ldots, c_t\right)\right)$.

21. What about the general case (cont-d)

- According to calculus, the maximum of a function is attained when all its partial derivatives are equal to 0, i.e., when for each $j = 1, \ldots, t$

$$\sum_{k=1}^{K} \frac{\partial}{\partial c_j} \left(\ln \left(f \left(X^{(k)}, c_1, \ldots, c_t \right) \right) \right) = 0.$$

- So, $\sum\limits_{k=1}^{K} a_j \left(X^{(k)} \right) = 0$, where $a_j(X) \stackrel{\text{def}}{=} \dfrac{\partial}{\partial c_j} \left(\ln \left(f \left(X, c_1, \ldots, c_t \right) \right) \right)$.
- For any function $a(X)$, a natural estimate of its mean value $\int a(X) \cdot f(X)\, dX$ based on sample is $\dfrac{a\left(X^{(1)}\right) + \ldots + a\left(X^{(K)}\right)}{K}$.
- The above formula implies that the sample average of the values $a_j(X)$ is equal to 0: $\dfrac{a_j\left(X^{(1)}\right) + \ldots + a_j\left(X^{(K)}\right)}{K} = 0.$

22. What about the general case (cont-d)

- Thus, the corresponding mean value $\int a_j(X) \cdot f(X)\, dX$ is also close to 0.
- In other words, the estimates c_j are equivalent to linear inequalities

$$-\varepsilon \le \int a_j(X) \cdot f(X)\, dX \le \varepsilon.$$

- So, in the general case, we can formulate any partial knowledge about a probability distribution in terms of linear inequalities.

23. What Do We Want to Estimate

- The main claim of this section is that:
  - the decisions of a rational decision makers are equivalent to
  - maximizing some expression $c = \int c(X) \cdot f(X)\, dX$ which is linear in terms of the probabilities $f(X)$.
- To justify this claim, we will use general ideas of decision theory.

24. How to describe preferences in numerical terms

- Computers have been invented to deal with numbers.
- Numbers are still what computers process most efficiently; so:
  - to enable computers to help us make decisions,
  - it is desirable to describe all available information into numbers.
- In particular, it is desirable to transform information about our preferences into numbers.
- To perform this transformation, we need to have a numerical scale for preferences.
- This scale can be constructed as follows. First, let us select two alternatives:
  - an alternative A_- which is worse than anything that we can potentially encounter; we will call this alternative *very bad*; and
  - an alternative A_+ which is better than anything that we can potentially encounter; we will call this alternative *very good*.

25. How to describe preferences in numerical terms (cont-d)

- Then, for all numbers p from the interval $[0, 1]$, we can form a lottery in which:
  - we get a very good alternative A_+ with probability p, and
  - we get a very bad alternative A_- with the remaining probability $1 - p$.
- We will denote this lottery by $L(p)$.
- Now, let us consider any of the actual alternatives A.
- When $p \approx 0$, the lottery $L(p)$ is close to the very bad alternative A_- and is, thus, worse than A; we will denote it by $L(p) < A$.
- When $p \approx 1$, the lottery $L(p)$ is close to the very good alternative A_+ and is, thus, better than A: $A < L(p)$.
- As the probability p of getting the very good alternative increases, the lottery $L(p)$ becomes better and better.
- At some point, we will switch from $L(p) < A$ to $A < L(p)$.

26. How to describe preferences in numerical terms (cont-d)

- This threshold value $u(A)$ is called the *utility* of the alternative A.
- For this value:
  - we have $L(p) < A$ for $p < u(A)$, and
  - we have $A < L(p)$ for $p > u(A)$
- By definition of the utility, for every $\varepsilon > 0$, we have

$$L(u(A) - \varepsilon) < A < L(u(A) + \varepsilon).$$

- When ε is sufficiently small, there is no way to practically distinguish probabilities $u(A)$ and $u(A) \pm \varepsilon$.
- Thus, from the practical viewpoint, the alternative A is equivalent to the lottery $L(u(A))$.
- We will denote this practical equivalence by $A \equiv L(u(A))$.

27. How to describe preferences in numerical terms (cont-d)

- For lotteries $L(p)$, the larger the probability, the better:

$$p < q \Leftrightarrow L(p) < L(q).$$

- Thus, in general, $A < B$ if and only $u(A) < u(B)$.
- So, utilities indeed provide a numerical representation of preferences.

28. How to describe preferences under uncertainty

- In practice, we often cannot predict with 100% certainty what will be the consequence of each action.
- Suppose that for some action a, possible results are $A_1, \ldots, A_r$ with probabilities $p_1, \ldots, p_r$.
- So, this action is equivalent to a lottery in which we get each alternative A_i with probability p_i.
- Each alternative A_i, in its turn, is equivalent to a lottery in which:
  - we get A_+ with probability $u(A_i)$, and
  - we get A_- with the remaining probability $1 - u(A_i)$.
- Thus, the action a is equivalent to a two-stage lottery, in which:
  - first, we select one of the alternatives A_i with probability p_i, and
  - then, depending on what we selected on the first stage, we select A_+ with probability $u(A_i)$ or A_- with probability $1 - u(A_i)$.

29. How to describe preferences under uncertainty (cont-d)

- As a result of this two-stage lottery, we get either A_+ or A_-.
- One can see that the probability of selecting A_+ in this two-stage lottery is equal to the sum $\sum_{i=1}^{r} p_i \cdot u(A_i)$.
- By definition of utility, this means that this formula describes the utility of the action a.
- So, this formula describes the quality of each action – as a linear combination of the corresponding probabilities.
- Thus, to decide which action is better, we need to estimate this expression for different actions.
- This expression is a linear function of probabilities.

30. How Can We Feasibly Estimate the Desired Quantities: General Case

- According to the above analysis, our knowledge of probabilities $p_{n_1,\ldots,n_m}$ can be described in terms of linear inequalities

$$\sum_{n_1=1}^{N_1} \ldots \sum_{n_m=1}^{N_m} a^{(\ell)}_{n_1,\ldots,n_m} \cdot p_{n_1,\ldots,n_m} \le b^{(\ell)}, \quad \ell = 1, \ldots, L.$$

- Based on this information, we want to estimate the value of the objective function $c = \sum_{n_1=1}^{N_1} \ldots \sum_{n_m=1}^{N_m} c_{n_1,\ldots,n_m} \cdot p_{n_1,\ldots,n_m}$.
- In general, due to uncertainty, the value of the objective function is not uniquely determined by the available data.
- We can have the whole range $[\underline{c}, \overline{c}]$ of possible values.
- The lower endpoint $\underline{c}$ of this range can be found if we minimize c under the above constraints.

31. How Can We Feasibly Estimate the Desired Quantities: General Case (cont-d)

- The upper endpoint $\overline{c}$ of the range can be found if we maximize c under above constraints.
- In both cases, we optimize a linear expression under linear inequalities.
- Such optimization problems are known as *linear programming* problems.
- Good news is that there exists feasible algorithms for solving linear programming problems.
- Thus, indeed, we can conclude that many practical uncertainty optimization problems are feasible.

32. Example

- Suppose that we know:
  - the joint probability of x_1 and x_2, and
  - the joint probability of x_2 and x_3.
- What can we then conclude about the joint probability of x_1 and x_3?
- In other words, for each n_1, n_2, and n_3, we know:
  - the values $p^{(1,2)}_{n_1,n_2} = \text{Prob}(x_1 = x_{1,n_1} \,\&\, x_2 = x_{2,n_2})$ and
  - the values $p^{(2,3)}_{n_2,n_3} = \text{Prob}(x_2 = x_{1,n_2} \,\&\, x_3 = x_{3,n_3})$.
- Based on this information, we need to estimate the values

$$p^{(1,3)}_{n_1,n_3} = \text{Prob}(x_1 = x_{1,n_1} \,\&\, x_3 = x_{3,n_3}).$$

- In other words, we need to find the ranges $\left[\underline{p}^{(1,3)}_{n_1,n_3}, \overline{p}^{(1,3)}_{n_1,n_3}\right]$.

33. Example (cont-d)

- In terms of the unknowns p_{n_1,n_2,n_3}, each desired value $p^{(1,3)}_{n_1,n_3}$ has the form $\sum_{n_2=1}^{N_2} p_{n_1,n_2,n_3}$.
- The available information has the form

$$\sum_{n_3=1}^{N_3} p_{n_1,n_2,n_3} = p^{(1,2)}_{n_1,n_2} \text{ and } \sum_{n_1=1}^{N_1} p_{n_1,n_2,n_3} = p^{(2,3)}_{n_2,n_3}.$$

34. Can we get non-trivial bounds this way?

- Sometimes, we may get trivial bounds $\underline{p}^{(1,3)}_{n_1,n_3} = 0$ and $\overline{p}^{(1,3)}_{n_1,n_3} = 1$.
- However, there are cases when we get non-trivial bounds.
- For example, suppose that for all three variables:
  - we have the same sequence of values $x_{1,i} = x_{2,i} = x_{3,i}$ for all i, and
  - we have $p^{(1,2)}_{i,j} = 0$ for $i \neq j$ and $p^{(2,3)}_{j,k} = 0$ when $j \neq k$.
- This means that:
  - with probability 1, $x_1 = x_2$, and
  - with probability 1, $x_2 = x_3$.
- In this case, we conclude that $x_1 = x_3$, i.e., that $p^{(1,3)}_{i,k} = 0$ for all $i \neq k$.
- One can see that in this situation, the above linear programming problems lead to $\underline{p}^{(1,3)}_{i,k} = \overline{p}^{(1,3)}_{i,k} = 0$ for all $i \neq k$.

35. Acknowledgments

- This work was supported in part by the National Science Foundation grants:
  - 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and
  - HRD-1834620 and HRD-2034030 (CAHSI Includes).
- It was also supported by the AT&T Fellowship in Information Technology.
- It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.