How to Estimate Amount of Useful Information, in Particular Under Imprecise Probability

Luc Longpré[1], Olga Kosheleva[2], and Vladik Kreinovich[1]
[1]Department of Computer Science
[2]Department of Teacher Education
University of Texas at El Paso
El Paso, TX 79968, USA
longpre@utep.edu, olgak@utep.edu, vladik@utep.edu

How to Gauge the . . .
Finite Case
Finite Case with . . .
How to Gauge . . .
Need to Distinguish . . .
Such Distinction Is . . .
Such Distinction Is . . .
How to Estimate the . . .
What If We Only Have . . .

1. How to Gauge the Amount of Information: General Idea

- Our ultimate goal is to gain a complete knowledge of the world.
- In practice, we usually have only *partial* information.
- In other words, in practice, we have *uncertainty*.
- Additional information allows us to decrease this uncertainty.
- It is therefore reasonable to:
  - gauge the amount of information in the new knowledge
  - by how much this information decreases the original uncertainty.
- Uncertainty means that for some questions, we do not have a definite answer.

2. Gauging Amount of Information (cont-d)

- Once we learn the answers to these questions, we thus decrease the original uncertainty.
- It is therefore reasonable to:
  - estimate the amount of uncertainty
  - by the number of questions needed to eliminate this uncertainty.
- Of course, not all questions are created equal:
  - some can have a simple binary "yes"-"no" answer;
  - some look for a more detailed answer – e.g., we can ask what is the value of a certain quantity.
- No matter what is the answer, we can describe this answer inside the computer.
- Everything in a computer is represented as 0s and 1s.

How to Gauge the . . .
Finite Case
Finite Case with . . .
How to Gauge . . .
Need to Distinguish . . .
Such Distinction Is . . .
Such Distinction Is . . .
How to Estimate the . . .
What If We Only Have . . .

3. Gauging Amount of Information (cont-d)

- Everything in a computer is represented as 0s and 1s.
- So, each answer is a sequence of 0s and 1s.
- Such a several-bits question can be represented as a sequence of on-bit questions:
  - we can first ask what is the first bit of the answer,
  - we can then ask what is the second bit of the answer, etc.
- So, every question can thus be represented as a sequence of one-bit ("yes"-"no") questions.
- So, it is reasonable to:
  - measure uncertainty
  - by the smaller number of such "yes"-"no" questions which are needed to eliminate this uncertainty.

How to Gauge the ...
Finite Case
Finite Case with ...
How to Gauge ...
Need to Distinguish ...
Such Distinction Is ...
Such Distinction Is ...
How to Estimate the ...
What If We Only Have ...

4. Finite Case

- Let us first consider the situation when we have finitely many N alternatives.
- If we ask one binary question, then we can get two possible answers (0 and 1).
- Thus, we can uniquely determine one of the two different states.
- If we ask 2 binary questions, then we can get four possible combinations of answers (00, 01, 10, and 11).
- In general, if we ask q binary questions, then we can get 2^q possible combinations of answers.
- Thus, we can uniquely determine one of 2^q states.
- So, to identify one of n states, we need to ask q questions, where $2^q \geq N$.
- The smallest such q is $\lceil \log_2(N) \rceil$.

How to Gauge the . . .
Finite Case
Finite Case with . . .
How to Gauge . . .
Need to Distinguish . . .
Such Distinction Is . . .
Such Distinction Is . . .
How to Estimate the . . .
What If We Only Have . . .

5. Finite Case with Known Probabilities

- So far, we considered the situation when we have n alternatives about whose frequency we know nothing.
- In practice, we often know the probabilities $p_1, \ldots, p_n$ of different alternatives; in this case:
  - instead of considering the *worst-case* number of binary questions needed to eliminate uncertainty,
  - it is reasonable to consider the *average* number of questions.
- This value can be estimated as follows.
- We have a large number N of similar situations with n-uncertainty.
- In $N \cdot p_1$ of these situations, the actual state is State 1.
- In $N \cdot p_2$ of them, the actual state is State 2, etc.

6. Case of Known Probabilities (cont-d)

- The average number of binary questions can be obtained if we divide:
  - the overall number of questions needed to determine the states in all N situations,
  - by N.
- There are $\binom{N}{N\cdot p_1} = \dfrac{N!}{(N\cdot p_1)!\cdot(N-N\cdot p_1)!}$ ways to select the situations in State 1.
- Out of these, there are many ways to to select $N\cdot p_2$ situations in State 2:

$$\binom{N-N\cdot p_1}{N\cdot p_2} = \frac{(N-N\cdot p_1)!}{(N\cdot p_2)!\cdot(N-N\cdot p_1-N\cdot p_2)!}.$$

- So, the number A of possible arrangements is:

$$\frac{N!}{(N\cdot p_1)!\cdot(N-N\cdot p_1)!}\cdot\frac{(N-N\cdot p_1)!}{(N\cdot p_2)!\cdot(N-N\cdot p_1-N\cdot p_2)!}\cdot\ldots$$

How to Gauge the . . .
Finite Case
Finite Case with . . .
How to Gauge . . .
Need to Distinguish . . .
Such Distinction Is . . .
Such Distinction Is . . .
How to Estimate the . . .
What If We Only Have . . .

7. Case of Known Probabilities (final)

- Thus, $A = \dfrac{N!}{(N \cdot p_1)! \cdot (N \cdot p_2)! \cdot \ldots \cdot (N \cdot p_n)!}$.
- To identify an arrangement, we need to ask the following number of binary questions:

$$Q = \log_2(A) = \log_2(N!) - \sum_{i=1}^{n} \log_2((N \cdot p_i)!).$$

- Here, $m! \sim \left(\dfrac{m}{e}\right)^m$, so

$$\log_2(m!) \sim m \cdot (\log_2(m) - \log_2(e)).$$

- As a result, we get the usual Shannon's formula:

$$\overline{q} = -\sum_{i=1}^{n} p_i \cdot \log_2(p_i).$$

8. How to Gauge Uncertainty: Continuous Case

- In the continuous case, when the unknown(s) can take any of the infinitely many values from some interval.
- So, we need infinitely many binary questions to uniquely determine the exact value.
- It thus makes sense to determine each value with a given accuracy $\varepsilon > 0$:
  - we divide the real line into intervals $[x_i - \varepsilon, x_i + \varepsilon]$, where $x_{i+1} = x_i + 2\varepsilon$, and
  - we want to find out to which of these intervals the actual value x belongs.
- For small ε, the probability p_i of belonging to the i-th interval is equal to $p_i \approx \rho(x_i) \cdot (2\varepsilon)$.
- Substituting this expression for p_i into Shannon's formula, we get the following formula:

9. Continuous Case (cont-d)

$$\overline{q} = -\sum_{i=1}^{n} p_i \cdot \log_2(p_i) = -\sum_{i=1}^{n} \rho(x_i) \cdot (2\varepsilon) \cdot \log_2(\rho(x_i) \cdot (2\varepsilon)), \text{ i.e.,}$$

$$\overline{q} = -\sum_{i=1}^{n} \rho(x_i) \cdot (2\varepsilon) \cdot \log_2(\rho(x_i)) - \sum_{i=1}^{n} \rho(x_i) \cdot (2\varepsilon) \cdot \log_2(2\varepsilon).$$

- The first term in this sum has the form

$$-\sum_{i=1}^{n} \rho(x_i) \cdot \log_2(\rho(x_i)) \cdot (2\varepsilon) = -\sum_{i=1}^{n} \rho(x_i) \cdot \log_2(\rho(x_i)) \cdot \Delta x_i.$$

- This term is an integral sum for the interval

$$-\int \rho(x) \cdot \log_2(\rho(x))\, dx.$$

- Thus, for small ε, it is practically equal to this interval.

How to Gauge the . . .
Finite Case
Finite Case with . . .
How to Gauge . . .
Need to Distinguish . . .
Such Distinction Is . . .
Such Distinction Is . . .
How to Estimate the . . .
What If We Only Have . . .

10. Continuous Case (final)

- Similarly, the second term has the form

$$-\sum_{i=1}^{n} \rho(x_i) \cdot (2\varepsilon) \cdot \log_2(2\varepsilon) = -\log_2(2\varepsilon) \cdot \sum_{i=1}^{n} \rho(x_i) \cdot \Delta x_i.$$

- The 2nd terms is, thus, an integral sum for

$$-\log_2(2\varepsilon) \cdot \int \rho(x)\,dx = -\log_2(2\varepsilon).$$

- So, the average number of binary questions $\overline{q}$ which is needed to determine x with accuracy ε is equal to

$$\overline{q} = -\int \rho(x) \cdot \log_2(\rho(x))\,dx - \log_2(2\varepsilon).$$

- The first term does not depend on ε, and is, thus, a good measure of how much uncertainty we have.
- This term is exactly Shannon's entropy.

How to Gauge the . . .
Finite Case
Finite Case with . . .
How to Gauge . . .
Need to Distinguish . . .
Such Distinction Is . . .
Such Distinction Is . . .
How to Estimate the . . .
What If We Only Have . . .

11. Need to Distinguish Between Useful and Unimportant Information

- A similar formula holds in the multi-D case:

$$S = -\int \rho(\vec{x}) \cdot \log_2(\rho(\vec{x}))\, d\vec{x}.$$

- Not all information is created equal:
  - some pieces of information are useful, while
  - other pieces of information are unimportant.
- Whether the information is useful or not depends on what we plan to do with this information:
  - if we want to predict weather, the smell of the fog is unimportant, while
  - if we are analyzing pollution level, this is a very useful information.

How to Gauge the . . .
Finite Case
Finite Case with . . .
How to Gauge . . .
Need to Distinguish . . .
Such Distinction Is . . .
Such Distinction Is . . .
How to Estimate the . . .
What If We Only Have . . .

12. Such Distinction Is Important for Privacy

- Ideally, no one can gain any information about a person without his or her explicit permission.
- Realistically, some information may be leaked.
- It is therefore important to distinguish the cases:
  - when an important information was leaked and
  - when an unimportant information was leaked.
- For example, disclosing the higher bits of the salaries would be a major violation of privacy.
- However, disclosing the lowest bits (number of cents) is mostly harmless.
- How to estimate the amount of *useful* information, that affects the utility of different alternatives?

How to Gauge the . . .
Finite Case
Finite Case with . . .
How to Gauge . . .
Need to Distinguish . . .
Such Distinction Is . . .
Such Distinction Is . . .
How to Estimate the . . .
What If We Only Have . . .

13. Such Distinction Is Important in Education

- Psychological studies show that (almost) all students are capable of learning, with $\pm 10\%$ difference.
- Groups originally viewed as inferior (e.g., girls) have shown equal abilities.
- However, the results of studying differ in orders of magnitude.
- To explain this difference, psychologists asked kids to recall everything they remember from the class.
- All kids recalled the same number of bits, but:
  - good students recalled the class material, while
  - failing students recalled mostly irrelevant details.
- This fact can be used to speed up learning, by blocking irrelevant information (e.g., no windows).

How to Gauge the . . .
Finite Case
Finite Case with . . .
How to Gauge . . .
Need to Distinguish . . .
Such Distinction Is . . .
Such Distinction Is . . .
How to Estimate the . . .
What If We Only Have . . .

14. How to Estimate the Amount of Useful Information: A Suggestion

- According to decision theory, the usefulness of a situation x to a user can be described by *utility* $u(x)$.
- So, we propose to count the number of binary questions that are needed to determine $u(x)$ with $\varepsilon > 0$.
- From this viewpoint, if some variable is irrelevant, then it does not affect the utility at all.
- So we should not waste binary questions trying to find the value of this variable.
- If some variable is slightly relevant, then its very crude estimate will give us ε-accuracy in $u(x)$.
- Therefore, few questions will be needed.
- On the other hand, if a variable is highly relevant, then we need exactly as many questions as before.

How to Gauge the ...
Finite Case
Finite Case with ...
How to Gauge ...
Need to Distinguish ...
Such Distinction Is ...
Such Distinction Is ...
How to Estimate the ...
What If We Only Have ...

15. Towards a Precise Definition: 1-D Case

- In the 1-D case:
  - if we know x with uncertainty Δx,
  - then we know the utility with accuracy

$$u(x+\Delta x)-u(x)\approx u'(x)\cdot\Delta x.$$

- Thus, to get $u(x)$ with accuracy ε, we must determine x with accuracy $\Delta x=\dfrac{\varepsilon}{|u'(x)|}$.
- In this case, we divide the real line into intervals $\left[x_i-\dfrac{\varepsilon}{|u'(x_i)|},x_i+\dfrac{\varepsilon}{|u'(x_i)|}\right]$, where $x_{i+1}=x_i+\dfrac{2\varepsilon}{|u'(x_i)|}$.
- For small ε, the probability p_i of belonging to the i-th interval is equal to

$$p_i\approx\rho(x_i)\cdot\Delta x_i=\rho(x_i)\cdot\frac{2\varepsilon}{|u'(x_i)|},\ \text{where } \Delta x_i\stackrel{\text{def}}{=}x_{i+1}-x_i.$$

How to Gauge the ...
Finite Case
Finite Case with ...
How to Gauge ...
Need to Distinguish ...
Such Distinction Is ...
Such Distinction Is ...
How to Estimate the ...
What If We Only Have ...

16. 1-D Case (cont-d)

- Substituting the expression for p_i into Shannon's formula, we get:

$$\overline{q} = -\sum_{i=1}^{n} p_i \cdot \log_2(p_i) = -\sum_{i=1}^{n} \rho(x_i) \cdot \Delta x_i \cdot \log_2\left(\rho(x_i) \cdot \frac{2\varepsilon}{|u'(x_i)|}\right) =$$

$$-\sum_{i=1}^{n} \rho(x_i) \cdot \Delta x_i \cdot \log_2\left(\frac{\rho(x_i)}{|u'(x_i)|}\right) - \sum_{i=1}^{n} \rho(x_i) \cdot \Delta x_i \cdot \log_2(2\varepsilon).$$

- The first term is an integral sum for

$$-\int \rho(x) \cdot \log_2\left(\frac{\rho(x)}{|u'(x)|}\right) dx.$$

- Thus, $\overline{q} = -\int \rho(x) \cdot \log_2\left(\frac{\rho(x)}{|u'(x)|}\right) dx - \log_2(2\varepsilon).$

How to Gauge the . . .
Finite Case
Finite Case with . . .
How to Gauge . . .
Need to Distinguish . . .
Such Distinction Is . . .
Such Distinction Is . . .
How to Estimate the . . .
What If We Only Have . . .

17. 1-D Case (final)

- We can thus view the corresponding term as an amount of useful information:

$$S_u \stackrel{\text{def}}{=} -\int \rho(x) \cdot \log_2\left(\frac{\rho(x)}{|u'(x)|}\right)\,dx.$$

- Here, $S_u = S + \int \rho(x) \cdot \log_2(|u'(x)|)\,dx$, where S is the traditional Shannon's entropy.
- The additional integral term is the mathematical expectation of $\log_2(|u'(x)|)$.
- When $u(x) = x$, the new expression coincides with the traditional Shannon's entropy formula.
- The smaller the derivative $|u'(x)|$:
  - the less relevant the variable x, and
  - the smaller the amount S_u of useful information.

18. Multi-D Case

- For each x_j, the interval that guarantees accuracy ε in $u(x)$ has the width $\Delta x_j = \dfrac{2\varepsilon}{|u_{,j}|}$, where $u_{,j} \stackrel{\text{def}}{=} \dfrac{\partial u}{\partial x_j}$.
- Thus, we divide the m-dimensional space into zones of volume $\Delta V = \dfrac{(2\varepsilon)^m}{\prod\limits_{j=1}^{m} |u_{,j}|}$ and prob. $p_i = \rho(\vec{x}_i) \cdot \Delta V$.
- Hence, $\overline{q} = -\sum p_i \cdot \log_2(p_i) = S_u - \log_2(2\varepsilon)$, where:

$$S_u \stackrel{\text{def}}{=} -\int \rho(\vec{x}) \cdot \log_2 \left(\frac{\rho(\vec{x})}{\prod\limits_{j=1}^{m} |u_{,j}(\vec{x})|} \right) d\vec{x} =$$

$$S + \sum_{i=1}^{m} \int \rho(\vec{x}) \cdot \log_2(|u_{,j}(\vec{x})|)\, d\vec{x}.$$

How to Gauge the ...
Finite Case
Finite Case with ...
How to Gauge ...
Need to Distinguish ...
Such Distinction Is ...
Such Distinction Is ...
How to Estimate the ...
What If We Only Have ...

19. What If We Only Have Partial Information About the Probabilities

- In practice, however, we only have partial information about the probabilities.
- Specifically, we do not know the exact value $\rho(\vec{x})$.
- Instead, we only know a lower bound $\underline{\rho}(\vec{x})$ and an upper bound $\overline{\rho}(\vec{x})$ on the actual (unknown) value $\rho(\vec{x})$:

$$\rho(\vec{x}) \in [\underline{\rho}(\vec{x}), \overline{\rho}(\vec{x})].$$

- Many different probability distributions are consistent with this interval information.
- For different such distributions, in general, we get different values for the amount S_u of useful information.
- We do not know which of the distributions are more probable and which are less probable.

How to Gauge the . . .
Finite Case
Finite Case with . . .
How to Gauge . . .
Need to Distinguish . . .
Such Distinction Is . . .
Such Distinction Is . . .
How to Estimate the . . .
What If We Only Have . . .

20. Case of Partial Information (cont-d)

- Thus, we do not know which values of S_u are more probable and which are less probable.
- It thus makes sense to characterize the uncertainty by the worst case scenario, i.e., by the largest S_u:

$$\overline{S}_u \stackrel{\text{def}}{=} \max\left\{S_u : \underline{\rho}(\vec{x}) \le \rho(\vec{x}) \le \overline{\rho}(\vec{x}) \text{ for all } x \text{ and } \int \rho(\vec{x})\,d\vec{x} = 1\right\}.$$

- To find $\overline{S}_u$, we can use efficient convex optimization algorithms, since:
  - the objective function S_u is concave and
  - the corresponding domain is convex:

$$\left\{\rho(\vec{x}) : \underline{\rho}(\vec{x}) \le \rho(\vec{x}) \le \overline{\rho}(\vec{x}) \text{ for all } x \text{ and } \int \rho(\vec{x})\,d\vec{x} = 1\right\}.$$

How to Gauge the . . .
Finite Case
Finite Case with . . .
How to Gauge . . .
Need to Distinguish . . .
Such Distinction Is . . .
Such Distinction Is . . .
How to Estimate the . . .
What If We Only Have . . .

21. Acknowledgements

This work was supported in part:

- by the National Science Foundation grants
  - HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and
  - DUE-0926721, and
- by an award from Prudential Foundation.

How to Gauge the . . .
Finite Case
Finite Case with . . .
How to Gauge . . .
Need to Distinguish . . .
Such Distinction Is . . .
Such Distinction Is . . .
How to Estimate the . . .
What If We Only Have . . .