How to Describe Relative Approximation Error? A New Justification for Gustafson's Logarithmic Expression

Martine Ceberio, Olga Kosheleva, and Vladik Kreinovich
University of Texas at El Paso, El Paso, Texas 79968, USA
mceberio@utep.edu, olgak@utep.edu, vladik@utep.edu

1. It is desirable to describe relative approximation error

- Suppose that we use a value a to approximate a value b.
- Then the natural number of accuracy of this approximation is the absolute value $|a-b|$ of the difference between these two values.
- This quantity is known as the *absolute approximation error*.
- In many practical cases, both a and b represent values of some physical quantity.
- In such cases, the absolute error changes when we replace the original measuring unit with the one which is $\lambda > 0$ times smaller.
- After this replacements, the numerical values describing the corresponding quantities get multiplied by λ:

$$a \mapsto a' = \lambda \cdot a \text{ and } b \mapsto b' = \lambda \cdot b.$$

- For example, if we replace meters by centimeters, 1.7 m becomes $100 \cdot 1.7 = 170$ cm.

2. It is desirable to describe relative approximation error

- In this case, the numerical value of the absolute approximation error also gets multiplied by λ :

$$|a' - b'| = |\lambda \cdot a - \lambda \cdot b| = \lambda \cdot |a - b|.$$

- It is sometimes desirable to provide a measure of approximation error that would not depend on the choice of the measuring unit.
- Such measures are known as *relative approximation error*.

3. Traditional description of relative approximation error and its limitations

- Usually, the relative approximation errors is described by the ratio

$$\frac{|a-b|}{b}.$$

- The problem is that intuitively, the value a approximates the value b with exactly the same accuracy as b approximates a.

- However, for the measure, this is not true; for example:
  - the value 0.8 approximates 1 with accuracy $\dfrac{|0.8-1|}{1}=0.2$,
  - while 1 approximates 0.8 with accuracy $\dfrac{|1-0.8|}{0.8}=0.25\neq 0.2$.

4. Logarithmic measure of relative accuracy

- How can we avoid the above-described asymmetry?
- John Gustafson proposed to use the following alternative expression for relative approximation accuracy $|\ln(a/b)|$.
- One can easily check that this expression is indeed symmetric:

$$|\ln(a/b)| = |\ln(b/a)|.$$

- A natural question is:
  - there can be several different symmetric measures,
  - why logarithmic one?
- In this talk, we provide a natural explanation for selecting the logarithmic measure.

5. What we want

- What we want is, in effect, a metric on the set $\mathbb{R}_+$ of all positive real numbers, i.e., a function $d(a,b) \geq 0$ for which:
  - $d(a,b) = 0$ if and only if $a = b$;
  - $d(a,b) = d(b,a)$ for all a and b, and
  - $d(a,c) \leq d(a,b) + d(b,c)$ for all a, b, and c.
- We want this metric to be *scale-invariant* in the following precise sense:
- We say that a metric $d(a,b)$ on the set of all positive real numbers is *scale-invariant* if

$$d(\lambda \cdot a, \lambda \cdot b) = d(a,b) \text{ for all } \lambda > 0, a > 0, \text{ and } b > 0.$$

6. What we want (cont-d)

- It is also reasonable to require that the desired metric – just like the usual Euclidean metric:
  - is uniquely generated by its local properties,
  - in the sense that the distance between every two points is equal to the length of the shortest path connecting these points.
- Let us describe this requirement in precise terms.
- Let M be a metric space with metric $d(a, b)$.
- By a *path* s from a point $a \in M$ to a point $b \in M$, we mean continuous mapping $s : [0, 1] \mapsto M$.

7. What we want (cont-d)

- We say that a path has *length* L if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that:
  - if we have a sequence $t_0 = 0 < t_1 < \ldots < t_{n-1} < t_n = 1$ for which $t_{i+1} - t_i \leq \delta$ for all i,
  - then $\left| \sum\limits_{i=1}^{n-1} d(t_i, t_{i+1}) - L \right| \leq \varepsilon$.
- We say that a metric $d(a, b)$ is *regular* if every two points $a, b \in M$ can be connected by a path of length $d(a, b)$.

8. Main Result

- **Proposition.**
  - *Every scale-invariant regular metric on the set of all positive real numbers has the form* $d(a,b)=k\cdot|\ln(b/a)|$ *for some* $k>0$.
  - *For every* $k>0$, *the metric* $d(a,b)=k\cdot|\ln(b/a)|$ *is regular and scale-invariant.*
- This result justifies the use of logarithmic metric.

9. Proof of the Main Result

- It is easy to see that the metric $d(a,b)=k\cdot|\ln(b/a)|$ is regular and scale-invariant.
- Let us prove that, vice versa, every scale-invariant regular metric on the set of all positive numbers has this form.
- Let us first note that on the shortest path, each point occurs only once.
- Indeed, otherwise:
  - if we had a point c repeated twice,
  - we could cut out the part of the path that connect the first and second occurrences of this point, and
  - thus get an even shorter path.

10. Proof of the Main Result (cont-d)

- On the set of all positive real numbers:
  - the only path between two points $a < b$ that does not contain repetitions
  - is a continuous monotonic mapping of the interval $[0, 1]$ into the interval $[a, b]$.
- So, this is the shortest path.
- On each shortest path between the point a and c for which $a < c$, for each intermediate point b, we have $d(a, c) = d(a, b) + d(b, c)$.
- Indeed, the sequence t_i contains a point t_j close to b.
- Each sum $\sum d(t_i, t_{i+1})$ is close to the sum of two subsums – before this point and after this point.

11. Proof of the Main Result (cont-d)

- When $\varepsilon \to 0$:
  - the first subsum – corresponding to the shortest path from a to b – tends to $d(a,b)$, and
  - the second subsum tends to $d(b,c)$.
- Thus, in the limit, for all $a < b < c$, we indeed have

$$d(a,c) = d(a,b) + d(b,c).$$

- By scale-invariance, for $\lambda = 1/a$, we have $d(a,b) = d(1, b/a)$.
- For all $x > 1$, let us denote $f(x) \stackrel{\text{def}}{=} d(1,x)$.
- In these terms, for $a < b$, we have $d(a,b) = f(b/a)$ and thus, the above equality takes the form $f(c/a) = f(b/a) + f(c/b)$.
- In particular, for each $x \ge 1$ and $y \ge 1$, we can take $a = 1$, $b = x$, and $c = x \cdot y$, and conclude that $f(x \cdot y) = f(x) + f(y)$.

12. Proof of the Main Result (cont-d)

- It is known that the only non-negative non-zero solutions to this functional equation are $f(x) = k \cdot \ln(x)$ for some $k > 0$.
- Thus, indeed, for $a < b$, we have $d(a, b) = f(b/a) = k \cdot \ln(b/a)$.
- Since $d(a, b) = d(b, a)$, for $a > b$, we get $d(a, b) = d(b, a) = k \cdot \ln(a/b)$, i.e., exactly $d(a, b) = k \cdot |\ln(b/a)|$.
- The proposition is proven.

13. Related Result

- So far, we considered situations in which we can select different measuring units.
- For some quantities, we can select different starting points.
- For example, when we measure time, we can start from any moment of time.
- For such quantities:
  - if we select a new starting point which is a_0 moments earlier,
  - then the numerical value corresponding to the same moment of time is shifted from a to $a \mapsto a' = a + a_0$.
- In such cases, it is reasonable to consider shift-invariant metrics.
- We say that a metric $d(a, b)$ on the set of all real numbers is *shift-invariant* if

$$d(a + a_0, b + a_0) = d(a, b) \text{ for all } a, b, \text{ and } a_0.$$

14. Related Result (cont-d)

- **Proposition.**
  - *Every shift-invariant regular metric on the set of all real numbers has the form* $d(a, b) = k \cdot |a - b|$ *for some* $k > 0$.
  - *For every* $k > 0$, *the metric* $d(a, b) = k \cdot |a - b|$ *is regular and shift-invariant.*
- So, in this case, we get the usual description of the absolute approximation error.

15. Proof of the Related Result

- It is easy to see that the metric $d(a,b) = k \cdot |a - b|$ is regular and shift-invariant.
- Let us prove that, vice versa, every shift-invariant regular metric on the set of all real numbers has this form.
- Similarly to the proof of the Main Result, we conclude that for all $a < b < c$, we have $d(a,c) = d(a,b) + d(b,c)$.
- By shift-invariance, for $a_0 = -b$, we have $d(a,b) = d(0, b-a)$.
- For all $x > 0$, let us denote $g(x) \stackrel{\text{def}}{=} d(0,x)$.
- In these terms, for $a < b$, we have $d(a,b) = g(b-a)$ and thus, the equality (4) takes the form $g(c-a) = g(b-a) + g(c-b)$.
- In particular, for each $x \geq 1$ and $y \geq 1$, we can take $a = 0$, $b = x$, and $c = x + y$, and conclude that $g(x+y) = g(x) + g(y)$.
- It is known that the only non-negative non-zero solutions to this functional equation are $g(x) = k \cdot x$ for some $k > 0$.

16. Proof of the Related Result (cont-d)

- Thus, indeed, for $a < b$, we have $d(a,b) = g(b/a) = k \cdot (b - a)$.
- Since $d(a,b) = d(b,a)$, for $a > b$, we get $d(a,b) = d(b,a) = k \cdot (b - a)$, i.e., exactly $d(a,b) = k \cdot |a - b|$.
- The proposition is proven.

17. 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.