Symmetries: A General Approach to Integrated Uncertainty Management

Vladik Kreinovich[1,2], Hung T. Nguyen[2,3], and Songsak Sriboonchitta[2]

[1]University of Texas at El Paso, USA, vladik@utep.edu
[2]Faculty of Economics, Chiang Mai University, Thailand
[3]New Mexico State University, Las Cruces, USA
hunguyen@nmsu.edu

1. Symmetry: a Fundamental Property of the Physical World

- *One of the main objectives of science:* prediction.
- *Basis for prediction:* we observed *similar* situations in the past, and we expect similar outcomes.
- *In mathematical terms:* similarity corresponds to *symmetry*, and similarity of outcomes – to *invariance.*
- *Example:* we dropped the ball, it fall down.
- *Symmetries:* shift, rotation, etc.
- *In modern physics:* theories are usually formulated in terms of symmetries (not diff. equations).
- *Natural idea:* let us use symmetry to describe uncertainty as well.

2. Basic Symmetries: Scaling and Shift

- *Typical situation:* we deal with the numerical values of a physical quantity.
- Numerical values depend on the *measuring unit.*
- *Scaling:* if we use a new unit which is λ times smaller, numerical values are multiplied by λ: $x \to \lambda \cdot x$.
- *Example:* x meters $= 100 \cdot x$ cm.
- *Another possibility:* change the starting point.
- *Shift:* if we use a new starting point which is s units before, then $x \to x + s$ (example: time).
- Together, scaling and shifts form *linear transformations* $x \to a \cdot x + b$.
- *Invariance:* physical formulas should not depend on the choice of a measuring unit or of a starting point.

3. Basic Nonlinear Symmetries

- Sometimes, a system also has *nonlinear* symmetries.
- If a system is invariant under f and g, then:
  - it is invariant under their composition $f \circ g$, and
  - it is invariant under the inverse transformation f^{-1}.
- In mathematical terms, this means that symmetries form a *group*.
- In practice, at any given moment of time, we can only store and describe finitely many parameters.
- Thus, it is reasonable to restrict ourselves to *finite-dimensional* groups.
- *Question* (N. Wiener): describe all finite-dimensional groups that contain all linear transformations.
- *Answer* (for real numbers): all elements of this group are fractionally-linear $x \to (a \cdot x + b)/(c \cdot x + d)$.

4. Symmetries Explain the Basic Formulas of Different Uncertainty Formalisms: Neural Networks

- *What needs explaining:* formula for the *activation function* $f(x) = 1/(1 + e^{-x})$.
- A change in the input starting point: $x \to x + s$.
- *Reasonable requirement:* the new output $f(x+s)$ equivalent to the $f(x)$ mod. appropriate transformation.
- *Reminder:* all appropriate transformations are fractionally linear.
- *Conclusion:* $f(x + s) = \dfrac{a(s) \cdot f(x) + b(s)}{c(s) \cdot f(x) + d(s)}$.
- Differentiating both sides by s and equating s to 0, we get a differential equation for $f(x)$.
- Its known solution is the above activation function – which can thus be explained by symmetries.

5. Symmetries Explain the Basic Formulas of Different Uncertainty Formalisms: Fuzzy Logic

- *Main quantity:* certainty degree $a = d(S)$.
- One way to define $d(S)$ is by polling n experts and taking the fraction $a = m/n$ of those who believe in S.
- To make this estimate more accurate, we can go beyond top experts and ask n' other experts as well.
- In the presence of top experts, other experts may
  - either remain shyly silent
  - or shyly confirm the majority's opinion.
- In the first case, the degree reduces from $a = m/n$ to $a' = m/(n+n')$, i.e., to $a' = \lambda \cdot a$, where $\lambda = n/(n+n')$.
- In the second case, a changes to $a' = (m+m')/(n+m')$ – a linear transformation.
- In general, we get all linear transformations.

6. Symmetries Explain the Basic Formulas of Different Uncertainty Formalisms: Fuzzy Logic (cont-d)

- *Fact:* we can describe the degree of certainty $d(S)$ in a statement S:
  - either by its own degree of certainty,
  - or by a degree of certainty in, say, $S \,\&\, S_0$ for some statement S_0.
- *Reasonable to require:* the corresponding transformation $d(S) \to d(S \,\&\, S_0)$ is appropriate.
- *Conclusion:* the transformation $d(S) \to d(S \,\&\, S_0)$ is fractionally linear.
- *Results:* this conclusion explains many empirically efficient t-norms and t-conorms.
- *Comment:* many other uncertainty-related formulas can also be similarly explained.

7. What Else We Do in This Paper

- *We have shown:* basic uncertainty-related formulas can be explained in terms of symmetries.
- *We show:* many other aspects of uncertainty can be explained in terms of symmetries:
  - heuristic and semi-heuristic approaches can be justified by appropriate natural symmetries, and
  - symmetries can help in designing optimal algorithms.

8. First Example: Practical Need for Uncertainty Propagation

- *Practical problem:* we are often interested in the quantity y which is difficult to measure directly.
- *Solution:*
  - estimate easier-to-measure quantities $x_1, \ldots, x_n$ which are related to y by a known algorithm $y = f(x_1, \ldots, x_n)$;
  - compute $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ based on the estimates $\widetilde{x}_i$.
- *Fact:* estimates are never absolutely accurate: $\widetilde{x}_i \neq x_i$.
- *Consequence:* the estimate $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ is different from the actual value $y = f(x_1, \ldots, x_n)$.
- *Problem:* estimate the uncertainty $\Delta y \stackrel{\text{def}}{=} \widetilde{y} - y$.

9. Propagation of Probabilistic Uncertainty

- *Fact:* often, we know the probabilities of different values of Δx_i.
- *Example:* Δx_i are independent normally distributed with mean 0 and known st. dev. σ_i.
- *Monte-Carlo approach:*
  - For $k = 1, \ldots, N$ times, we:
    * simulate the values $\Delta x_i^{(k)}$ according to the known probability distributions for x_i;
    * find $x_i^{(k)} = \widetilde{x}_i - \Delta x_i^{(k)}$;
    * find $y^{(k)} = f(x_1^{(k)}, \ldots, x_n^{(k)})$;
    * estimate $\Delta y^{(k)} = y^{(k)} - \widetilde{y}$.
  - Based on the sample $\Delta y^{(1)}, \ldots, \Delta y^{(N)}$, we estimate the statistical characteristics of Δy.

10. Propagation of Interval Uncertainty

- *In practice:* we often do not know the probabilities.
- *What we know:* the upper bounds Δ_i on the measurement errors Δx_i: $|\Delta x_i| \le \Delta_i$.
- *Enter intervals:* once we know $\widetilde{x}_i$, we conclude that the actual (unknown) x_i is in the interval

$$\mathbf{x}_i = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i].$$

- *Problem:* find the range $\mathbf{y} = [\underline{y}, \overline{y}]$ of possible values of y when $x_i \in \mathbf{x}_i$:

$$\mathbf{y} = f(\mathbf{x}_1, \ldots, \mathbf{x}_n) \stackrel{\text{def}}{=}$$

$$\{f(x_1, \ldots, x_n) \,|\, x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\}.$$

- *Fact:* this *interval computation* problem is, in general, NP-hard.

11. Propagation of Fuzzy Uncertainty

- In many practical situations, the estimates $\widetilde{x}_i$ come from experts.
- Experts often describe the inaccuracy of their estimates by natural language terms like "approximately 0.1".
- A natural way to formalize such terms is to use membership functions $\mu_i(x_i)$.
- For each α, we can determine the α-*cut*

$$\mathbf{x}_i(\alpha) = \{x_i \,|\, \mu_i(x_i) \geq \alpha\}.$$

- Natural idea: find $\mu(y)$ for which, for each α,

$$\mathbf{y}(\alpha) = f(\mathbf{x}_1(\alpha), \ldots, \mathbf{x}_1(\alpha)).$$

- So, the problem of propagating fuzzy uncertainty can be reduced to several interval propagation problems.

12. Need for Faster Algorithms for Uncertainty Propagation

- For propagating probabilistic uncertainty, there are efficient algorithms such as Monte-Carlo simulations.
- In contrast, the problems of propagating interval and fuzzy uncertainty are computationally difficult.
- It is therefore desirable to design faster algorithms for propagating interval and fuzzy uncertainty.
- The problem of propagating fuzzy uncertainty can be reduced to the interval case.
- Hence, we mainly concentrate on faster algorithms for propagating interval uncertainty.

13. Linearization

- In many practical situations, the errors Δx_i are small, so we can ignore quadratic terms:

$$\Delta y = \widetilde{y} - y = f(\widetilde{x}_1, \ldots, \widetilde{x}_n) - f(x_1, \ldots, x_n) =$$
$$f(\widetilde{x}_1, \ldots, \widetilde{x}_n) - f(\widetilde{x}_1 - \Delta x_1, \ldots, \widetilde{x}_n - \Delta x_n) \approx$$
$$c_1 \cdot \Delta x_1 + \ldots + c_n \cdot \Delta x_n,$$

  where $c_i \stackrel{\text{def}}{=} \dfrac{\partial f}{\partial x_i}(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.

- For a linear function, the largest Δy is obtained when each term $c_i \cdot \Delta x_i$ is the largest:

$$\Delta = |c_1| \cdot \Delta_1 + \ldots + |c_n| \cdot \Delta_n.$$

- Due to the linearization assumption, we can estimate each partial derivative c_i as

$$c_i \approx \frac{f(\widetilde{x}_1, \ldots, \widetilde{x}_{i-1}, \widetilde{x}_i + h_i, \widetilde{x}_{i+1}, \ldots, \widetilde{x}_n) - \widetilde{y}}{h_i}.$$

14. Linearization: Algorithm

To compute the range $\mathbf{y}$ of y, we do the following.

- First, we apply the algorithm f to the original estimates $\widetilde{x}_1, \ldots, \widetilde{x}_n$, resulting in the value $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.
- Second, for all i from 1 to n,
  - we compute $f(\widetilde{x}_1, \ldots, \widetilde{x}_{i-1}, \widetilde{x}_i + h_i, \widetilde{x}_{i+1}, \ldots, \widetilde{x}_n)$ for some small h_i and then
  - we compute

$$c_i = \frac{f(\widetilde{x}_1, \ldots, \widetilde{x}_{i-1}, \widetilde{x}_i + h_i, \widetilde{x}_{i+1}, \ldots, \widetilde{x}_n) - \widetilde{y}}{h_i}.$$

- Finally, we compute $\Delta = |c_1| \cdot \Delta_1 + \ldots + |c_n| \cdot \Delta_n$ and the desired range $\mathbf{y} = [\widetilde{y} - \Delta, \widetilde{y} + \Delta]$.
- *Problem:* we need $n+1$ calls to f, and this is often too long.

15. Cauchy Deviate Method: Idea

- For large n, we can further reduce the number of calls to f if we Cauchy distributions, w/pdf

$$\rho(z) = \frac{\Delta}{\pi \cdot (z^2 + \Delta^2)}.$$

- Known property of Cauchy transforms:
  - if $z_1, \ldots, z_n$ are independent Cauchy random variables w/parameters $\Delta_1, \ldots, \Delta_n$,
  - then $z = c_1 \cdot z_1 + \ldots + c_n \cdot z_n$ is also Cauchy distributed, w/parameter

$$\Delta = |c_1| \cdot \Delta_1 + \ldots + |c_n| \cdot \Delta_n.$$

- This is exactly what we need to estimate interval uncertainty!

16. Cauchy Deviate Method: Towards Implementation

- To implement the Cauchy idea, we must answer the following questions:
  - how to simulate the Cauchy distribution; and
  - how to estimate the parameter Δ of this distribution from a finite sample.
- Simulation can be based on the functional transformation of uniformly distributed sample values:
$$\delta_i = \Delta_i \cdot \tan(\pi \cdot (r_i - 0.5)), \text{ where } r_i \sim U([0,1]).$$
- To estimate Δ, we can apply the Maximum Likelihood Method $\rho(\delta^{(1)}) \cdot \rho(\delta^{(2)}) \cdot \ldots \cdot \rho(\delta^{(N)}) \to \max$, i.e., solve
$$\frac{1}{1+\left(\frac{\delta^{(1)}}{\Delta}\right)^2} + \ldots + \frac{1}{1+\left(\frac{\delta^{(N)}}{\Delta}\right)^2} = \frac{N}{2}.$$

17. Cauchy Deviates Method: Algorithm

- Apply f to $\widetilde{x}_i$; we get $\widetilde{y} := f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.
- For $k = 1, 2, \ldots, N$, repeat the following:
  - use the standard RNG to draw $r_i^{(k)} \sim U([0,1])$, $i = 1, 2, \ldots, n$;
  - compute Cauchy distributed values $c_i^{(k)} := \tan(\pi \cdot (r_i^{(k)} - 0.5))$;
  - compute $K := \max_i |c_i^{(k)}|$ and normalized errors $\delta_i^{(k)} := \Delta_i \cdot c_i^{(k)}/K$;
  - compute the simulated "actual values" $x_i^{(k)} := \widetilde{x}_i - \delta_i^{(k)}$;
  - compute simulated errors of indirect measurement: $\delta^{(k)} := K \cdot \left(\widetilde{y} - f\left(x_1^{(k)}, \ldots, x_n^{(k)}\right)\right)$;
- Compute Δ by applying the bisection method to solve the Maximum Likelihood equation.

18. Important Comment

- To avoid confusion, we should emphasize that:
  - in contrast to the Monte-Carlo solution for the probabilistic case,
  - the use of Cauchy distribution in the interval case is a computational trick,
  - it is *not* a truthful simulation of the actual measurement error Δx_i.
- Indeed:
  - we know that the actual value of Δx_i is always inside the interval $[-\Delta_i, \Delta_i]$, but
  - a Cauchy distributed random attains values outside this interval as well.

19. Cauchy Deviate Method: Need for Intuitive Explanation

- *Fact:* the Cauchy deviate method is mathematically valid.
- *Problem:* this method is somewhat counterintuitive:
  - we want to analyze errors which are located *instead* a given interval $[-\Delta, \Delta]$, but
  - this analysis use Cauchy simulated errors which are located *outside* this interval.
- It is therefore desirable to come up with an intuitive explanation for this technique.
- In this talk, we show that such an explanation can be obtained from neural networks.

20. Werbos's Idea: Use Neurons

- *Traditionally:* neural networks are used to simulate a deterministic dependence.
- *Paul Werbos* suggested that the same neural networks can be used to describe stochastic dependencies as well.
- *How:* as one of the inputs, we take a random number $r \sim U([0,1])$.
- *Simplest case:* a single neuron.
- *In this case:* we apply the activation (input-output) function $f(y)$ to the random number r.
- *What we do:* let us analyze the resulting distribution of $f(r)$.
- *Question:* which $f(y)$ should we use?

21. We Must Choose a Family of Functions, Not a Single Function

- *Changing units:* if $f \in F$, then $k \cdot f \in F$.
- *Conclusion:* in mathematical terms, we choose a *family* F of functions f.
- *Changing starting point:* if $f \in F$, then $f + c \in F$.
- *Non-linear changes:* since NN are useful in non-linear case, we consider $f(y) \to g(f(y))$ for non-linear $g \in G$.
- *Natural requirement:* G is closed under composition and depends on finitely many parameters.
- *Result:* any finite-D group G containing all linear f-s has fractional-linear ones.
- *Conclusion:* $F = \{g(f(x)) : g \in G\}$.

22. Which Family is the Best?

- *Optimality criterion* is not necessary numerical:
  - we can choose F with smallest approximation error,
  - among such F, the fastest to compute.
- *General idea:* a partial (pre-)order.
- *Shift-invariance:* if $F > G$, then $T_a(F) > T_a(G)$, where $T_a(F) = \{f(x+a) \mid f \in F\}$.
- *Finality:*
  - if several families are optimal w.r.t. some criterion,
  - we can use this non-uniqueness to select the one with some additional good qualities;
  - in effect, we this change a criterion to a new one in which the optimal family is unique;
  - thus, in the *final* criterion, there is only one optimal family.

23. Main Result

Theorem.

- *Let a F be optimal in the sense of some optimality criterion that is final and shift-invariant.*
- *Then $f \in F$ has the form $a + b \cdot s_0(K \cdot y + l)$ for some a, b, K and l, where $s_0(y)$ is*
  - *either a linear or fractional-linear function,*
  - *or $s_0(y) = \exp(y)$,*
  - *or the logistic function $s_0(y) = 1/(1 + \exp(-y))$,*
  - *or $s_0(y) = \tan(y)$.*

Comments.

- The logistic function is indeed the most popular activation in NN, but others are also used.
- $\tan(r)$ leads to the desired Cauchy distribution.

24. Second Example: Many Practical Situations Eventually Reach Equilibrium

- In *economics*,
  - a situation changes;
  - prices start changing (often fluctuating);
  - eventually, prices reach an equilibrium between supply and demand.
- In *transportation*,
  - a new road is built;
  - some traffic moves to this road to avoid congestion on the other roads;
  - this move causes congestion on the new road;
  - as a result, some drivers to go back to their previous routes;
  - etc.

25. It Is Often Desirable to Predict the Corresponding Equilibrium

- For the purposes of the long-term planning, it is desirable to find the corresponding equilibrium.
- *Economic example:* how, in the long run, oil prices will change if we start exploring new oil fields in Alaska?
- *Transportation example:* to what extent the introduction of a new road will relieve the traffic congestion?
- *General objective:* solve the practically important problem of predicting the equilibrium.
- *First step:* describe the equilibrium prediction problem in precise terms.

26. Finding Equilibrium as a Mathematical Problem

- *Non-equilibrium states*: economic example:
  - *situation:* oil price is too high;
  - *result:* profitable to explore difficult-to-extract oil fields;
  - *new result:* the supply of oil increases, and prices drop.
- *Non-equilibrium states:* transportation example:
  - *situation:* too many cars move to a new road;
  - *results:* the new road becomes congested;
  - *new result:* drivers abandon the new road.
- *General description:* given a current state x, we can determine the state $f(x)$ at the next moment of time.
- *Equilibrium:* a state that does not change $f(x) = x$ (i.e., a fixed point).

27. Towards an Optimal Algorithm for Computing Fixed Points

- *Idea:* when iterations $x_{k+1} = f(x_k)$ do not converge, $x_{k+1} = x_k + \alpha \cdot (f(x_k) - x_k) = (1 - \alpha_k) \cdot x_k + \alpha_k \cdot f(x_k)$.
- *Question:* which choice of α_k is best?
- *Idea:* this is a discrete approximation to a continuous-time system $\dfrac{dx}{dt} = \alpha(t) \cdot (f(x) - x)$.
- *Scale invariance:* the system should not change if we use a different discretization, i.e., re-scale t to $t' = t/\lambda$:
$$\frac{dx}{dt'} = (\lambda \cdot \alpha(\lambda \cdot t')) \cdot (f(x) - x).$$
- *Conclusion:* $\lambda \cdot \alpha(\lambda \cdot t') = a(t')$, so for $\lambda = 1/t'$, we get $\alpha(t') = \dfrac{c}{t'}$ for some c.
- *Fact:* this is indeed empirically best.

28. Acknowledgments

This work was supported in part:

- by NSF grant HRD-0734825 and
- by Grant 1 T36 GM078000-01 from the National Institutes of Health.