Algebraic Approach to Data Processing: Techniques and Applications

Julio C. Urenda

Department of Computer Science
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
jcurenda@utep.edu

1. Introduction

- How do we gain knowledge about the world?
- For example, how did we learn that if we drop an object, it will fall with the acceleration of 9.81 m/sec^2?
- Well, the scientists dropped an object once, and observed this fall.
- Then they moved to a different location and repeated the same experiment – and got the exact same result.
- Then they turned by an angle – and also got the same result.
- After several such experiments, they concluded that:
  - the result of this experiment
  - does not change if we move to a different location or turn by some angle.
- In other words, they concluded that this process is *invariant* with respect to shifts and rotations.

2. Introduction (cont-d)

- In other cases, other transformations are appropriate.
- For example:
  - the whole idea of a wind tunnel – in which smaller-size airplane models used to be tested
  - is that the corresponding processes do not change if we re-scale the objects.
- In electrodynamics, all interactions remain the same if we replace all positive charges with negative ones, and vice versa.
- According to Special Relativity theory, processes do not change if everything starts moving with a constant speed, etc.
- In all these cases, we have some transformations with respect to which processes are invariant.
- In mathematics, studying the classes of such transformation is classified as part of *algebra*.

3. Introduction (cont-d)

- From this viewpoint, algebraic approach to designing (and optimizing) data processing algorithms is a very natural idea.
- In this thesis, we describe applications of this idea:
  - to various aspects of algorithmics,
  - to dynamic systems,
  - to physics,
  - to engineering,
  - to medicine,
  - to economics,
  - to social sciences,
  - to education, and
  - to mathematics.
- In this talk, I will give some examples of these applications.

Part I

Why Squashing Functions in Multi-Layer Neural Networks

4. A Short Introduction

- In their successful applications, deep neural networks use a non-linear transformation $s(z) = \max(0, z)$.
- It is called a *rectified linear* activation function.
- Sometimes, more general transformations – called *squashing functions* – lead to even better results.
- In this talk, we provide a theoretical explanation for this empirical fact.
- To provide this explanation, let us first briefly recall:
  - why we need machine learning in the first place,
  - what are deep neural networks, and
  - what activation functions these neural networks use.

5. Machine Learning Is Needed

- For some simple systems, we know the equations that describe the system's dynamics.
- These equations may be approximate, but they are often good enough.
- With more complex systems (such as systems of systems), this is often no longer the case.
- Even when we have a good approximate model for each subsystem, the corresponding inaccuracies add up.
- So, the resulting model of the whole system is too inaccurate to be useful.
- We also need to use the records of the actual system's behavior when making predictions.
- Using the previous behavior to predict the future is called *machine learning*.

6. Deep Learning

- The most efficient machine learning technique is *deep learning*: the use of multi-layer neural networks.
- In general, on a layer of a neural network, we transform signals $x_1, \ldots, x_n$ into a new signal $y = s\left(\sum\limits_{i=1}^{n} w_i \cdot x_i + w_0\right)$.
- The coefficient w_i (called *weights*) are to be determined during training.
- $s(z)$ is a non-linear function called *activation function*.
- Most multi-layer neural networks use $s(z) = \max(z, 0)$ known as *rectified linear* function.

7. Shall We Go Beyond Rectified Linear?

- Preliminary analysis shows that for some applications:
  - it is more advantageous to use different activation functions for different neurons;
  - specifically, this was shown for a special family of *squashing* activation functions

$$S^{(\beta)}_{a,\lambda}(z)=\frac{1}{\lambda\cdot\beta}\cdot\ln\frac{1+\exp(\beta\cdot z-(a-\lambda/2))}{1+\exp(\beta\cdot z-(a+\lambda/2))};$$

  - this family contains rectified linear neurons as a particular case.
- We explain their empirical success of squashing functions by showing that:
  - their formulas
  - follow from reasonably natural symmetries.

8. How This Talk Is Structured

- First, we recall the main ideas of symmetries and invariance.
- Then, we recall how these ideas can be used to explain the efficiency of the sigmoid activation function

$$s_0(z) = \frac{1}{1 + \exp(-z)}.$$

- This function is used in the traditional 3-layer neural networks.
- Finally, we use this information to explain the efficiency of squashing activation functions.

9. Which Transformations Are Natural?

- From the mathematical viewpoint, we can apply any non-linear transformation.
- However, some of these transformations are purely mathematical, with no clear physical interpretation.
- Other transformation are *natural* in the sense that they have physical meaning.
- What are natural transformations?

10. Numerical Values Change When We Change a Measuring Unit And/Or Starting Point

- In data processing, we deal with numerical values of different physical quantities.
- Computers just treat these values as numbers.
- However, from the physical viewpoint, the numerical values are not absolute; they change:
  - if we change the measuring unit and/or
  - the starting point for measuring the corresponding quantity.
- The corresponding changes in numerical values are clearly physically meaningful, i.e., natural.
- For example, we can measure a person's height in meters or in centimeters.

11. Numerical Values Change (cont-d)

- The same height of 1.7 m, when described in centimeters, becomes 170 cm.
- In general, if we replace the original measuring unit with a new unit which is λ times smaller, then:
  - instead of the original numerical value x,
  - we get a new numerical value $\lambda \cdot x$ – while the actual quantity remains the same.
- Such a transformation $x \to \lambda \cdot x$ is known as *scaling*.
- For some quantities, e.g., for time or temperature, the numerical value also depends on the starting point.
- For example, we can measure the time from the moment when the talk started.
- Alternatively, we can use the usual calendar time, in which Year 0 is the starting point.

12. Numerical Values Change (cont-d)

- In general, if we replace the original starting point with the new one which is x_0 units earlier, than:
  - each original numerical value x
  - is replaced by a new numerical value $x + x_0$.
- Such a transformation $x \to x + x_0$ is known as *shift*.
- In general, if we change both the measuring unit and the starting point, we get a linear transformation:

$$x \to \lambda \cdot x + x_0.$$

- A usual example of such a transformation is a transition from Celsius to Fahrenheit temperature scales:

$$t_F = 1.8 \cdot t_C + 32.$$

13. Invariance

- Changing the measuring unit and/or starting point:
  - changes the numerical values but
  - does not change the actual quantity.
- It is therefore reasonable to require that physical equations do not change if we simply:
  - change the measuring unit and/or
  - change the starting point.
- Of course, to preserve the physical equations:
  - if we change the measuring unit and/or starting point for one quantity,
  - we may need to change the measuring units and/or starting points for other quantities as well.
- For example, there is a well-known relation $d = v \cdot t$ between distance d, velocity v, and time t.

14. Invariance (cont-d)

- If we change the measuring units for measuring distance and time:
  - this formula remains valid –
  - but only if we accordingly change the units for velocity.
- For example:
  - if we replace kilometers with meters and hours with seconds,
  - then, to preserve this formula, we also need to change the unit for velocity from km/h to m/sec.

15. Natural Transformations Beyond Linear Ones

- In some cases, the relation between different scales is non-linear.
- For example, we can measure the earthquake energy:
  - in Joules (i.e., in the usual scale) or
  - in a logarithmic (Richter) scale.
- Which nonlinear transformation are natural?
- First, as we have argued, all linear transformations are natural.
- Second:
  - if we have a natural transformation $f(x)$ from scale A to another B,
  - then the inverse transformation $f^{-1}(x)$ from scale B to scale A should also be natural.

16. Natural Transformations (cont-d)

- Third:
  - if $f(x)$ and $g(x)$ are natural scale transformation,
  - then we can apply first $g(x)$ to get $y = g(x)$ and then f to get $f(y) = f(g(x))$.
- Thus, the composition $f(g(x))$ of two natural transformations should also be natural.
- The class of transformations that satisfies the 2nd and 3rd properties is called a *transformation group*.
- We also need to take into account that in a computer:
  - at any given moment of time,
  - we can only store the values of finitely many parameters.
- Thus, the transformations should be determined by a finite number of parameters.

17. Natural Transformations (cont-d)

- The smallest number of parameters needed to describe a family is known as the *dimension* of this family.
- E.g., that we need 3 coordinates to describe any point in space means that the physical space is 3-dimensional.
- In these terms, the transformation group T must be finite-dimensional.

18. Let Us Describe All Natural Transformations

- Interestingly, the above requirements uniquely determine the class of all possible natural transformation.
- This result can be traced back to Norbert Wiener, the father of cybernetics.
- In his seminal book *Cybernetics*, he noticed that:
  - when we approach an object form afar,
  - our perception of this object goes through several distinct phases.
- First, we see a blob; this means that:
  - at a large distance,
  - we cannot distinguish between images obtained each other by all possible continuous transformations.
- This phase corresponds to the group of all possible continuous transformations.

19. All Natural Transformations (cont-d)

- As we get closer, we start distinguishing angular parts from smooth parts, but still cannot compare sizes.
- This corresponds to the group of all projective transformations.
- After that, we become able to detect parallel lines.
- This corresponds to the group of all transformations that preserve parallel lines.
- These are linear (= affine) transformations.
- When we get even closer, we become able to detect the shapes, sizes, etc.

20. All Natural Transformations (cont-d)

- Wiener argued that there are no other transformation groups – since:
  - if there were other transformation groups,
  - after billions years of evolution, we would use them.
- In precise terms, he conjectured that:
  - the only finite-dimensional transformation group that contain all linear transformations
  - is the groups of all projective transformations.
- This conjecture was later proven.
- For transformations of the real line, projective transformations are simply fractional-linear transformations

$$f(x) = \frac{a \cdot x + b}{c \cdot x + d}.$$

- So, natural transformations are fractional-linear ones.

21. Traditional Neural Networks (NN)

- Let us recall why traditional neural networks appeared in the first place.
- The main reason, in our opinion, was that computers were too slow.
- A natural way to speed up computations is to make several processors work in parallel.
- Then, each processor only handles a simple task, not requiring too much computation time.
- For processing data, the simplest possible functions to compute are linear functions.

22. Traditional Neural Networks (cont-d)

- However, we cannot only use linear functions – because then:
  - no matter how many linear transformations we apply one after another,
  - we will only get linear functions, and many real-life dependencies are nonlinear.
- So, we need to supplement linear computations with some nonlinear ones.
- In general, the fewer inputs, the faster the computations.
- Thus, the fastest to compute are functions with one input, i.e., functions of one variable.

23. Traditional Neural Networks (cont-d)

- So, we end up with a parallel computational device that has:
  - linear processing units (L) and
  - nonlinear processing units (NL) that compute functions of one variable.
- First, the input signals come to a layer of such devices; we will call such a layer a *d-layer*; d for *d*evice.
- Then, the results of this d-layer go to another d-layer, etc.
- The fewer d-layers we have, the faster the computations.

24. How Many d-Layers Do We Need?

- It can be proven that:
  - 1-d-layer schemes (L or NL) are not sufficient to approximate any possible dependence, and
  - 2-d-layer schemes (L-NL, linear layer followed by non-linear layer, or NL-L) are also not enough.
- Thus, we need at least 3-d-layer networks – and 3-d-layer networks can be proven to be sufficient.
- In a 3-d-layer network:
  - we cannot have two linear layers or two nonlinear d-layers following each other,
  - that would be equivalent to having one d-layer since, e.g., a composition of two L functions is also L.
- So, our only options are L-NL-L and NL-L-NL.

25. How Many d-Layers Do We Need (cont-d)

- Since linear transformations are faster to compute, the fastest scheme is L-NL-L.
- In this scheme:
  - first, each neuron k in the L d-layer combines the inputs into a linear combination
$$z_k = \sum_{i=1}^{n} w_{ki} \cdot x_i + w_{k0};$$
  - then, in the next d-layer, each such signal is transformed into $y_k = s_k(z_k)$ for some non-linear f-n;
  - finally, in the last linear d-layer, we form a linear combination of the values y_k: $y = \sum\limits_{k=1}^{K} W_k \cdot y_k + W_0$.

26. How Many d-Layers Do We Need (cont-d)

- The resulting transformation takes the form

$$y = \sum_{k=1}^{K} W_k \cdot s_k\left(\sum_{i=1}^{n} w_{ki} \cdot x_i + w_{k0}\right) + W_0.$$

- Usually, we use the same function $s(z)$ for all transformations.
- This is indeed the usual formula of the traditional neural network.

27. Traditional NN Mostly Used Sigmoid

- Originally, the sigmoid function was selected because:
  - it provides a reasonable approximation to
  - how biological neurons process their inputs.
- Several other nonlinear activation functions have been tried.
- However, in most cases, the sigmoid $s_0(z)$ leads to the best approximation results.
- A partial explanation for this empirical success is that:
  - neural networks using sigmoid activation function $s_0(z)$ have proven to be universal approximators;
  - i.e., the corresponding neural networks can approximate any continuous function.
- However, many other non-linear activation functions have the same universal approximation property.

28. So, Why Sigmoid?

- We have mentioned that the values of physical quantities change when we:
  - change the starting point,
  - i.e., shift all the data points by the same constant x_0.
- At first glance, it may seem that this does not apply to neural data processing, since usually:
  - before we apply a neural network,
  - we *normalize* the data, i.e., transform all the input values into the some fixed interval (e.g., $[0, 1]$).
- This normalization is based on all the values of the corresponding quantity that have been observed so far.
- The smallest of these values corresponds to 0 and the largest to 1.

29. Why Sigmoid (cont-d)

- However, as we will show, shift still makes sense even for the normalized data.
- Indeed, in real life, signals come with noise, in particular, with background noise.
- Often, a significant part of this noise is a constant which is added to all the measured signals.
- This constant noise component is, in general, different for different situations.
- We can try to get rid of this constant noise component by subtracting the corresponding constant.
- So, we replace:
  - each original numerical value x_i
  - with a corrected value $x_i - n_i$.

30. Why Sigmoid (cont-d)

- After this correction, instead of the original value z_k, we get a corrected value

$$z'_k = \sum_{i=1}^{n} w_{ki} \cdot (x_i - n_i) + w_{k0} = z_k - h'_k.$$

- Here, we denoted $h'_k \stackrel{\text{def}}{=} \sum\limits_{i=1}^{n} w_{ki} \cdot n_i$.

- The trouble is that we do not know the exact values of these constants n_i.

- So, depending on our estimates, we may subtract different values n_i and thus, different values h'_k:
  - if we change from one value h'_k to another one h''_k,
  - then the resulting value of z_k is shifted by the difference $h_k \stackrel{\text{def}}{=} h'_k - h''_k$: $z''_k = z'_k + h_k$.

31. Why Sigmoid (cont-d)

- This is exactly the same formula as for the shift corresponding to the change in the starting point.

- Since we do not know what shift is the best, all shifts within a certain range are equally possible.

- It is therefore reasonable to require that the formula $y = s(z)$ for the nonlinear activation function:
  - should work for all possible shifts,
  - i.e., this formula should be, in this sense, *shift-invariant.*

- In other words:
  - if we start with the formula $y = s(z)$ and we shift from z to $z' = z + h$,
  - then we should have the same relation $y' = s(z')$ for an appropriately transformed $y' = f(y)$.

32. Why Sigmoid (cont-d)

- For different shifts h, we will have, in general, different natural transformations $f(y)$.
- We have mentioned that all natural transformations $f(y)$ are fractionally linear.
- Thus, for each h, $y' = s(z+h)$ should be fractional-linear in $y = s(z)$:

$$s(z+h) = \frac{a(h) \cdot s(z) + b(h)}{c(h) \cdot s(z) + d(h)}.$$

- It turns out that this implies the sigmoid $s_0(z)$.

33. Why Sigmoid: Derivation

- For $h = 0$, we should have $s(z+h) = s(z)$, thus, we should have $d(0) \neq 0$.
- It is reasonable to require that the function $d(h)$ is continuous.
- In this case, $d(h)$ is different from 0 for all small h.
- Then, we can divide both numerator and denominator of the above formula by $d(h)$ and get a simpler formula:

$$s(z+h) = \frac{A(h) \cdot s(z) + B(h)}{C(h) \cdot s(z) + 1}, \text{ where } A(h) = a(h)/d(h), \ldots$$

- For $h = 0$, we have $s(z+h) = s(z)$, so $A(h) = 1$ and $B(h) = C(h) = 0$.
- It is also reasonable to require that the activation function $s(z)$ be defined and smooth for all z.

34. Why Sigmoid: Derivation (cont-d)

- Indeed, on each interval, every continuous function:
  - can be approximated, with any desired accuracy,
  - by a smooth one – even by a polynomial.
- So, from the practical viewpoint, it is sufficient to only consider smooth activation functions.
- Multiplying both sides of the above formula by the denominator, we get:

$$s(z+h) = A(h)\cdot s(z) + B(h) - C(h)\cdot s(z+h)\cdot s(z).$$

- Let us take three different values z_i.
- Then, for each h, we get 3 linear equations for three unknown $A(h)$, $B(h)$, and $C(h)$:

$$s(z_i+h) = A(h)\cdot s(z_i) + B(h) - C(h)\cdot s(z_i+h)\cdot s(z_i),\ \ i = 1, 2, 3.$$

35. Why Sigmoid: Derivation (cont-d)

- Due to Cramer's rule, the solution to this system is:
  - a ratio of two determinants,
  - i.e., a ration of two polynomials of the coefficients.
- Thus, $A(h)$, $B(h)$, and $C(h)$ are smooth functions of the values $s(z_i + h)$.
- Since the function $s(z)$ is smooth, we conclude that all three functions $A(h)$, $B(h)$, and $C(h)$ are also smooth.
- Thus, we can differentiate both sides of the above equation by h and get

$$s'(z+h) = \frac{N(h)}{(C(h)\cdot s(z)+1)^2}, \quad \text{where}$$

$$N(h) \overset{\text{def}}{=} (A'(h)\cdot s(z) + B'(h))\cdot (C(h)\cdot s(z)+1) -$$

$$(A(h)\cdot s(z) + B(h))\cdot (C'(h)\cdot s(z)).$$

36. Why Sigmoid: Derivation (cont-d)

- In particular, for $h = 0$, taking into account that $A(h) = 1$ and $B(h) = C(h) = 0$, we conclude that
$$s'(z) = a_0 + a_1 \cdot s(z) + a_2 \cdot (s(z))^2, \text{ where } a_0 = B'(0), \ldots$$
- So, $\dfrac{ds}{dz} = a_0 + a_1 \cdot s + a_2 \cdot s^2$ and
$$\frac{ds}{a_0 + a_1 \cdot s + a_2 \cdot s^2} = dz.$$
- We can now integrate both sides of this formula and get an explicit expression of $z(s)$.
- Based on this expression, we can find the explicit formula for the dependence of s on z.

37. Why Sigmoid: Derivation (cont-d)

- The only non-linear dependencies $s(z)$ are:
  - the sigmoid (plus some linear transformations before and after) and
  - the sigmoid's limit case $\exp(z)$.
- So, the sigmoid $s_0(z)$ is the only shift-invariant activation function.
- This explains its efficiency in traditional neural networks.

38. We Need Multi-Layer Neural Networks

- The problem with traditional neural networks is that they waste a lot of bits:
  - for K neurons,
  - any of $K!$ permutations results in exactly the same function.
- To decrease this duplication, we need to decrease the number of neurons K in each layer.
- So, instead of placing all nonlinear neurons in one layer, we place them in several consecutive layers.
- This is one of the main idea behind deep learning.

39. Which Activation Function Should We Use

- In the first nonlinear d-layer, we make sure that:
  - a shift in the input – corresponding to a different estimate of the constant noise component,
  - does not change the processing formula,
  - i.e., that results $s(z+c)$ and $s(z)$ can be obtained from each other by an appropriate transformation.
- We already know that this idea leads to the sigmoid function $s_0(z)$.
- This logic doesn't work if we try to find out what activation function we should use in the *next* NL d-layer.
- Indeed, the input to the 2nd NL d-layer is the output of the 1st NL d-layer.
- This input is *no longer* shift-invariant.

40. Which Activation Function (cont-d)

- This input is invariant with respect to some more *complex* (fractional linear) transformations.
- We know what to do when the input is shift-invariant.
- So a natural idea is to perform some *additional* transformation that will make the results shift-invariant.
- If we do that, then:
  - we will again be able to apply the sigmoid activation function $s_0(z)$,
  - then again the additional transformation, etc.
- These additional transformations should transform generic fractional-linear operations into shift.

41. Which Activation Function (cont-d)

- Thus, the inverse of such a transformation should transform shifts into fractional-linear operations.
- But this is exactly what we analyzed earlier – transformations that transform shifts into fractional-linear.
- We already know the formulas $s(z)$ for these transformations.
- In general, they are formed as follows:
  - first, we apply some linear transformation to the input z, resulting in a linear combination

    $$Z = p \cdot z + q;$$

  - then, we compute $Y = \exp(Z)$; and
  - finally, we apply some fractional-linear transformation to the resulting value Y, getting y.

42. Which Activation Function (cont-d)

- So, to get the inverse transformation, we need to reverse all three steps, starting with the last one:
  - first, we apply a fractional-linear transformation to y, getting Y;
  - then, we compute $Z = \ln(Y)$; and
  - finally, we apply a linear transformation to Z, resulting in z.

43. This Leads Exactly to Squashing Functions

- What happens if we:
  - first apply a sigmoid-type transformation moving us from shifts to tractional-linear operations,
  - and then an inverse-type transformation?
- The last step of the sigmoid transformation and the first step of the inverse are fractional-linear.
- The composition of fractional-linear transformations is fractional-linear.
- So, we can combine these 2 steps into a single step.

44. This Leads to Squashing Functions (cont-d)

- Thus, the resulting combined activation function can thus be described as follows:
  - first, we apply some linear transformation L_1 to the input z, resulting in a linear combination
  $$Z = L_1(z) = p \cdot z + q;$$
  - then, we compute $E = \exp(Z) = \exp(L_1(z))$;
  - then, we apply a fractional-linear transformation F to $E = \exp(Z)$, getting $T = F(E) = F(\exp(L_1(z))$;
  - then, we compute $Y = \ln(T) = \ln(F(\exp(L_1(z)))$;
  - and finally, we apply a linear transformation L_2 to Y, resulting in the final value
  $$y = s(z) = L_2(Y) = L_2(\ln(F(\exp(L_1(z))))).$$

45. This Leads to Squashing Functions (cont-d)

- One can check that these are exactly squashing function!
- Thus, squashing functions can indeed be naturally explained by the invariance requirements.

46. Example

- Let us provide a family of squashing functions that tend to the rectified linear activation function $\max(z, 0)$.

- For this purpose, let us take:

  - $L_1(z) = k \cdot z$, with $k > 0$, so that
  $$E = \exp(L_1(z)) = \exp(k \cdot z);$$
  - $F(E) = 1 + E$, so that $T = F(E) = \exp(k \cdot z) + 1$ and $Y = \ln(T) = \ln(\exp(k \cdot z) + 1)$; and
  - $L_2(Y) = \dfrac{1}{k} \cdot Y$, so that the resulting activation function takes the form $s(z) = \dfrac{1}{k} \cdot \ln(\exp(k \cdot z) + 1)$.

- Let us show that this expression tends to the rectified linear activation function when $k \to \infty$.

- When $z < 0$, then $\exp(k \cdot z) \to 0$, so $\exp(k \cdot z) + 1 \to 1$, $\ln(\exp(k \cdot z) + 1) \to 0$ and so $s(z) \to 0$.

47. Example (cont-d)

- On the other hand, when $z > 0$, then

$$\exp(k \cdot z) + 1 = \exp(k \cdot z) \cdot (1 + \exp(-k \cdot z)).$$

- Thus, $\ln(\exp(k \cdot z) + 1) = k \cdot z + \ln(1 + \exp(-k \cdot z))$ and

$$s(z) = \frac{1}{k} \cdot \ln(\exp(k \cdot z) + 1) = z + \frac{1}{k} \cdot \ln(1 + \exp(-k \cdot z)).$$

- When $k \to \infty$, we have $\exp(-k \cdot z) \to 0$, hence

$$1 + \exp(-k \cdot z) \to 1, \quad \ln(1 + \exp(-k \cdot z)) \to 0.$$

- So $\dfrac{1}{k} \cdot \ln(1 + \exp(-k \cdot z)) \to 0$ and indeed $s(z) \to z$.

Part II
Natural Invariance Explains Empirical Success of Specific Membership Functions, Hedge Operations, and Negation Operations

48. Fuzzy Techniques: A Brief Reminder

- In many applications, we have knowledge formulated:
  - in terms of imprecise ("fuzzy") terms from natural language,
  - like "small", "somewhat small", etc.
- To translate this knowledge into computer-understandable form, Lotfi Zadeh proposes *fuzzy techniques*.
- According to these techniques, each imprecise property like "small" can be described by assigning:
  - to each value x of the corresponding quantity,
  - a degree $\mu(x)$ to which, according to the expert, this property is true.

49. Fuzzy Techniques (cont-d)

- These degrees are usually selected from the interval $[0, 1]$, so that:
  - 1 corresponds to full confidence,
  - 0 to complete lack of confidence, and
  - values between 0 and 1 describe intermediate degrees of confidence.
- The resulting function $\mu(x)$ is known as a *membership function.*
- In practice, we can only ask finitely many questions to the expert.
- So we only elicit a few values $\mu(x_1)$, $\mu(x_2)$, etc.
- Based on these values, we need to estimate the values $\mu(x)$ for all other values x.

50. Fuzzy Techniques (cont-d)

- For this purpose, usually:
  - we select a family of membership functions – e.g., triangular, trapezoidal, etc. – and
  - we select a function from this family which best fits the known values.
- For terms like "somewhat small", "very small", the situation is more complicated.
- We can add different "hedges" like "somewhat", "very", etc., to each property.
- As a result, we get a large number of possible terms.

51. Fuzzy Techniques (cont-d)

- It is not realistically possible to ask the expert about each such term; instead:
  - practitioners estimate the degree to which, e.g., “somewhat small” is true
  - based on the degree to which “small” is true.
- In other words, with each linguistic hedge, we associate a function h from $[0,1]$ to $[0,1]$ that:
  - transforms the degree to which a property is true
  - into an estimate for the degree to which the hedged property is true.

52. Fuzzy Techniques (cont-d)

- Similarly to the membership functions:
  - we can elicit a few values $h(x_i)$ of the hedge operation from the experts, and
  - then we extrapolate and/or interpolate to get all the other values of $h(x)$.
- Usually, a family of hedge operations is pre-selected.
- Then we select a specific operation from this family which best fits the elicited values $h(x_i)$.

53. Fuzzy Techniques (cont-d)

- Similarly:
  - instead of asking experts for their degrees of confidence in statements like “not small”,
  - we estimate these degrees based on their degrees of confidence in the positive statements.
- The corresponding operation $n(x)$ is known as the *negation operation.*

54. Need to Select Proper Membership Functions, Hedge Operations, And Negation Operations

- Fuzzy techniques have been successfully applied to many application areas.
- However, this does not necessarily mean that every time we try to use fuzzy techniques, we get a success.
- The success (or not) often depends on which membership functions etc. we select:
  - for some selections, we get good results (e.g., good control),
  - for other selections, the results are not so good.
- There is a lot of empirical data about which selections work better.
- In this talk, we provide a general explanation for several of these empirically best selections.

55. Need to Select Proper Functions (cont-d)

- This explanation is based on the natural concepts of invariance.
- For symmetric membership functions that describe properties like "small",
  - for which $\mu(x)=\mu(-x)$ and the degree $\mu(|x|)$ decreases with $|x|$,
  - in many practical situations, the most empirically successful are so-called *distending* functions:

$$\mu(x)=\frac{1}{1+a\cdot |x|^b}.$$

- Among hedge and negation operations, often, the most efficient are fractional linear functions:

$$h(x)=\frac{a+b\cdot x}{1+c\cdot x}.$$

56. Re-Scaling

- The variable x describes the value of some physical quantity, such a distance, height, etc.
- When we process these values, we deal with numbers.
- Numbers depend on the selection of the measuring unit:
  - if we replace the original measuring unit with a new one which is λ times smaller,
  - then all the numerical values will be multiplied by λ: $x \to X = \lambda \cdot x$.
- For example, 2 meters become $2 \cdot 100 = 200$ cm.
- This transformation from one measuring scale to another is known as *re-scaling*.

57. Scale-Invariance: Idea

- In many physical situations, the choice of a measuring unit is rather arbitrary.
- In such situations, all the formulas remain the same no matter what unit we use.
- For example, the formula $y = x^2$ for the area of the square with side x remains valid:
  - if we replace the unit for measuring sides from meters with centimeters,
  - of course, we then need to appropriately change the unit for y, from m^2 to cm^2.

58. Scale-Invariance (cont-d)

- In general, invariance of the formula $y = f(x)$ means that:
  - for each re-scaling $x \to X = \lambda \cdot x$, there exists an appropriate re-scaling $y \to Y$
  - for which the same formula $Y = f(X)$ will be true for the re-scaled variables X and Y.

59. Let Us Apply This Idea to the Membership Function

- It is reasonable to require that:
  - the selection of the best membership functions
  - should also not depend on the choice of the unit for measuring the corresponding quantity x.
- So, it is reasonable to require that for each $\lambda > 0$:
  - there should exist some reasonable transformation $y \to Y = T(y)$ of the degree of confidence
  - for which $y = \mu(x)$ implies $Y = \mu(X)$.

60. So, What Are Reasonable Transformations of the Degree of Confidence?

- One way to measure the degree of confidence is to have a poll:
  - ask N experts how many of them believe that a given value x is, e.g., small,
  - count the number M of whose who believe in this, and
  - take the ratio M/N as the desired degree $y = \mu(x)$.
- As usual with polls, the more people we ask, the more adequately we describe the general opinion.
- So, to get a more accurate estimate for $\mu(x)$, it is reasonable to ask more people.
- When we have a limited number of people to ask, it is reasonable to ask top experts in the field.

61. Reasonable Transformations (cont-d)

- When we start asking more people:
  - we are thus adding people who are less experienced,
  - and who may therefore be somewhat intimidated by the opinions of the top experts.
- This intimidation can be expressed in different ways.
- Some new people may be too shy to express their own opinion, so they will keep quiet; as a result:
  - if we add A people to the original N, we sill still have the same number M of people voting "yes",
  - and the new ratio is $Y = \dfrac{M}{N+A}$.
- Here, $Y = a \cdot y$, where $a \stackrel{\text{def}}{=} \dfrac{N}{N+A}$.
- Some new people will be too shy to think on their own and will vote with the majority.

62. Reasonable Transformations (cont-d)

- So when $M > N/2$, we will have $Y = \dfrac{M+A}{N+A}$.
- Since $M = y \cdot N$, we will have $Y = \dfrac{y \cdot N + A}{N+A} = a \cdot y + b$, where a is the same as before and $b = \dfrac{A}{N+A}$.
- We may also have a situation in which:
  - a certain proportion c of the new people keep quiet while
  - the others vote with the majority.
- In this case, we have $Y = \dfrac{M + (1-c) \cdot A}{N+A} = a \cdot y + b$, where $a = (1-c) \cdot \dfrac{A}{N+A}$.

63. Reasonable Transformations (cont-d)

- In all these cases, we have a linear transformation

$$Y = a \cdot y + b.$$

- So, it seems reasonable to identify reasonable transformations with linear ones.
- We will call the corresponding scale-invariance L-scale-invariance (L for Linear).

64. What Membership Functions We Consider

- We consider symmetric properties, for which
$$\mu(-x) = \mu(x).$$
- So it is sufficient to consider only positive values x.
- We consider properties like "small" for which $\mu(x)$ decreases with x and $\lim\limits_{x \to \infty} \mu(x) = 0$.
- We will call such membership functions s-membership functions (s for small).
- We say that an s-membership function $\mu(x)$ is *L-scale-invariant* if:
  - for every $\lambda > 0$, there exist values $a(\lambda)$ and $b(\lambda)$
  - for which $y = \mu(x)$ implies $Y = \mu(X)$, where
$$X = \lambda \cdot x \text{ and } Y = a(\lambda) \cdot y + b(\lambda).$$

65. What Membership Functions (cont-d)

- Unfortunately, this does not solve our problem:
- **Proposition 1.** *The only L-scale-invariant s-membership functions are constant functions* $\mu(x) = \text{const}$.
- What does this result mean?
- We considered two possible types of reasonable transformations of the degrees of confidence.
- They both turned out to be linear.
- This was not enough.
- So probably there are other reasonable transformations of degrees of confidence.
- How can we describe such transformations?

66. What Membership Functions (cont-d)

- Clearly, if we have a reasonable transformation, then its inverse is also reasonable.
- Also, a composition of two reasonable transformations should be a reasonable transformation too.
- So, in mathematical terms, reasonable transformations should form a *group*.
- This group should be finite-dimensional, i.e.:
  - different transformations should be uniquely determined
  - by a finite number of parameters – since in the computer, we can store only finitely many parameters.

67. What Membership Functions (cont-d)

- We also know that linear transformations are reasonable; so, we are looking for:
  - a finite-dimensional group of transformations from real numbers to real numbers
  - that contains all linear transformations.
- It is known that all such transformations are piece-wise linear: $\mu \to \dfrac{a \cdot \mu + b}{1 + c \cdot \mu}$.
- Thus, we arrive at the following definitions.

68. Definitions and the Main Result

- We say that an s-membership function $\mu(x)$ is *scale-invariant* if:
  - for every $\lambda > 0$, there exist $a(\lambda)$, $b(\lambda)$, and $c(\lambda)$
  - for which $y = \mu(x)$ implies $Y = \mu(X)$, where

$$X = \lambda \cdot x \text{ and } Y = \frac{a(\lambda) \cdot y + b(\lambda)}{1 + c(\lambda) \cdot y}.$$

- **Proposition 2.** *The only scale-invariant s-membership functions are distending membership functions.*
- This result explains the empirical success of distending functions.

69. Which Hedge Operations and Negation Operations Should We Select

- We would like hedging and negation operations $y = h(x)$ to be also invariant, i.e., that:
  - for each natural transformation $X = T(x)$, there should be a transformation $Y = S(y)$
  - for which $y = h(x)$ implies $Y = h(X)$.
- Now we know what are natural transformations of membership degrees – they are fractional-linear functions.
- Let us call this h-scale-invariance.
- **Proposition 3.** *The only h-scale-invariant functions are fractionally linear ones.*
- This result explains the empirical success of fractional-linear hedge operations and negation operations.

70. Proof of Proposition 1

- We will prove this result by contradiction.
- Let us assume that the function $\mu(x)$ is not a constant, and let us derive a contradiction.
- Let us substitute the expressions for X, Y, and $y = \mu(x)$ into the formula $Y = \mu(X)$.
- Then, we conclude that for every x and for every λ, we have $\mu(\lambda \cdot x) = a(\lambda) \cdot \mu(x) + b(\lambda)$.
- It is known that monotonic functions are almost everywhere differentiable; due to the above formula:
  - if a function $\mu(x)$ is differentiable at $x = x_0$,
  - it is also differentiable at any point of the type $\lambda \cdot x_0$ for every $\lambda > 0$,
  - and thus, that it is differentiable for all $x > 0$.

71. Proof of Proposition 1 (cont-d)

- Since the function $\mu(x)$ is not constant, there exist values $x_1 \neq x_2$ for which $\mu(x_1) \neq \mu(x_2)$.
- For these values, the above formula has the form

$$\mu(\lambda \cdot x_1) = a(\lambda) \cdot \mu(x_1) + b(\lambda); \quad \mu(\lambda \cdot x_2) = a(\lambda) \cdot \mu(x_2) + b(\lambda).$$

- Subtracting the two equations, we get

$$\mu(\lambda \cdot x_1) - \mu(\lambda \cdot x_2) = a(\lambda) \cdot (\mu(x_1) - \mu(x_2)), \text{ thus}$$

$$a(\lambda) = \frac{\mu(\lambda \cdot x_1) - \mu(\lambda \cdot x_2)}{\mu(x_1) - \mu(x_2)}.$$

- Since the function $\mu(x)$ is differentiable, we can conclude that the function $a(\lambda)$ is also differentiable.
- Thus, the function $b(\lambda) = \mu(\lambda \cdot x) - a(\lambda) \cdot \mu(x)$ is differentiable too.
- So, all three functions $\mu(x)$, $a(\lambda)$, and $b(\lambda)$ are differentiable.

72. Proof of Proposition 1 (cont-d)

- So, we can differentiate both sides of the equality
$$\mu(\lambda\cdot x)=a(\lambda)\cdot\mu(x)+b(\lambda) \text{ with respect to } \lambda.$$
- If we substitute $\lambda=1$, we get $x\cdot\mu'(x)=A\cdot\mu(x)+B$, where we denoted $A\stackrel{\text{def}}{=}a'(1)$, $B\stackrel{\text{def}}{=}b'(1)$.
- Here, $\mu'(x)$, as usual, indicates the derivative.
- Thus, $x\cdot\dfrac{d\mu}{dx}=A\cdot\mu+B$.
- We cannot have $A=0$ and $B=0$, since then $\mu'(x)=0$ and $\mu(x)$ would be a constant.
- Thus, in general, the expression $A\cdot\mu+B$ is not 0, so
$$\frac{d\mu}{A\cdot\mu+B}=\frac{dx}{x}.$$
- If $A=0$, then integration leads to $\dfrac{1}{B}\cdot\mu(x)=\ln(x)+c$, where c_0 is the integration constant.

73. Proof of Proposition 1 (cont-d)

- Thus, $\mu(x) = B \cdot \ln(x) + B \cdot c_0$.
- This expression has negative values for some x, while all the values $\mu(x)$ are in the interval $[0,1]$.
- So, this case is impossible.
- If $A \neq 0$, then we have $d(A \cdot \mu + B) = A \cdot d\mu$, hence
$$\frac{d(A \cdot \mu + B)}{A \cdot \mu + B} = A \cdot \frac{dx}{x}.$$
- Integration leads to $\ln(A \cdot \mu(x) + B) = A \cdot \ln(x) + c_0$.
- By applying $\exp(z)$ to both sides, we get $A \cdot \mu(x) + B = \exp(c_0) \cdot x^A$, i.e., $\mu(x) = A^{-1} \cdot \exp(c_0) \cdot x^A - B/A$.
- This expression tends to infinity either for $x \to \infty$ (if $A > 0$) or for $x \to 0$ (if $A < 0$).
- In both cases, we get a contradiction with our assumption that $\mu(x)$ is within the interval $[0,1]$. Q.E.D.

74. Proof of Proposition 2

- Let us substitute the expressions for X, Y, and $y = \mu(x)$ into the formula $Y = \mu(X)$.
- Then, we conclude that for every x and for every λ:

$$\mu(\lambda \cdot x) = \frac{a(\lambda) \cdot \mu(x) + b(\lambda)}{1 + c(\lambda) \cdot \mu(x)}.$$

- Similarly to the previous proof, we can conclude that the function $\mu(x)$ is differentiable for all $x > 0$.
- Multiplying both sides of the above equality by the denominator, we conclude that

$$\mu(\lambda \cdot x) + c(\lambda) \cdot \mu(x) \cdot \mu(\lambda \cdot x) = a(\lambda) \cdot \mu(x) + b(\lambda).$$

- So, for three different values x_i, we have the following three equations:

$$\mu(\lambda \cdot x_i) + c(\lambda) \cdot \mu(x_i) \cdot \mu(\lambda \cdot x_i) = a(\lambda) \cdot \mu(x_i) + b(\lambda), \quad i = 1, 2, 3.$$

75. Proof of Proposition 2 (cont-d)

- We thus have a system of three linear equations for three unknowns $a(\lambda)$, $b(\lambda)$, and $c(\lambda)$.
- By Cramer's rule:
  - the solution to such a system
  - is a rational (hence differentiable) function of the coefficients and the right-hand sides.
- So, since $\mu(x)$ is differentiable, we can conclude that $a(\lambda)$, $b(\lambda)$, and $c(\lambda)$ are differentiable.
- All the functions $\mu(x)$, $a(\lambda)$, $b(\lambda)$, and $c(\lambda)$ are differentiable.
- So, we can differentiate both sides of the above formula with respect to λ.
- Let us substitute $\lambda = 1$ and take into account that for $\lambda = 1$, we have $a(1) = 1$ and $b(1) = c(1) = 0$.

76. Proof of Proposition 2 (cont-d)

- Then, we get $x \cdot \dfrac{d\mu}{dx} = A \cdot \mu + B - C \cdot \mu^2$, where A and B are the same as in the previous proof and $C \stackrel{\text{def}}{=} c'(1)$.
- For $x \to \infty$, we have $\mu(x) \to 0$, so $\mu'(x) \to 0$, and thus $B = 0$ and $x \cdot \dfrac{d\mu}{dx} = A \cdot \mu - C \cdot \mu^2$.
- So, $\dfrac{d\mu}{B \cdot \mu - C \cdot \mu^2} = \dfrac{dx}{x}$.
- As we have shown in the previous proof, we cannot have $C = 0$, so $C \neq 0$.
- One can easily see that

$$\frac{1}{\mu - \dfrac{B}{C}} - \frac{1}{\mu} = \frac{\dfrac{B}{C}}{\mu \cdot \left(\mu - \dfrac{B}{C}\right)} = \frac{-B}{B \cdot \mu - C \cdot \mu^2}.$$

77. Proof of Proposition 2 (cont-d)

- Thus, by multiplying the equality $\dfrac{d\mu}{B\cdot\mu - C\cdot\mu^2} = \dfrac{dx}{x}$ by $-B$, we get: $\dfrac{d\mu}{\mu - \dfrac{B}{C}} - \dfrac{d\mu}{\mu} = -B\cdot\dfrac{dx}{x}$.
- Integrating both sides, we get
$$\ln\left(\mu(x) - \frac{B}{C}\right) - \ln(\mu) = -B\cdot\ln(x) + c_0.$$
- By applying $\exp(z)$ to both sides, we get
$$\frac{\mu(x) - \dfrac{B}{C}}{\mu(x)} = C_0\cdot x^{-B}.\text{ so } 1 - \frac{B/C}{\mu} = C_0\cdot x^{-B}.$$
- Hence $\dfrac{B/C}{\mu} = 1 - C_0\cdot x^{-B}$ and $\mu(x) = \dfrac{B/C}{1 - C_0\cdot x^{-B}}$.
- From the condition that $\mu(0) = 1$, we conclude that $B < 0$ and $B/C = 1$.

78. Proof of Proposition 2 (cont-d)

- From $\mu(x) \le 1$, we conclude that $C_0 < 0$.
- So, we get the desired formula $\mu(x) = \dfrac{1}{1 + |C_0| \cdot x^{|B|}}$.
- The proposition is proven.

79. Proof of Proposition 3

- For constant functions the statement is trivial.
- Therefore, it is sufficient to prove for non-constant functions $h(x)$.
- Similarly to the previous proof, we can prove that the function $h(x)$ is differentiable.
- Let $x \in D$, and let λ and x_0 from an open neighborhood of 1 and 0 respectively be such that

$$\lambda \cdot x \in D \text{ and } x + x_0 \in D.$$

- Since the function $h(x)$ is h-scale-invariant, there exist fractional-linear transformations for which

$$h(x + x_0) = \frac{a(x_0) \cdot h(x) + b(x_0)}{1 + c(x_0) \cdot h(x)} \text{ and}$$

$$h(\lambda \cdot x) = \frac{d(\lambda) \cdot h(x) + e(\lambda)}{1 + f(\lambda) \cdot h(x)}.$$

80. Proof of Proposition 3 (cont-d)

- Similarly to the previous proof, we can prove that the functions $a(x_0)$, …, are differentiable.
- So, we can differentiate the λ-formula with respect to λ and take $\lambda = 1$, then we get:
$$x \cdot h' = D \cdot h + E - F \cdot h^2.$$
- Similarly, differentiating the h_0-formula with respect to x_0 and taking $x_0 = 0$, we get:
$$h' = A \cdot h + B - C \cdot h^2.$$
- Let us consider two cases: $C \neq 0$ and $C = 0$.
- Let us first consider the case when $C \neq 0$.
- By completing the square, we get $h' = A \cdot h + B - C \cdot h^2 = \widehat{A} - C \cdot (h - h_0)^2$ for some $\widehat{A}$ and h_0, i.e.,
$$h' = \widehat{A} - C \cdot H^2, \text{ where } H \stackrel{\text{def}}{=} h - h_0.$$

81. Proof of Proposition 3 (cont-d)

- Substituting $h = H + h_0$ into the right-hand side, we conclude that

$$x \cdot h' = \widehat{D} \cdot H + \widehat{E} - F \cdot H^2 \text{ for some } \widehat{D} \text{ and } \widehat{E}.$$

- Dividing the two equations, we get

$$x = \frac{\widehat{D} \cdot H + \widehat{E} - F \cdot H^2}{\widehat{A} - C \cdot H^2}, \text{ so } \frac{dx}{dH} =$$

$$\frac{(\widehat{D} - 2F \cdot H)(\widehat{A} - C \cdot H^2) - (\widehat{D} \cdot H + \widehat{E} - F \cdot H^2)(-2C \cdot H)}{(\widehat{A} - C \cdot H^2)^2} =$$

$$\frac{\widehat{A} \cdot \widehat{D} - 2(\widehat{A} \cdot F - C \cdot \widehat{E}) \cdot H + C \cdot \widehat{D} \cdot H^2}{(\widehat{A} - C \cdot H^2)^2}.$$

- On the other hand,

$$\frac{dx}{dH} = \frac{1}{\dfrac{dH}{dx}} = \frac{1}{\widehat{A} - C \cdot H^2}.$$

82. Proof of Proposition 3 (cont-d)

- The right-hand sides of these two formulas must be equal, so for all H, we have

$$\widehat{A} \cdot \widehat{D} - 2(\widehat{A} \cdot F - C \cdot \widehat{E}) \cdot H + C \cdot \widehat{D} \cdot H^2 = \widehat{A} - C \cdot H^2.$$

- Since the two polynomials of H are equal, the coefficients at 1, H, and H^2 must coincide.
- Comparing the coefficients at H^2, we get $C \cdot \widehat{D} = -C$.
- Since $C \neq 0$, we conclude that $\widehat{D} = -1$.
- Comparing the coefficients at 1, we get $\widehat{A} \cdot \widehat{D} = \widehat{A}$, i.e., $-\widehat{A} = \widehat{A}$ and thus $\widehat{A} = 0$.
- Comparing the coefficients at H and taking into account that $\widehat{A} = 0$, we get $0 = \widehat{A} \cdot F - C \cdot \widehat{E} = -C \cdot \widehat{E}$.
- Since $C \neq 0$, this implies $\widehat{E} = 0$.

83. Proof of Proposition 3 (cont-d)

- So, the above formula for x takes the form
$$x = \frac{\widehat{D} \cdot H - F \cdot H^2}{-C \cdot H^2} = \frac{\widehat{D} - F \cdot H}{-C \cdot H}.$$
- Thus x is a fractional linear function of H.
- Hence H (and therefore $h = H + h_0$) is also a fractional linear function of x.
- Let us now consider the case when $C = 0$.
- Then, $h' = A \cdot h + B$ and $x \cdot h' = D \cdot h + E - F \cdot h^2$, so:
$$x = \frac{x \cdot h'}{h'} = \frac{D \cdot h + E - F \cdot h^2}{A \cdot h + B}.$$
- If $F = 0$, then x is a fractional linear function of $h(x)$ and hence, h is also a fractional-linear function of x.
- So, it is sufficient to consider the case when $F \neq 0$.

84. Proof of Proposition 3 (cont-d)

- In this case, by completing the square, we can find constants $\widehat{D}$, h_0, and $\widehat{B}$ for which, for $H = h - h_0$:
$$x \cdot h' = D \cdot h + E - F \cdot h^2 = \widetilde{D} - F \cdot H^2 \text{ and}$$
$$h' = A \cdot h + B = A \cdot H + \widehat{B}.$$
- Dividing the first equation by the second one, we have
$$x = \frac{\widetilde{D} - F \cdot H^2}{A \cdot H + \widehat{B}}, \text{ thus}$$
$$\frac{dx}{dH} = \frac{(-2F \cdot H) \cdot (A \cdot H + \widehat{B}) - (\widehat{D} - F \cdot H^2) \cdot A}{(A \cdot H + \widehat{B})^2}$$
$$= \frac{-A \cdot \widehat{D} - 2\widehat{B} \cdot F \cdot H - A \cdot F \cdot H^2}{(A \cdot H + \widehat{B})^2}.$$
- On the other hand, $\dfrac{dx}{dH} = \dfrac{1}{\dfrac{dH}{dx}} = \dfrac{1}{A \cdot H + \widehat{B}}$.

85. Proof of Proposition 3 (cont-d)

- By equating the two expressions for the derivative and multiplying both sides by $(A \cdot H + \widehat{B})^2$, we get:

$$-A \cdot \widehat{D} - 2\widehat{B} \cdot F \cdot H - A \cdot F \cdot H^2 = A \cdot H + \widehat{B}.$$

- Thus $A \cdot F = 0$, $A = -2\widehat{B} \cdot F$, and $-A \cdot \widehat{D} = \widehat{B}$.
- If $A = 0$, then we have $\widehat{B} = 0$, so $h' = 0$ and h is a constant.
- However, we consider the case when the function $h(x)$ is not a constant.
- Thus, $A \neq 0$, hence $F = 0$, and the above formula describes x as a fractional-linear function of H.
- Both for $C \neq 0$ and $C = 0$, x is fractionally linear in H (hence in h).
- Since the inverse of a fractional linear is fractional linear, the function $h(x)$ is also fractional linear. Q.E.D.

Part III

Why Ellipsoids in Mechanical Analysis of Wood Structures

86. Formulation of the Problem

- Many constructions are made of wood.
- Wood is one of the oldest materials used in construction.
- During the past millennia, people have developed a lot of skills for working with wood.
- However, in spite of this experience, wood remains one of the most difficult materials to handle.
- The main reason for this difficulty is that:
  - in contrast to many other construction materials which are mostly homogeneous and isotropic,
  - wood is highly inhomogeneous and anisotropic.

87. Formulation of the Problem (cont-d)

- At each point in the wooden beam:
  - both the average values and fluctuations of the local mechanical properties
  - depend on whether the direction is longitudinal, radial or tangential with respect to the grain.
- In designing wooden constructions, it is important:
  - to properly describe and to properly take into account
  - this inhomogeneity and anisotropy.
- How can we describe local fluctuations of mechanical characteristics?
- These fluctuations are caused by many different relatively small factors.

88. Formulation of the Problem (cont-d)

- It is known that the distribution of the joint effect of a large number of small factors is close to Gaussian.
- This follows from the Central Limit Theorem, according to which:
  - this distribution tends to Gaussian
  - when the number of factors increases.
- To describe a Gaussian distribution, it is sufficient to describe its first and second moments.
- For a general random field $f(x)$, this means that we need to describe:
  - its mean values $E[f(x)]$ (where $E[\cdot]$ denotes the expected value) and
  - its covariances $E[f(x) \cdot f(y)]$.

89. Formulation of the Problem (cont-d)

- For fluctuations, the mean is 0, so we only need to describe covariances.
- In statistics, it is often convenient:
  - instead of explicitly describing covariances,
  - to describe the standard deviations and correlations:

$$\sigma[f(x)] \stackrel{\text{def}}{=} \sqrt{E[(f(x)^2]}; \quad \rho(x,y) \stackrel{\text{def}}{=} \frac{E[f(x)\cdot f(y)]}{\sigma[f(x)]\cdot\sigma[f(y)]}.$$

- Then, covariances can be reconstructed as

$$E[f(x)\cdot f(y)] = \sigma[f(x)]\cdot\sigma[f(y)]\cdot\rho(x,y).$$

- An interesting property of the corresponding correlation functions was recently empirically found.

90. Formulation of the Problem (cont-d)

- This property is about:
  - *iso-correlation* surfaces corresponding to each spatial location x,
  - i.e., surfaces formed by all the points y for which the correlation $\rho(x, y)$ is equal to a constant ρ_0.
- Empirical analysis shows that:
  - for each point x,
  - the corresponding surfaces are well approximated by concentric homothetic ellipsoids.
- This property helps narrow down possible functions $\rho(x, y)$ when we analyze mechanical properties of wood.
- Thus, it has a potential to make mechanical analysis of wooden structures more efficient.

91. Formulation of the Problem (cont-d)

- The problem is that so far, this property was purely empirical, it had no theoretical justification.
- Thus, engineers were reluctant to use it.
- It is known that sometimes:
  - empirical properties found under some conditions
  - do not work well when conditions change.
- We want to make this property more reliable and thus, more practically useful.
- It is therefore desirable to come up with a theoretical explanation.
- In this talk, we provide a desired theoretical explanation for this empirical fact.

92. Our Explanation: Main Idea

- We show that there exists the smallest dimension d for which:
  - it is possible to have an affine-invariant optimality criterion
  - on the space of all such d-dimensional classes.
- We also show that for any such criterion, the optimal family consists of concentric homothetic ellipsoids.
- Thus, such families of ellipsoids provide the optimal approximation to the actual surfaces:
  - at least in the *first* approximation, i.e.,
  - approximation corresponding to the smallest possible number of parameters.

93. Family of Sets: Towards a Precise Definition

- For each spatial point x, we would like to describe:
  - for each possible value ρ_0 of the correlation $\rho(x,y)$,
  - the set $S_{\rho_0}(x)=\{y:\rho(x,y)\ge\rho_0\}$.
- What are the natural properties of these families of sets?
- The first property is coverage.
- For each y, there is some value of $\rho(x,y)$.
- So for this x, the union of all these sets $S_{\rho_0}(x)$ coincides with the whole space.
- The second property is monotonicity.
- If $\rho(x,y)\ge\rho_0$ and $\rho_0\ge\rho'_0$, then $\rho(x,y)\ge\rho'_0$.
- So, the sets $S_{\rho_0}(x)$ should be inclusion-monotonic:

$$\text{if } \rho_0\le\rho'_0, \text{ then } S_{\rho'_0}(x)\subseteq S_{\rho_0}(x).$$

94. Family of Sets (cont-d)

- The third property is boundedness.
- From the physical viewpoint:
  - the further away is the point y from the point x,
  - the less the physical quantities corresponding to these points are correlated.
- As the distance increases, this correlation should tend to 0.
- Thus, each set $S_{\rho_0}(x)$ is *bounded*.
- The fourth property is continuity.
- In physics:
  - most processes are continuous,
  - with the exception of processes like fracturing, which we do not consider here.

95. Family of Sets (cont-d)

- We can therefore conclude that the correlation $\rho(x, y)$ continuously depends on y, so:
  - if we have $\rho(x, y_n) \geq \rho_0$ for some sequence of points y_n that converges to a point y $(y_n \to y)$,
  - then we should have $\rho(x, y) = \lim\limits_{n \to \infty} \rho(x, y_n) \geq \rho_0$.
- Thus, if $y_n \in S_{\rho_0}(x)$ and $y_n \to y$, then $y \in S_{\rho_0}(x)$.
- So, each set $S_{\rho_0}(x)$ is *closed*.
- Similarly, it is reasonable to conclude that the set $S_{\rho_0}(x)$ should continually depend on ρ_0:
  - if the two values ρ_0 and ρ'_0 are close,
  - then the corresponding sets $S_{\rho_0}(x)$ and $S_{\rho'_0}(x)$ should also be close.
- A natural way to describe closeness between (bounded closed) sets is to use the so-called Hausdorff distance.

96. Family of Sets (cont-d)

- We say that the sets A and B are ε-*close* if:
  - every point $a \in A$ is ε-close to some point $b \in B$, i.e., $d(a, b) \le \varepsilon$, and
  - every point $b \in B$ is ε-close to some point $a \in A$.
- The Hausdorff distance $d_H(A, B)$ is defined as the smallest ε for which the sets A and B are ε-closed.
- It can be shown that this distance can be equivalently defined as follows:

$$d_H(A, B) = \max\left(\sup_{a \in A} d(a, B), \sup_{b \in B} d(b, A)\right), \text{ where}$$

$$d(a, B) \overset{\text{def}}{=} \inf_{b \in B} d(a, b).$$

- What is the set of possible values of the parameter?
- In this family of sets, correlation value is a parameter.

97. Family of Sets (cont-d)

- Correlations can take any value from -1 to 1.
- When $y = x$, the correlation is clearly equal to 1.
- When $y \to \infty$, we get values close to 0.
- Since the function $\rho(x, y)$ is continuous, this function takes all intermediate values.
- So, the possible values of the correlation form some interval.
- In some cases, we may have all possible negative values.
- In other cases, only some negative values, in yet other cases, we only have non-negative values.
- So, in general, we will consider all possible intervals of possible value of ρ_0.
- This interval may be closed – if there are points with limit correlation, or is can be open.

98. Definition

- So, we arrive at the following definition.
- Let $N \geq 2$ be an integer.
- Let I be an interval.
- By a *family of sets*, we mean a set $\{S_c : c \in I\}$ of bounded closed sets $S_c \subseteq \mathbb{R}^N$ for which:
  - the dependence of S_c on c is continuous: if $c_n \to c$, then $d_H(S_{c_n}, S_c) \to 0$;
  - the family S_c is monotonic: if $c < c'$, then $S_{c'} \subseteq S_c$; and
  - the union of all the sets S_c coincides with the whole space.

99. Comments

- According to this definition, the family remains the same if we simply re-parameterize the family.

- For example:

  - instead of the original parameter c,
  - we can use a new parameter $c' = c + c_0$ or $c' = \lambda \cdot c$ for some constants c_0 and λ.

- In our specific problem, we are interested in the 3-D case $N = 3$.

- However, we can envision similar problem in the plane $N = 2$ or in higher-dimensional spaces.

- So, in this talk, we consider the general case $N \geq 2$.

100. Comments (cont-d)

- We are specifically interested:
  - in *concentric homothetic families of ellipsoids*, i.e.,
  - in families of the type $S_c = c \cdot E + a$, where a is a given vector, and E is an ellipsoid with center 0.

101. Class of Families of Sets

- For different situation, in general:
  - we get different correlations and thus,
  - we get different families of sets.
- We would like to find a general class of such families that would, ideally, cover all such situations.
- We can use different parameters to differentiate different families from this class.
- In other words, a class can be described as a method for assigning:
  - to each possible combination of values of these parameters,
  - a specific family.
- As before, it makes sense to require that the resulting mapping is continuous.

102. Class of Families of Sets (cont-d)

- Here is a precise definition.
- Let $N \geq 2$ and $r > 0$ be integers.
- By an *r-parametric class of families of sets*, we mean a mapping that assigns,
  - to each element $p = (p_1, \ldots, p_r)$ from an open r-dimensional set $D \subseteq \mathbb{R}^r$,
  - a family $\{S_c(p)\}$ so that the dependence of $S_c(p)$ on c and p is continuous.

103. Optimality Criteria: General Idea

- Out of all possible classes, we want to select a class which is, in some reasonable sense, optimal.
- For this, we need to be able to describe when some classes are better than others.
- In other words, we need to have an *order* on the set of all the classes.
- It would be nice to have a *total* (*linear*) order, in the sense that:
  - for every two classes,
  - we should be able to tell which one is better.
- However, it may be sufficient to have a *partial* order – as long as this order enables us to select the best class.
- It is OK if for some not-best classes, we do not have an opinion of which of them is better.

104. Optimality Criteria: General Idea (cont-d)

- In practice, usually, optimality criteria are described in numerical form:
  - we have an objective function $f(a)$ that assigns a numerical value to each possible alternative a, and
  - we want to select an alternative for which this value is the largest possible,
  - or, depending on the context, the smallest possible.
- For example:
  - a company wants to maximize its profit,
  - a city wants to upgrade its road system so as to minimize the average travel time, etc.
- However, often, we need to go somewhat beyond this approach.

105. Optimality Criteria: General Idea (cont-d)

- For example, a company may have two (or more) different projects that lead to the same expected profit.

- In this case, we can use this non-uniqueness to optimize something else.

- For example:
  - out of all most-profitable projects,
  - we can select the one that leads to the smallest possible long-term environmental impact.

- In this case, we have a more complex criterion for comparing alternatives: we say that a is better if:
  - either $f(a) > f(a')$
  - or $f(a) = f(a')$ and $g(a) > g(a')$, for some other numerical criterion $g(a)$.

106. Optimality Criteria: General Idea (cont-d)

- If this still does not select us a unique alternative, we can optimize yet something else, etc.

- In view of this possibility, in this talk, we do not restrict ourselves to numerical optimization criteria.

- Instead, we use the most general definition of the optimality criterion, when:
  - for some pairs of alternatives a and a', we know that a is better (we will denote it by $a' < a$),
  - for some pairs of alternatives a and a', we know that a' is better ($a < a'$), and
  - for some pairs of alternatives a and a', a and a' are of the same value (we will denote it by $a \sim a'$).

- Clearly, if a' is better than a, and a'' is better than a', then a'' should be better than a, etc.

107. Optimality Criteria: General Idea (cont-d)

- Thus, we arrive at the following definition
- Let A be a set; elements of this set will be called *alternatives*.
- By an *optimality criterion*, we mean a pair of binary relations $(<, \sim)$ on the set A for which:
  - if $a < a'$ and $a' < a''$, then $a < a''$;
  - if $a < a'$ and $a' \sim a''$, then $a < a''$;
  - if $a \sim a'$ and $a' < a''$, then $a < a''$;
  - if $a \sim a'$ and $a' \sim a''$, then $a \sim a''$;
  - if $a \sim a'$, then $a' \sim a$;
  - if $a < a'$, then we cannot have $a' < a$ or $a \sim a'$.
- Such a pair of relations is sometimes called a *partial pre-order*.

108. Optimality Criteria: General Idea (cont-d)

- Let $(<, \sim)$ be an optimality criterion on a set A.
- An alternative a_{opt} is called *optimal* with respect to this criterion if for every alternative $a \in A$, we have

$$a < a_{\text{opt}} \text{ or } a \sim a_{\text{opt}}.$$

109. We Need A Final Optimality Criterion

- For the optimality criterion to be useful, it must select *at least one* optimal alternative.
- If the criterion selects *several* alternatives as optimal, this means that this criterion is not final.
- We can use the resulting non-uniqueness:
  - to optimize something else,
  - i.e., in effect, to come up with a better optimality criterion.
- If for this better criterion, we still have several optimal alternatives, we should modify this criterion again.
- Finally, we get a criterion for which there is exactly one optimal alternative.
- We will call such criteria *final*.

110. For Our Problem, an Optimality Criterion Must Be Affine-Invariant

- In our case, we want to compare different classes (of families of sets).
- In selecting optimality criteria, it is reasonable to take into account that:
  - while we want to deal with sets of points in physical space,
  - from the mathematical viewpoint, we deal with sets of tuples of real numbers.
- Real numbers (coordinates) describing each point depend on what coordinate system we use.
- If we select a different starting point, then all the coordinates are shifted $x_i \to x_i + a_i$.

111. Affine-Invariant (cont-d)

- If we select different axes for the coordinates, we get a rotation $x_i \to \sum\limits_{j=1}^{N} r_{ij} \cdot x_j$ for an appropriate matrix r_{ij}.
- These transformations make sense for the *isotropic* case, when:
  - all the properties of a material
  - are the same in all directions.
- Wood is an example of an *anisotropic* material.
- For example, it is easier to cut it along the orientation of the original tree than across that orientation.
- It is known that in many cases:
  - the description of an anisotropic material can be reduced to the isotropic case
  - if we apply an appropriate affine transformation.

112. Affine-Invariant (cont-d)

- This usually comes from the fact that, e.g.:
  - mechanical properties of a body can be described by a symmetric matrix, and
  - a symmetric matrix becomes symmetric if we use its eigenvectors as new axes.
- In view of this, it is reasonable to require that our optimality criterion is invariant:
  - not only with respect to shifts and rotations,
  - but also with respect to all possible affine (linear) transformations.
- Thus, we arrive at the following definitions.
- Let $N > 2$ be an integer.

113. Affine-Invariant (cont-d)

- By an *affine transformation*, we mean $(Tx)_i = a_i + \sum\limits_{j=1}^{N} b_{ij} \cdot x_j$ for some reversible matrix b_{ij}.
- Let T be an affine transformation.
- Let $S \subseteq \mathbb{R}^N$ be a set.
- By the *result* $T(S)$ of applying T to S, we mean the set $\{T(s) : s \in S\}$.
- Let $F = \{S_c : c \in I\}$ be a family of sets.
- By the *result* $T(F)$ of applying T to F, we mean the family $\{T(S_c) : c \in I\}$.
- Let $C = \{S_c(p)\}$ be class of families.
- By the *result* $T(C)$ of applying T to C, we mean the class $\{T(S_c(p))\}$.

114. Affine-Invariant (cont-d)

- Let A be a set of alternatives, let $(<, \sim)$ be an optimality criterion of the set A.
- Let $\mathcal{T}$ be a class of transformations $A \to A$.
- We say that $(<, \sim)$ is $\mathcal{T}$-*invariant* if for all $T \in \mathcal{T}$ and $a, a' \in A$, we have:
  - if $a < a'$ then $T(a) < T(a')$, and
  - If $a \sim a'$, then $T(a) \sim T(a)$.

115. Main Result

- Let $N > 0$ and $r > 0$ be integers.
- We consider sets in $\mathbb{R}^N$.
- Let $(<, \sim)$ be a final affine-invariant optimality criterion on the set of all r-parametric classes of families.
- Then $r \geq r_{\min} \stackrel{\text{def}}{=} \dfrac{N \cdot (N+3)}{2} - 1$, and:
  - for $r = r_{\min}$,
  - the optimal class consists of concentric homothetic families of ellipsoids.
- This result indeed shows that:
  - the class of concentric homothetic families of ellipsoids
  - is the simplest (= fewer parameters) of all possible optimal classes.

116. Proof

- Since the optimality criterion is final, there exists exactly one optimal class $C_{\rm opt}$ for which:

$$C < C_{\rm opt} \text{ or } C \sim C_{\rm opt} \text{ for all classes } C.$$

- Let us prove that the optimal class $C_{\rm opt}$ is itself affine-invariant, i.e., that $T(C_{\rm opt}) = C_{\rm opt}$ for each affine T.
- Indeed, due to optimality, for each class C and for each affine transformation class T, for $T^{-1}(C)$, we have:

$$T^{-1}(C) < C_{\rm opt} \text{ or } T^{-1}(C) \sim C_{\rm opt}.$$

- Since the criterion is affine-invariant, we have:

$$T(T^{-1}(C)) < T(C_{\rm opt}) \text{ or } T(T^{-1}(C)) \sim T(C_{\rm opt}).$$

- Here, by the definition of the inverse transformation:

$$T(T^{-1}(C)) = C.$$

117. Proof (cont-d)

- So we conclude that for every class C, we have:

$$C < T(C_{\text{opt}}) \text{ or } C \sim T(C_{\text{opt}}).$$

- By definition of optimality, this means that the class $T(C_{\text{opt}})$ is optimal.
- However, our optimality criterion is final, which means that there is only one optimal class.
- Thus, indeed, $T(C_{\text{opt}}) = C_{\text{opt}}$.
- Since the optimal class is affine-invariant, with each family F this class also contains the family $T(F)$.
- This means that for each set S_c from each family, some family from the optimal class contains the set $T(S_c)$.
- Let us show that $r \geq \dfrac{N \cdot (N+3)}{2} - 1$.

118. Proof (cont-d)

- Indeed, it is known that:
  - for every non-degenerate bounded set S (i.e., not contained in a proper subspace),
  - among all ellipsoids that contain S, there exists a unique ellipsoid of the smallest volume.
- It is also known that this correspondence between a set and the corresponding ellipsoid is affine-invariant:
  - if an ellipsoid E corresponds to the set S_c, then,
  - for each affine transformation T, to the set $T(S_c)$ there corresponds the ellipsoid $T(E)$.
- It is known that every two ellipsoids can be obtained from each other by an affine transformation.

119. Proof (cont-d)

- Thus:
  - the family of all ellipsoids corresponding to all the sets from all the families
  - consists of all the ellipsoids.
- How many ellipsoids are there?
- A general ellipsoid can be determine by a quadratic formula $\sum\limits_{ij} a_{ij} \cdot x_i \cdot x_j + \sum\limits_{i=1}^{N} a_i \cdot x_i \le 1.$
- Here, a_{ij} is a symmetric matrix a_{ij} and a_i is a vector.
- It is easy to see that different combinations of the matrix and the vector lead to different ellipsoids.
- We need N values $a_1, \ldots, a_N$ to describe a vector.

120. Proof (cont-d)

- Out of N^2 elements of the matrix:
  - we need N values to describe its diagonal values a_{ii}, and
  - we need $\dfrac{N^2 - N}{2}$ to describe non-diagonal elements.
- We divide by two since the matrix is symmetric:

$$a_{ij} = a_{ji}.$$

- Thus, overall, we need

$$N + N + \frac{N^2 - N}{2} = \frac{N \cdot (N+3)}{2} \text{ values.}$$

- So, the set of all ellipsoids is:

$$\frac{N \cdot (N+3)}{2}\text{-dimensional.}$$

121. Proof (cont-d)

- To each set S_c from families from the optimal class, we assign an ellipsoid.
- Thus, the dimension of the set of such sets should also be at least $\dfrac{N\cdot(N+3)}{2}$-dimensional.
- These sets are divided into 1-parametric families.
- So the dimension r of the class of such families cannot be smaller than the above dimension minus 1.
- Thus, indeed, $r \geq \dfrac{N\cdot(N+3)}{2} - 1$.

122. Proof (cont-d)

- Let us now prove that:
  - for the smallest possible dimension
    $$r = r_{\min} \stackrel{\text{def}}{=} \frac{N \cdot (N+3)}{2} - 1,$$
  - all the sets S_c from the each family of the optimal class are ellipsoids.
- Indeed, we showed that each ellipsoid is associated with some set S_c from one of these families.
- The unit ball with a center at 0 is clearly an ellipsoid.
- Let us consider the set S_c which is associated with this unit ball.
- A unit ball is invariant with respect to all the rotations around its center.

123. Proof (cont-d)

- If the associated set S_c is not equal to the unit ball, this means that:
  - this set is not invariant
  - with respect to at least some rotations.
- In other words:
  - the group of all rotations that leave this set invariant
  - is a proper subgroup of the group of all rotations.
- This implies that the dimension of this group is smaller than the dimension of the group of all rotations.
- Thus, that there exists at least 1-parametric family $\mathcal{R}$ of rotations R w.r.t. which the set S_c is not invariant.
- The optimal class is affine-invariant.

124. Proof (cont-d)

- Thus, all the sets $R(S_c)$ are also sets from some family from the optimal class.
- For all these sets, the same unit ball is the smallest-volume ellipsoid.
- Thus, for this particular ellipsoid – the unit ball:
  - we have at least a 1-dimensional family of sets S_c
  - associated with this ellipsoid.
- By applying a generic affine transformation:
  - we can find a similar at-least-1-dimensional family of sets
  - corresponding to each ellipsoid.

125. Proof (cont-d)

- Thus:
  - the dimension of the set of all sets S_c
  - is at least one larger than the dimension of the family of all ellipsoids,
  - i.e. at least $\dfrac{N \cdot (N+3)}{2} + 1 = r_{\min} + 2.$
- However, we have a $r_{\min}$-dimensional class of 1-dimensional families of sets.
- So the overall dimension of the set of all the sets S_c cannot be larger than $r_{\min} + 1$.
- This contradiction shows that the set S_c cannot be different from the enclosing minimal-volume ellipsoid.
- Thus, indeed, each set from each family from the optimal class is an ellipsoid.

126. Completing the Proof

- To complete the proof, we need to prove that ellipsoids in each family are concentric and homothetic.
- We have proven that each ellipsoid appears as an appropriate smallest-volume set.
- We know that each set S_c coincides with its smallest-volume enclosure.
- So, each ellipsoid appears as one of the sets S_c from one of the families from the optimal class.
- Let us again consider the unit ball centered at 0:
  - if the 1-dimensional family F_0 containing this ball is not invariant with respect to all possible rotations,
  - then we have at least a 1-dimensional group of different families containing the same ellipsoid.

127. Completing the Proof (cont-d)

- We have:
  - an $r_{\min}$-dimensional class of 1-dimensional families
  - covering the whole $(r_{\max}+1)$-dimensional family of ellipsoids.
- Thus, all elements of all families are different.
- So we cannot have several families containing the same ellipsoid.
- This argument shows that the family F_0 containing the unit ball *should be* rotation-invariant.
- All the sets from this family are included in each other and thus, cannot be rotated into each other.
- This means that each ellipsoid from this family F_0 must be rotation-invariant.

128. Completing the Proof (cont-d)

- This means that each ellipsoid from this family must be a ball concentric with our selected unit ball.
- Thus, it be homothetic to the original ball.
- For any other family F:
  - by selecting any ellipsoid E from this family and
  - by applying the affine transformation that transforms the above unit ball into E,
  - we get a new family $T(F_0)$ of concentric homothetic ellipsoids.
- An ellipsoid can only belong to one family.
- We thus conclude that the family F also consists of concentric homothetic ellipsoids.
- The result is proven.

129. Conclusions

- Wood is one the oldest construction materials; however:
  - in spite of several thousand years of experience with wooden constructions,
  - predicting and estimating mechanical properties of wooden constructions remains a difficult problem.
- One of the main reasons for this difficulty is that:
  - in contrast to many other constructions materials which are largely homogeneous and isotropic,
  - wood is highly inhomogeneous and anisotropic.
- Recently, a new property of wooden materials was discovered.
- It has a potential to make mechanical analysis of wooden structures more efficient.

130. Conclusions (cont-d)

- Namely, for wood:
  - iso-correlation surfaces (i.e., surfaces of equal correlation)
  - are well-approximated by concentric homothetic ellipsoids.
- The problem is that this property is purely empirical.
- It has no theoretical explanation and thus, engineers are understandably reluctant to rely on it.
- In this talk, we provide a theoretical explanation for this empirical fact.
- Thus, we make this property more reliable and therefore more useable.

Part IV

Why Spiking Neural Networks Are Efficient: A Theorem

131. Why Spiking Neural Networks (NN)

- At this moment, artificial neural networks are the most successful – and the most promising – direction in AI.
- Artificial neural networks are largely patterned after the way the actual biological neural networks work.
- This patterning makes perfect sense:
  - after all, our brains are the result of billions of years of improving evolution,
  - so it is reasonable to conclude that many features of biological neural networks are close to optimal,
  - not very efficient features would have been filtered out in this long evolutionary process.
- However, there is an important difference between the current artificial NN and biological NN.

132. Why Spiking NN (cont-d)

- In hardware-implemented artificial NN each value is represented by the intensity of the signal.
- In contrast, in the biological neural networks, each value is represented by a frequency instantaneous spikes.
- Since simulating many other features of biological neural networks has led to many successes.
- So, a natural idea is to also try to emulate the spiking character of the biological neural networks.

133. Spiking Neural Networks Are Indeed Efficient

- Interestingly, adding spiking to artificial neural networks has indeed led to many successful applications.
- They were especially successful in processing temporal (and even spatio-temporal) signals.
- A biological explanation of the success of spiking neural networks makes perfect sense.
- However, it would be nice to supplement it with a clear mathematical explanation.
- It is especially important since:
  - in spite of all the billions years of evolution,
  - we humans are not perfect as biological beings,
  - we need medicines, surgeries, and other artificial techniques to survive, and
  - our brains often make mistakes.

134. Looking for Basic Functions

- In general, to represent a signal $x(t)$ means to approximate it as a linear combination of some basic functions.
- For example, it is reasonable to represent a periodic signal as a linear combination of sines and cosines.
- Often, it makes sense to represent the observed values as a linear combination of:
  - functions t, t^2, etc., representing the trend and
  - sines and cosines that describe the periodic part of the signal.
- We can also take into account that the amplitudes of the periodic components can also change with time.
- So, we end up with terms of the type $t \cdot \sin(\omega \cdot t)$.

135. Looking for Basic Functions (cont-d)

- For radioactivity, the observed signal is:
  - a linear combination of functions $\exp(-k \cdot t)$
  - that represent the decay of different isotopes.
- So, in precise terms, selecting a representation means selecting an appropriate family of basic functions.
- In general, elements $b(t)$ of a family can be described as $b(t) = B(c_1, \ldots, c_n, t)$ corr. to diff. $c = (c_1, \ldots, c_n)$.
- Sometimes, there is only one parameter, as in sines and cosines.
- In control, typical are functions $\exp(-k \cdot t) \cdot \sin(\omega \cdot t)$, with two parameters k and ω, etc.

136. Dependence on Parameters Is Continuous

- We want the dependence $B(c_1, \ldots, c_n, t)$ to be computable.
- It is known that all computable functions are, in some reasonable sense, continuous.
- Indeed, in real life, we can only determine the values of all physical quantities c_i with some accuracy.
- Measurements are always not 100% accurate, and computations always involve some rounding.
- For any given accuracy, we can provide the value with this accuracy.
- Thus, the approximate values of c_i are the only thing that $B(c_1, \ldots, c_n, t)$-computing algorithm can use.
- This algorithm can ask for more and more accurate values of c_i.

137. Dependence Is Continuous (cont-d)

- However, at some point it must produce the result.
- At this point, we only known approximate values of c_i.
- So, we only know the interval of possible values of c_i.
- And for all the values of c_i from this interval:
  - the result of the algorithm provides, with the given accuracy,
  - the approximation to the desired value $B(c_1, \ldots, c_n, t)$.
- This is exactly what continuity is about!
- One has to be careful here, since the real-life processes may actually be discontinuous.
- Sudden collapses, explosions, fractures do happen.

138. Dependence Is Continuous (cont-d)

- For example, we want to make sure that:
  - a step-function which is equal to 0 for $t < 0$ and to 1 for $t \ge 0$ is close to
  - an "almost" step function which is equal to 0 for $t < 0$, to 1 for $t \ge \varepsilon$ and to t/ε for $t \in (0, \varepsilon)$.
- In such situations:
  - we cannot exactly describe the value at moment t,
  - since the moment t is also measured approximately.
- What we can describe is its values at a moment close to t.

139. Dependence Is Continuous (cont-d)

- In other words, we can say that the two functions $a_1(t)$ and $a_2(t)$ are ε-close if:
  - for each t_1, there are ε-close t_{21}, t_{22} such that $a_1(t_1)$ is ε-close to a convex combination of $a_2(t_{2i})$;
  - for each t_2, there are ε-t_{11}, t_{12} such that $a_2(t_2)$ is ε-close to a convex combination of $a_1(t_{1i})$.

140. Additional Requirement

- We consider linear combinations of basic functions.
- So, it does not make sense to have two basic functions that differ only by a constant.
- If $b_2(t) = C \cdot b_1(t)$, then there is no need to consider the function $b_2(t)$ at all.
- In each linear combination we can replace $b_2(t)$ with
$$C \cdot b_1(t).$$

141. We Would Like to Have the Simplest Possible Family

- How many parameters c_i do we need? The fewer parameters:
  - the easier it is to adjust the values of these parameters, and
  - the smaller the probability of *overfitting* – a known problem of machine learning and data analysis in general.
- We cannot have a family with no parameters at all; this would mean, in effect, that:
  - we have only one basic function $b(t)$ and
  - we approximate every signal by an expression $C \cdot b(t)$ obtained by its scaling.

142. Simplest Possible Family (cont-d)

- This will be a very lousy approximation to real-life processes:
  - these processes are all different,
  - they do not resemble each other at all.
- So, we need at least one parameter.
- We are looking for the simplest possible family.
- So, we should therefore consider families depending on a single parameter c_1.
- In precise terms, we need functions $b(t) = B(c_1, t)$ corresponding to different values of the parameter c_1.

143. Most Observed Processes Are Limited in Time

- From our viewpoint, we may view astronomical processes as going on forever.
- In reality, even they are limited by billions of years.
- In general, the vast majority of processes that we observe and that we want to predict are limited in time.
- A thunderstorm stops, a hurricane ends, after-shocks of an earthquake stop, etc.
- From this viewpoint:
  - to get a reasonable description of such processes,
  - it is desirable to have basic functions which are also limited in time,
  - i.e., which are equal to 0 outside some finite time interval.

144. Limited in Time (cont-d)

- This need for finite duration is one of the main reasons in many practical problems:
  - a decomposition into wavelets performs much better than
  - a more traditional Fourier expansion into linear combinations of sines and cosines.

145. Shift- and Scale-Invariance

- Processes can start at any moment of time.
- Suppose that we have a process starting at moment 0 which is described by a function $x(t)$.
- What if we start the same process t_0 moments earlier?
- At each moment t, the new process $x'(t)$ has been happening for the time period $t+t_0$, so $x'(t)=x(t+t_0)$.
- There is no special starting point.
- So it is reasonable to require that the class of basic function not change if we change the starting point:

$$\{B(c_1,t+t_0)\}_{c_1}=\{B(c_1,t)\}_{c_1}.$$

- Similarly, processes can have different speed.

146. Shift- and Scale-Invariance (cont-d)

- Some processes are slow, some are faster:
  - if a process starting at 0 is $x(t)$,
  - then a λ times faster process is characterized by the function $x'(t) = x(\lambda \cdot t)$.
- There is no special speed.
- So it is reasonable to require that the class of basic function not change if we change the process's speed:

$$\{B(c_1, \lambda \cdot t\}_{c_1} = \{B(c_1, t)\}_{c_1}.$$

- Now, we are ready for the formal definitions.

147. Definitions and the First Result

- We say that a function $b(t)$ is *limited in time* if it equal to 0 outside some interval.
- We say that a function $b(t)$ is a *spike* if it is different from 0 only for a single value t.
- This non-zero value is called the *height* of the spike.
- Let $\varepsilon > 0$ be a real number.
- We say that the numbers a_1 and a_2 are ε-*close* if

$$|a_1 - a_2| \leq \varepsilon.$$

- We already had a definition of the functions $a_1(t)$ and $a_2(t)$ being ε-*close*.

148. Definitions and the First Result (cont-d)

- We say that a mapping $B(c_1, t)$ is *continuous* if, for every c_1 and $\varepsilon > 0$, there exists $\delta > 0$ such that:
  - if c'_1 is δ-close to c_1,
  - then the function $b(t) = B(c_1, t)$ is ε-close to the function $b'(t) = B(c'_1, t)$.
- By a *family of basic functions*, we mean a continuous mapping for which:
  - for each c_1, the function $b(t) = B(c_1, t)$ is limited in time, and
  - if $c_1 \neq c'_1$, then $B(c'_1, t) \not\equiv C \cdot B(c_1, t)$.
- We say that a family $B(c_1, t)$ is *shift-invariant* if for each t_0: $\{B(c_1, t)\}_{c_1} = \{B(c_1, t + t_0)\}_{c_1}$.
- We say that a family $B(c_1, t)$ is *scale-invariant* if for each $\lambda > 0$: $\{B(c_1, t)\}_{c_1} = \{B(c_1, \lambda \cdot t)\}_{c_1}$.

149. The First Result (cont-d)

- **Proposition.** *If a family of basic functions $B(c_1, t)$ is shift- and scale-invariant, then:*
  - *for every c_1, the corresponding function $b(t) = B(c_1, t)$ is a spike, and*
  - *all these spikes have the same height.*
- This result provides a possible explanation for the efficiency of spikes.

150. Proof

- Let us assume that the family of basic functions $B(c_1, t)$ is shift- and scale-invariant.
- Let us prove that all the functions $b(t) = B(c_1, t)$ are spikes.
- First, we prove that none of the functions $B(c_1, t)$ is identically 0.
- Indeed, the zero function can be contained from any other function by multiplying by 0.
- This would violate the definition of a family of basic functions).
- Let us prove that each function from the given family is a spike.
- Indeed, each of the functions $b(t) = B(c_1, t)$ is not identically zero, i.e., it attains non-zero values for some t.

151. Proof (cont-d)

- By definition, each of these functions is limited in time.
- So, the values t for which the function $b(t)$ is non-zero are bounded by some interval.
- Thus, the values $t_- \stackrel{\text{def}}{=} \inf\{t : b(t) \neq 0\}$ and $t_+ \stackrel{\text{def}}{=} \sup\{t : b(t) \neq 0\}$ are finite, with $t_- \leq t_+$.
- Let us prove that we cannot have $t_- < t_+$.
- Indeed, in this case, the interval $[t_-, t_+]$ is non-degenerate; thus:
  - by an appropriate combination of shift and scaling,
  - we will be able to get this interval from any other non-degenerate interval $[a, b]$.
- The family is shift- and scale-invariant.
- Thus, the correspondingly re-scaled function $b'(t) = b(\lambda \cdot t + t_0)$ also belongs to the family $B(c_1, t)$.

152. Proof (cont-d)

- For this function, the corresponding values t'_- and t'_+ will coincide with a and b.
- All these functions are different – so, we will have a 2-dimensional family of functions.
- This contradicts to our assumption that the family $B(c_1, t)$ is one-dimensional.
- We cannot have $t_- < t_+$, so $t_- = t_+$, i.e., every function from our family is a spike.
- Let us prove that all the spikes have the same height.
- Indeed, let $b_1(t)$ and $b_2(t)$ be any two functions from the family.

153. Proof (cont-d)

- Both functions are spikes, so:
  - the value $b_1(t)$ is only different from 0 for some value t_1, its height is $h_1 \stackrel{\text{def}}{=} b_1(t_1)$;
  - similarly, the value $b_2(t)$ is only different from 0 for some value t_2, its height is $h_2 \stackrel{\text{def}}{=} b_2(t_2)$.
- Since the family $\mathcal{B}$ is shift-invariant, for $t_0 \stackrel{\text{def}}{=} t_1 - t_2$, the shifted function $b_1'(t) \stackrel{\text{def}}{=} b_1(t + t_0)$ is also in $\mathcal{B}$.
- The shifted function is non-zero when $t + t_0 = t_1$, i.e., when $t = t_1 - t_0 = t_2$, and it has the same height h_1.
- If $h_1 \neq h_2$, we would have $b_1'(t) = C \cdot b_2(t)$ for $C \neq 1$.
- Thus, the heights must be the same.
- The proposition is proven.

154. But Are Spiked Neurons Optimal?

- We showed that spikes naturally appear if we require reasonable properties like shift- and scale-invariance.
- This provides some justification for the spiked neural networks.
- However, the ultimate goal of neural networks is to solve practical problems.
- A practitioner is not interested in invariance or other mathematical properties.
- A practitioner wants to optimize some objective function.
- So, from the practitioner's viewpoint, the main question is: are spiked neurons optimal?

155. Different Practitioners Have Different Optimality Criteria

- In principle:
  - we can pick one such criterion (or two or three) and
  - analyze which families of basic functions are optimal with respect to these particular criterion.
- However, this will not be very convincing to a practitioner who has a different optimality criterion.
- An ideal explanation should work for *all* reasonable optimality criteria.
- To achieve this goal, let us analyze which optimality criteria can be considered reasonable.

156. What Is an Optimality Criterion: Analysis

- At first glance, the answer to this question may sound straightforward,
- We have an objective function $J(a)$ that assigns, to each alternative a, a numerical value $J(a)$
- We want to select an alternative for which the value of this function is the largest possible.
- If we are interested in minimizing losses, the value is the smallest possible.
- This formulation indeed describes many optimality criteria, but not all of them.
- Indeed, assume, for example, we are looking for the best method a for approximating functions.
- A natural criterion may be to minimize the mean squared approximation error $J(a)$ of the method a.

157. What Is an Optimality Criterion (cont-d)

- If there is only one method with the smallest possible mean squared error, then this method is selected.
- But what if there are several different methods with the same mean squared error.
- This, by the way, is often the case.
- In this case, we can use this non-uniqueness to optimize something else; e.g., we can select:
  - out of several methods with the same mean squared error,
  - the method for which the average computation time $T(a)$ is the smallest.
- The actual optimality criterion cannot be described by single objective function, it is more complex.

158. What Is an Optimality Criterion (cont-d)

- Namely, we say that a method a' is better than a method a if:
  - either $J(a) < J(a')$,
  - or $J(a) = J(a')$ and $T(a) < T(a')$.
- This additional criterion may still leave us with several equally good methods.
- We can use this non-uniqueness to optimize yet another criterion: e.g., worst-case computation time, etc.
- This criterion must enable us to decide which alternatives are better (or of the same quality).
- Let us denote this by $a \le a'$.
- Clearly, if $a \le a'$ and $a' \le a''$, then $a \le a''$, so the relation $\le$ must be transitive (a.k.a. *pre-orders*).

159. An Optimality Criterion Must Be Final

- In terms of the relation $\leq$, optimal means better than (or of the same quality as) all other alternatives:

$$a \leq a_{\text{opt}} \text{ for all } a.$$

- If we have several optimal alternatives, then we can use this non-uniqueness to optimize something else.
- So, the corresponding criterion is not final.
- For a *final* criterion, we should have only one optimal alternative.

160. An Optimality Criterion Must Be Invariant

- In real life, we deal with real-life processes $x(t)$, in which values of different quantities change with time t.
- The corresponding numerical values of time t depend:
  - on the starting point that we use for measuring time and
  - on the measuring unit.
- For example, 1 hour is equivalent to 60 minutes.
- Numerical values are different, but from the physical viewpoint, this is the same time interval.
- We are interested in a universal technique for processing data.

161. Criterion Must Be Invariant (cont-d)

- It is therefore reasonable to require that:
  - the relative quality of different techniques should not change
  - if we change the starting point for measuring time or a measuring unit.
- Let us describe all this in precise terms.

162. Definitions and the Main Result

- Let a set A be given; its elements will be called *alternatives*.
- By an *optimality criterion* $\le$ on the set A, we mean a transitive relation (i.e., a *pre-order*) on this set.
- An element a_{opt} is called *optimal* with respect to the criterion $\le$ is for all $a \in A$, we have $a \le a_{\text{opt}}$.
- An optimality criterion is called *final* if there exists exactly one optimal alternative.
- For each family $B(c_1, t)$ and for each t_0, by its *shift* $T_{t_0}(B)$, we mean a family $B(c_1, t + t_0)$.
- We say that an optimality criterion on the class of all families is *shift-invariant* if
  - for every two families B and B' and for each t_0,
  - $B \le B'$ implies that $T_{t_0}(B) \le T_{t_0}(B')$.

163. Definitions and the Main Result (cont-d)

- For each family $B(c_1, t)$ and for each $\lambda > 0$, by its *scaling* $S_\lambda(B)$, we mean a family $B(c_1, \lambda \cdot t)$.
- We say that an optimality criterion on the class of families is *scale-invariant* if:
  - for every two families B and B' and for each $\lambda > 0$,
  - $B \leq B'$ implies that $S_\lambda(B) \leq S_\lambda(B')$.
- **Proposition.**
  - *Let $\leq$ be a final shift- and scale-invariant optimality criterion on the class of all families of basic f-s.*
  - *Then, all elements of the optimal family are spikes of the same height.*

164. Discussion

- Techniques based on representing signals as a linear combination of spikes are known to be very efficient.
- In different applications, efficiency mean different things: faster computations, more accurate results, etc.
- In different situations, we may have different optimality criteria.
- Our result shows that no matter what optimality criterion we use, spikes are optimal.
- This explains why spiking NN have been efficient in several different situations, with different criteria.

165. Proof

- Let us prove that the optimal family B_{opt} is itself shift- and scale-invariant.
- Then this result will follow from the previous Proposition.
- Indeed, let us consider any transformation T – be it shift or scaling.
- By definition of optimality, for any other family B, we have $B \le B_{\text{opt}}$.
- In particular, for every B, this is true for $T^{-1}(B)$, i.e., $T^{-1}(B) \le B_{\text{opt}}$.
- Here, T^{-1} denotes the inverse transformation.
- Due to invariance, $T^{-1}(B) \le B_{\text{opt}}$ implies that $T(T^{-1}(B)) \le T(B_{\text{opt}})$, i.e., that $B \le T(B_{\text{opt}})$.

166. Proof (cont-d)

- This is true for each family B, thus the family $T(B_{\text{opt}})$ is optimal.
- However, our optimality criterion is final, i.e., there is only one optimal family.
- Thus, we have $T(B_{\text{opt}}) = B_{\text{opt}}$.
- So, the optimal family B_{opt} is indeed invariant with respect to any of the shifts and scalings.
- Now, by applying the previous Proposition, we conclude the proof of this proposition.

167. Conclusions

- A usual way to process signals is to approximate each signal by a linear combinations of basic functions.
- Examples: sinusoids, wavelets, etc.
- In the last decades, a new approximation turned out to be very efficient in many practical applications.
- Namely, approximation of a signal by a linear combination of spikes.
- In this talk, we provide a possible theoretical explanation for this empirical success.

168. Conclusions (cont-d)

- Our main explanation is that:
  - for every reasonable optimality criterion on the class of all possible families of basic functions,
  - the optimal family is the family of spikes,
  - provided that the optimality criterion is scale- and shift-invariant.

Part V
Why Most Empirical Distributions Are Few-Modal

169. Empirical Distributions: We Expect Them to Be Multi-Modal

- Continuous distributions are characterized by their probability density functions $\rho(x)$.
- In principle, a probability density function can be any non-negative function.
- The only condition is that the overall probability should be equal to 1, i.e., that $\int \rho(x)\,dx = 1$.
- In such situations, it is natural to expect that:
    - in general, we will observe generic functions with this property,
    - e.g., functions which are random with respect to some reasonable measure on the set of all functions.

170. Empirical Distributions (cont-d)

- The first such measure was Wiener measure, corresponding to random walk.
- Later, many other random measures have been proposed.
- In most of these random measures, almost all functions are truly random, similar to random walk.
- They are very "wiggly", they have infinitely many local maxima and minima.
- In probabilistic terms, we expect the empirical probability density functions to be multi-modal.

171. Empirical Distributions Are Mostly Few-Modal

- In reality, empirical distributions are mostly either unimodal, or bimodal, or – in rare cases – trimodal.
- In other words, they are usually few-modal.
- Why?
- In science and engineering, the few-modality is often easy to explain.
- E.g., the distributions are normal or Gamma, or, in general, follow some theoretically justified law.
- But few-modal distributions are ubiquitous also:
  - in situations where we do not have exact equations,
  - such as econometrics.
- Why?
- In this talk, we provide a theoretical explanation for the few-modality of empirical distributions.

172. Main Idea

- Of course, the space of all possible probability density functions is infinite-dimensional.
- So to exactly describe each such function, we need to describe the values of infinitely many parameters.
- In practice, at each moment of time, we can only use finitely many parameters.
- So, we need to look into appropriate finite-dimensional families of probability density functions.
- And we need explain why functions from this appropriate family are few-modal.
- To answer this question, let us describe natural properties of such families F of distributions $\rho(c_1, \ldots, c_n.x)$.
- How do we gain the information about the distributions?

173. We Want Smoothness

- It is reasonable to require that:
  - small changes in the values of the parameters c_i and/or small changes in x
  - should lead to small changes in the probability density.
- In other words, we want the function $\rho(c_1, \ldots, c_n, x)$ to be smooth.

174. Combinining Pieces of Knowledge

- Suppose that:
  - one piece of evidence is described by a probability density function (pdf) $\rho_1(x)$, and
  - another – independent – piece of evidence – leads to pdf $\rho_2(x)$.
- If these were evidences about two different quantities x_1 and x_2, then:
  - due to independence, we would conclude that
  - the distribution of the pair (x_1, x_2) follows a product distribution $\rho_1(x_1) \cdot \rho_2(x_2)$.
- In our case, however, we know that this is the same quantity, i.e., that $x_1 = x_2$.
- Thus, to get the resulting distribution, we need to restrict the product distribution to the case $x_1 = x_2$.

175. Combinining Pieces of Knowledge (cont-d)

- In precise terms, we need to consider conditional distribution under the condition that $x_1 = x_2$.
- This means that we need to consider the distribution

$$\rho(x) = c \cdot \rho_1(x) \cdot \rho_2(x).$$

- Here c is a normalizing constant – which can be determined by the condition that $\int \rho(x)\,dx = 1$.
- Thus, it is reasonable to require that:
  - for every two distribution $\rho_1(x)$ and $\rho_2(x)$ from the desired family F,
  - their normalized product $c \cdot \rho_1(x) \cdot \rho_2(x)$ should also belongs to this family.

176. Knowledge Can Come In Parts

- Sometimes, we gain the knowledge right away.
- In many other cases, knowledge comes in small steps.
- Suppose that:
  - the resulting knowledge is described by a probability density function $\rho(x)$, and
  - it comes via several (n) independent similar pieces of knowledge,
  - each step characterized by some probability density function $\rho_1(x)$.
- Then, based on the previous subsection, we can conclude that $\rho(x) = c \cdot (\rho_1(x))^n$ for some constant c.
- So, $\rho_1(x) = c_1 \cdot (\rho(x))^{1/n}$ for an appropriate normalizing coefficient c_1.

177. Knowledge Can Come In Parts (cont-d)

- Thus, it is reasonable to require that:
  - for every distribution $\rho_1(x)$ from the desired family F and
  - for every natural number $n > 1$,
  - the normalized distribution $c_1 \cdot (\rho(x))^{1/n}$ should also belong to the family.

178. Scale- and Shift-Invariance

- The numerical value of a quantity depends:
  - on the starting point for measuring this quantity
  - and on the measuring unit.
- When we change numerical values, the expression for the probability distribution also changes.
- It is reasonable to require that:
  - if we simply change the starting point and/or the measuring unit in a distribution from the family F,
  - then we should still get a distribution from the same family.
- What if we change the starting point, i.e.,
  - we replace the original starting point
  - with a new one which is a units larger.

179. Scale- and Shift-Invariance (cont-d)

- Then in the new units $y = x - a$, the distribution:
  - described by pdf $\rho(x)$
  - will now be described by $\rho_1(y) = \rho(y + a)$.
- Similarly, we can change the measuring unit, i.e.:
  - replace the original measuring unit
  - with a new one which is λ times larger.
- Then in the new units $y = x/\lambda$, the distribution
  - described by the pdf $\rho(x)$
  - will now be described by $\rho_1(y) = \lambda \cdot \rho(\lambda \cdot y)$.

180. Definitions

- Let n be a natural number.
- By an *n-parametric family of distributions*, we mean a family $F = \{f(c_1, \ldots, c_n, x)\}_{c_1,\ldots,c_n}$ of pdfs, where:
  - the values $(c_1, \ldots, c_n)$ go over some set U, and
  - the function $f(c_1, \ldots, c_n, x)$ is continuously differentiable over the closure of this set.
- We say that a family F *allows combining knowledge* if:
  - for every two functions $\rho_1(x), \rho_2(x) \in F$,
  - there exists a real number $c > 0$ for which the product $c \cdot \rho_1(x) \cdot \rho_2(x)$ also belongs to F.

181. Definitions (cont-d)

- We say that a family F *allows partial knowledge* if:
  - for every function $\rho(x)$ from this family and for every natural number n,
  - there exists a real number $c > 0$ for which the function $c \cdot (\rho(x))^{1/n}$ also belongs to F.
- We say that a family F is *shift-invariant* if:
  - for every function $\rho(x)$ from this family and for every real number a,
  - the function $\rho(x + a)$ also belongs to F.
- We say that a family F is *scale-invariant* if:
  - for every function $\rho(x)$ from this family and for every real number $\lambda > 0$,
  - the function $\lambda \cdot \rho(\lambda \cdot x)$ also belongs to F.

182. Main Result

- **Proposition.**
  - *Let F be a shift- and scale-invariant n-parametric family that allows combining and partial knowledge.*
  - *Then, every function $\rho \in F$ has the form $\rho(x) = \exp(P(x))$ for some polynomial of degree $\le n$.*
- **Corollary.**
  - *Let F be a shift- and scale-invariant n-parametric family that allows combining and partial knowledge.*
  - *Then, every function $\rho \in F$ has no more than $n - 1$ local maxima and local minima.*
- This explain why empirical distributions are few-modal.

183. Proof of the Corollary

- Indeed, at local maxima and minima, the derivative $\rho'(x) = \exp(P(x)) \cdot P'(x)$ is equal to 0.
- This is equivalent to $P'(x) = 0$.
- The derivative $P'(x)$ is a polynomial of degree $\leq n - 1$.
- Such polynomials can have no more than $n - 1$ zeros.

184. Proof of the Main Result

- Let F be a family that satisfies all the given properties.
- To simplify the problem, let's consider a family G of all the functions $c \cdot \rho(x)$, where $c > 0$ and $\rho(x) \in F$.
- By definition, every function from the family F is also an element of G.
- To show this, it is sufficient to take $c = 1$.
- We will prove the desired form for all the function from the class G.
- This will automatically imply that all the functions from the family F also have this property.
- What is the dimension of the family G?
- I.e., how many parameters do we need to specify each function from this family?

185. Proof (cont-d)

- To describe a function from G, we need to specify:
  - the value c (1 parameter), and
  - the function $\rho(x) \in F$ – which requires n parameters.
- Thus, $n+1$ parameters are sufficient, and the dimension of the family G is $\leq n+1$.
- For the family G, allowing combining knowledge leads to a simpler property: that
  - for every two functions $f_1(x), f_2(x) \in G$
  - their product $f_1(x) \cdot f_2(x)$ also belong to G.
- Indeed, $f_i(x) \in G$ means that $f_i(x) = c_i \cdot \rho_i(x)$ for some $c_i > 0$ and $\rho_i(x) \in F$.
- Thus, the product $f(x) = f_1(x) \cdot f_2(x)$ of these functions has the from $f(x) = c_1 \cdot c_2 \cdot \rho_1(x) \cdot \rho_2(x)$.

186. Proof (cont-d)

- By the property of allowing combining knowledge, for some $c > 0$, we have $\rho_0(x) = c \cdot \rho_1(x) \cdot \rho_2(x) \in F$.
- Thus, $f(x) = \dfrac{c_1 \cdot c_2}{c} \cdot (c \cdot \rho_1(x) \cdot \rho_2(x)) = c_0 \cdot \rho_0(x)$, where $c_0 \stackrel{\text{def}}{=} \dfrac{c_1 \cdot c_2}{c}$.
- So indeed, $f(x) \in G$.
- Similarly, from the other properties of the family F, we can make the following conclusions:
  - that for every function $f(x) \in G$ and for every natural number n, we have $(f(x))^{1/n} \in G$;
  - that for every function $f(x) \in G$ and for every real number a, we have $f(x + a) \in G$;
  - that for every function $f(x) \in G$ and for every real number $\lambda > 0$, we have $f(\lambda \cdot x) \in G$.

187. Proof (cont-d)

- We can simplify the problem even more if:
  - instead of the family G,
  - we consider the family g of all the functions of the type $F(x) = \ln(f(x))$, where $f(x) \in G$.
- To such functions, we also add the limit functions.
- Adding limit cases does not increase the dimension, so the dimension of the family g is still $\leq n+1$.
- In terms of this new family, we need to prove that all the functions from g are polynomials of order $\leq n$.
- The fact that the family G is closed under multiplication means that the family g is closed under addition.

188. Proof (cont-d)

- The fact that the family G is closed under taking the n-th root means that:
  - the family g is closed
  - under multiplication by $1/n$ for each natural number n.
- Together with closing under addition, this means that:
  - for every two natural numbers m and n,
  - the function $\dfrac{m}{n}\cdot F(x)=\dfrac{1}{n}\cdot F(x)+\ldots+\dfrac{1}{n}\cdot F(x)$ (m times) also belongs to the family g.
- In other words, for every $F(x)\in g$ and for every rational number r, we have $r\cdot F(x)\in g$.
- Every real number is a limit of rational numbers.
- E.g., it is a limit of numbers obtained if we only keep the first N digits in the decimal or binary expansion.

189. Proof (cont-d)

- Since we added all limit cases, we can conclude that $r \cdot F(x) \in g$ for all non-negative real numbers r as well.
- One can easily show that shift- and scale-invariance properties are also satisfied for the new family:
  - that for every function $F(x) \in g$ and for every real number a, we have $F(x+a) \in g$;
  - that for every function $F(x) \in g$ and for every real number $\lambda > 0$, we have $F(\lambda \cdot x) \in g$.
- As a final simplification, we consider the family h of all the differences $d(x) = F_1(x) - F_2(x)$ between $F_i(x) \in g$.
- To describe each of the functions $F_1(x)$ and $F_2(x)$, we need $n+1$ parameters.
- So the dimension of the new family does not exceed $2 \cdot (n+1)$.

190. Proof (cont-d)

- For every function $F(x) \in g$, the function $2F(x)$ also belongs to the family g.
- So, we can conclude that the difference $F(x) = (2F(x)) - F(x)$ also belongs to the family h. Thus, $g \subseteq h$.
- The family h is also closed under addition.
- Indeed, if $d_1(x) = F_{11}(x) - F_{12}(x)$ and $d_2(x) = F_{21}(x) - F_{22}(x)$ for some $F_{ij}(x) \in g$, then

$$d_1(x) + d_2(x) = (F_{11}(x) - F_{12}(x)) + (F_{21}(x) - F_{22}(x)) =$$

$$(F_{11}(x) + F_{21}(x)) - (F_{12}(x) + F_{22}(x)).$$

- Since g is closed under addition, the sums $F_{11}(x)+F_{21}(x)$ and $F_{12}(x)+F_{22}(x)$ also belong to g.
- Thus, indeed, the sum $d_1(x)+d_2(x)$ is a difference between two functions from g and is, thus, in h.

191. Proof (cont-d)

- We can also prove that the family h is closed under multiplication by any real number c.
- Indeed, let $d(x) = F_1(x) - F_2(x)$.
- If $c > 0$, then $c \cdot d(x) = (c \cdot F_1(x)) - (c \cdot F_2(x))$, where both $c \cdot F_1(x)$ and $c \cdot F_2(x)$ belong to the family g.
- If $c < 0$, then $c \cdot F(x) = |c| \cdot F_2(x) - |c| \cdot F_1(x)$, where also $|c| \cdot F_2(x)$ and $|c| \cdot F_1(x)$ belong to the family g.
- So, the family h is closed under addition and under multiplication by any real number.
- Thus, h is a linear space.
- Let $d \leq 2n + 2$ denote the dimension of this linear space.
- Let us select a basis $e_1(x), \ldots, e_d(x)$.

192. Proof (cont-d)

- This means that all functions from the space g have the form $c_1 \cdot e_1(x) + \ldots + c_d \cdot e_d(x)$.
- We know that the family g is shift- and scale-invariant.
- Thus, we can conclude that the family h is also shift- and scale-invariant.
- Shift-invariance means that for each $d(x) \in h$ and for each real number a, we have $d(x+a) \in h$.
- In particular, this is true for the basis functions

$$e_1(x), \ldots, e_d(x).$$

- Thus, for each i and a, there exist coefficients $c_{ij}(a)$ depending on a for which

$$e_i(x+a) = c_{i1}(a) \cdot e_1(x) + \ldots + c_{id}(a) \cdot e_d(x).$$

193. Proof (cont-d)

- In particular, for each i, we can select d different values
$$x_1, \ldots, x_d.$$
- Then we get the following system of d linear equations for determining the coefficients $c_{ij}(a)$:
$$e_i(x_1 + a) = c_{i1}(a) \cdot e_1(x_1) + \ldots + c_{id}(a) \cdot e_d(x_1),$$
$$\ldots$$
$$e_i(x_d + a) = c_{i1}(a) \cdot e_1(x_d) + \ldots + c_{id}(a) \cdot e_d(x_d).$$
- Here, the coefficients $e_j(x_k)$ are constants.
- So the values $c_{ij}(a)$ are linear combinations of the right-hand sides $e_i(x_k + a)$.
- Since the functions $e_i(x)$ are differentiable, the values $c_{ij}(a)$ are also differentiable functions of a.

194. Proof (cont-d)

- So, both sides of the following equality are differentiable: $e_i(x+a) = c_{i1}(a)\cdot e_1(x) + \ldots + c_{id}(a)\cdot e_d(x)$.

- Thus, we can differentiate them with respect to a and then plug in $a = 0$.

- As a result, we get the following system of differential equations, where $C_{ij} \stackrel{\text{def}}{=} c'_{ij}(0)$:

$$e'_1(x) = C_{11}\cdot e_1(x) + \ldots + C_{1d}\cdot e_d(x),$$

$$\ldots$$

$$e'_d(x) = C_{d1}\cdot e_1(x) + \ldots + C_{dd}\cdot e_d(x),$$

- In other words, for $e_i(x)$, we get a system of linear differential equations with constant coefficients.

195. Proof (cont-d)

- It is known that each solution of such system is a linear coefficient of the functions $x^p \cdot \exp(\alpha \cdot x)$, where:
  - the value p is a natural number and
  - α is a – possible complex – eigenvalue of the matrix C_{ij}.
- Similarly, scale-invariance means that for each function $d(x) \in h$ and for each positive real number $\lambda > 0$, we have $d(\lambda \cdot x) \in h$.
- In particular, this is true for the basis functions $e_i(x)$.
- For an auxiliary variable $X \stackrel{\text{def}}{=} \ln(x)$:
  - replacing x with $\lambda \cdot x$ corresponds to
  - replacing X with $X + a$, where $a \stackrel{\text{def}}{=} \ln(\lambda)$.

196. Proof (cont-d)

- So, for the correspondingly re-scaled functions $E_i(X) \stackrel{\text{def}}{=} e_i(\exp(X))$, we conclude that:
  - for each such function and for each real number a,
  - the function $E_i(X+a)$ is a linear combination of functions $E_1(X)$, $\ldots$, $E_d(X)$.
- We already know, from the previous parts of this proof, that this implies that:
  - each function $E_i(X)$
  - is a linear combination of the functions $X^p \cdot \exp(\alpha \cdot X)$.
- Thus, each function $e_i(x) = E_i(\ln(x))$ is a linear combination of expressions

$$(\ln(x))^p \cdot \exp(\alpha \cdot \ln(x)) = (\ln(x))^p \cdot x^\alpha.$$

197. Proof (cont-d)

- One can see that:
  - the only possibility for a function to be represented in both forms
  - is to avoid logarithms and exponential functions altogether.
- So, $e_i(x)$ is a linear combination of the terms x^p for natural p, i.e., a polynomial.
- Thus, each function from the class g is a polynomial, as a linear combination of d polynomials $e_i(x)$.
- Since $g \subseteq h$, all functions from the class g are also polynomials.
- What is the order of these polynomials?
- Let D be the order of a polynomial $F(x)$ from the g.

198. Proof (cont-d)

- For a polynomial of order D, in general, $F(x)$, $F(x+h)$, $F(x+2h)$, $\ldots$, $F(x+D\cdot h)$ are linearly independent.
- Indeed, for $h \to 0$, this is equivalent to linear independence of x^D, x^{D-1}, $\ldots$, 1.
- Thus, in the generic case, the corresponding determinant is different from 0.
- Since we have $D+1$ independent functions, thus, the family g has dimension $D+1$.
- But we know that the dimension of this family is $\leq n+1$.
- From $D+1 \leq n+1$, we conclude that $D \leq n$.
- Thus, all functions $F(x) = \ln(f(x))$ from the class g are polynomials of order $\leq n$.

199. Proof (cont-d)

- Thus, all functions $F(x) = \ln(f(x))$ from the class g are polynomials of order $\le n$.
- Hence, each function $f(x) = \exp(F(x))$ from the class F has the desired form.
- The proposition is proven.

Part VI

Why a Classification Based on Linear Approximation to Dynamical Systems Often Works Well in Nonlinear Cases

200. Dynamical Systems Are Ubiquitous

- We want to describe the state of a real-life system at any given moment of time.
- So, we need to know the values $x = (x_1, \ldots, x_n)$ of all the quantities that characterize this system.
- For example:
  - to describe the state of a mechanical system consisting of several pointwise objects,
  - we need to know the position and velocities of all these objects.
- To describe the state of an electric circuit, we need to know the currents and voltages, etc.

201. Dynamical Systems Are Ubiquitous (cont-d)

- In many real-life situation, the corresponding systems are deterministic – in the sense that:
  - the future states of the system
  - are uniquely determined by its current state.
- Sometimes, to make the system deterministic:
  - we need to enlarge its description
  - so that it incorporates all the objects that affect its dynamics.
- For example:
  - the system consisting of Earth and Moon is not deterministic in its original form,
  - since the Sun affects its dynamics.
- However, once we add the Sun, we get a system with a deterministic behavior.

202. Dynamical Systems Are Ubiquitous (cont-d)

- That the future dynamics of the system is uniquely determined by its current state means, in particular:
  - that the rate $\dot{x}$ with which the system changes is also uniquely determined by its current state,
  - i.e., that we have $\dot{x} = f(x)$, for some function $f(x)$.
- This equation can be described coordinate-wise, as

$$\dot{x}_i = f_i(x_1, \ldots, x_n).$$

- Systems that satisfy such equations are known as *dynamical systems*.

203. Simplest Case: Linear Systems

- The simplest case is when the rate of change $f_i(x_1, \ldots, x_n)$ of each variables is a linear function, i.e., when

$$\dot{x}_i = a_{i0} + \sum_{j=1}^{n} a_{ij} \cdot x_j.$$

- In almost all such cases, the matrix a_{ij} is non-degenerate.
- Then, we can select constants s_i so that:
  - for the shifted variables $y_i = x_i + s_i$,
  - the system gets a simpler form $\dot{y}_i = \sum\limits_{j=1}^{n} a_{ij} \cdot y_j$.
- Indeed, substituting $x_i = y_i - s_i$ into the above formula, and taking into account that $\dot{y}_i = \dot{x}_i$, we conclude that

$$\dot{x}_i = a_{i0} + \sum_{j=1}^{n} a_{ij} \cdot (y_j - s_j) = a_{i0} + \sum_{j=1}^{n} a_{ij} \cdot y_j - \sum_{j=1}^{n} a_{ij} \cdot s_j.$$

204. Simplest Case: Linear Systems (cont-d)

- Thus, if we select the value s_j for which $a_{i0} = \sum\limits_{j=1}^{n} a_{ij} \cdot s_j$ for each i, we will indeed get the desired formula.
- For the linear equation, the general solution is known: it is a linear combination of expressions $t^k \cdot \exp(\lambda \cdot t)$:
  - where λ is an eigenvalue of the matrix $\|a_{ij}\|$ – which is, in general, a complex number $\lambda = a + \mathrm{i} \cdot b$,
  - and k is a natural number which does not exceed the multiplicity of this eigenvalue.
- In real-number terms, we get a linear combination of the expressions $t^k \cdot \exp(a \cdot t) \cdot \sin(b \cdot t + \varphi)$.
- Depending on the values of λ, we have the following types of behavior.

205. Simplest Case: Linear Systems (cont-d)

- When $a < 0$ for all the eigenvalues, then the system is *stable*:
  - no matter what state we start with,
  - it asymptotically tends to the state
$$y_1 = \ldots = y_n = 0.$$
- When $a > 0$ for at least one eigenvalue, then the system is *unstable*.
- In this case, the deviation from the 0 state exponentially grows with time.
- When $a = 0$ and $b \neq 0$, we get an *oscillatory behavior*.

206. Simplest Case: Linear Systems (cont-d)

- When $a = b = 0$, we get a *transitional behavior*, when a system:
  - linearly (or quadratically etc.) moves
  - from one state to another.
- Interestingly, a similar classification works well for nonlinear dynamical systems as well, but why?
- In this talk, we will try to explain this fact.

207. We Need Finite-Dimensional Approximations

- We want to describe how the state $x(t) = (x_1(t), \ldots, x_n(t))$ of a dynamical system changes with time t.
- In general, the set of all possible smooth functions $x_i(t)$ is infinite-dimensional.
- In other words, we need infinitely many parameters to describe it.
- However, in practice, at any given moment, we can only have finitely many parameters.
- Thus, it is reasonable to look for finite-parametric approximations.

208. Finite-Dimensional Approximations (cont-d)

- A natural idea is:
  - to fix some smooth functions $e_k(t) = (e_{k1}(t), \ldots, e_{kn}(t))$, $1 \le k \le K$, and
  - consider linear combinations

$$x(t) = \sum_{k=1}^{K} c_k \cdot e_k(t).$$

209. Shift-Invariance

- For dynamical systems, there is no fixed moment of time.
- The equations remain the same:
  - if we change the starting point for measuring time,
  - i.e., if we replace the original temporal variable t with the new variable $t' = t + t_0$.
- It is therefore reasonable to require that:
  - the approximating family be invariant
  - with respect to the same transformation.
- In other words, we require that all shifted functions $e_k(t+t_0)$ can also be represented in the same form.
- Let us show that this reasonable requirement explains the above phenomenon.

210. Towards the Explanation

- Reminder: for each i, we have $x_i(t) = \sum\limits_{k=1}^{K} c_k \cdot e_{ki}(t)$.
- The fact that shifted functions can be represented in this form means that for each k, i, and t_0, we have

$$e_{ki}(t + t_0) = \sum_{\ell=1}^{K} c_{k\ell i}(t_0) \cdot e_{\ell i}(t), \text{ for some coefficients } c_{k\ell i}(t_0).$$

- Let us fix i and k and select K different moments of time t_m, $m = 1, \ldots, K$.
- For these moments of time, we get:

$$e_{ki}(t_m + t_0) = \sum_{\ell=1}^{K} c_{k\ell i}(t_0) \cdot e_{\ell i}(t_m).$$

- Thus, we get K linear equations for determining K unknowns $c_{k1i}(t_0)$, $\ldots$, $c_{kKi}(t_0)$.

211. Towards the Explanation (cont-d)

- Cramer's formula:
  - describes the solution to a system of linear equations
  - as a rational (and thus, smooth) function of its coefficients and right-hand sides.
- Thus, each coefficient $c_{k\ell i}(t_0)$ is a smooth function of the values $e_{ki}(t_m + t_0)$ and $e_{\ell i}(t_m)$.
- Since the functions $e_{ki}(t)$ are smooth, the dependence of the coefficients $c_{k\ell i}(t_0)$ on t_0 is also differentiable.
- All the functions involved in the formula $x_i(t) = \sum\limits_{k=1}^{K} c_k \cdot e_{ki}(t)$ are differentiable.
- So we can differentiate this formula with respect to t_0 and get $\dot{e}_{ki}(t + t_0) = \sum\limits_{\ell=1}^{K} \dot{c}_{k\ell i}(t_0) \cdot e_{\ell i}(t)$.

212. Towards the Explanation (cont-d)

- In particular, for $t_0 = 0$, we get

$$\dot e_{ki}(t) = \sum_{\ell=1}^{K} a_{k\ell i} \cdot e_{\ell i}(t), \text{ where } a_{k\ell i} \stackrel{\text{def}}{=} \dot c_{k\ell i}(t_0).$$

- So, the functions $e_{ki}(t)$ satisfy the system of linear differential equations with constant coefficients.

- We have already mentioned that:
  - the solutions to such systems
  - are exactly the functions leading to a known classification of linear dynamical system behaviors.

- This explains why for nonlinear systems, we also naturally observe similar types of behavior.

Part VII

When Revolutions Happen: Algebraic Explanation

213. When Revolutions Happen

- People usually believe that revolutions happen when life under the old regime becomes intolerable.
- However, a historical analysis shows that the usual understanding is wrong.
- Most revolutions happen *not* when the situation is at its worst.
- They usually happen when the situation has been improving for some time and then suddenly gets worse.
- Although, by the way, it never gets as bad as it was before the improvement started.

214. How Can We Explain This?

- Experiments show that in most situations, people act rationally:
  - the more their needs are satisfied, in general,
  - the happier they are.
- So why right before the revolution:
  - when the level of living is higher (often much higher) than in the recent past,
  - people are so much less happy that they start a revolution?
- How can we explain this unexpected (and somewhat counterintuitive) behavior?

215. Traditional Decision Theory: A Brief Reminder

- In traditional decision theory, people's preferences are described by numerical values called *utilities*.
- The actions of a person are determined:
  - not just by this person's current level of satisfaction – as described by the current utility value u_0,
  - but also by the expected future utility values u_1, u_2, etc.
- If we have m dollars, we can place it in a bank and get $(1+\alpha)^t \cdot m$ at time t, where α is the interest rate.
- Thus, \$1 at time t is equivalent to q^t dollars now, where $q \stackrel{\text{def}}{=} \dfrac{1}{1+\alpha}$.
- So, if we get m_0 now, m_1 in the next year, etc., this is equivalent to getting the following amount now:

$$m_0 + q \cdot m_1 + q^2 \cdot m_2 + \ldots$$

216. This General Approach Requires Extrapolation

- The future amounts are based on extrapolation:
  - we select a family of functions characterized by a few parameters $u_t = f(p_1, \ldots, p_n, t)$,
  - then we find the values $\widehat{p}_1, \ldots, \widehat{p}_n$ of the parameters that best fit the observed data u_0, u_{-1}, u_{-2}, etc.,
  - and then we use these values to predict future values as $f(\widehat{p}_1, \ldots, \widehat{p}_n, t)$.
- Let's use models that linearly depend on p_i:
  - then, matching parameters to data means easy-to-solve solving systems of linear equations,
  - while solving systems of nonlinear equations is, in general, NP-hard.
- Thus, we consider models $u_t = \sum\limits_{i=1}^{n} p_i \cdot f_i(t)$, where $f_i(t)$ are given functions, and p_i are parameters.

217. Which Basis Functions $f_i(t)$ Should We Choose?

- Most transitions are smooth; so, it's reasonable to require that all the functions $f_i(t)$ are smooth.
- Another reasonable requirement is related to the fact that the numerical value of time depends:
  - on the choice of a measuring unit – years or months,
  - and on the choice of a starting time.
- If we change a measuring unit by a new one which is a times smaller, then $t \to a \cdot t$.
- If we replace the original starting point with the new one, b units in the past $t \to t + b$.
- The general formulas for extrapolation should not depend on such an arbitrary things as:
  - selecting a unit of time or
  - selecting a starting point.

218. Choosing $f_i(t)$ (cont-d)

- It is therefore reasonable to assume that the approximating family $\left\{\sum\limits_{i=1}^n p_i\cdot f_i(t)\right\}$ will not change:

$$\left\{\sum_{i=1}^n p_i\cdot f_i(a\cdot t)\right\}_{p_1,\ldots,p_n}=\left\{\sum_{i=1}^n p_i\cdot f_i(t+b)\right\}_{p_1,\ldots,p_n}=$$

$$\left\{\sum_{i=1}^n p_i\cdot f_i(t)\right\}_{p_1,\ldots,p_n}.$$

- It turns out that under these conditions, all the basic functions are polynomials.

- So, all their linear combinations are polynomials.

- Thus, it is reasonable to approximate the actual history by a polynomial.

219. Two Simple Situations

- We will compare two simple situations:
  - a situation in which the level of living is consistently bad $u_0 = u_{-1} = \ldots = u_{-k} = \ldots = c_1$ for small c_1;
  - a situation in which the level of living used to be much better, but now somewhat decreased:

$$u_{-1} = u_{-2} = \ldots = c_+ \text{ but } u_0 = c_- < c_+.$$

- In the first situation, of course, a reasonable extrapolation should lead to the exact same small value $u_0 = c$.
- Thus, the overall utility is equal to

$$u_0 + q \cdot u_1 + \ldots = c \cdot (1 + q + q^2 + \ldots) = \frac{c}{1 - q}.$$

- But what to expect in the second situation?
- Let us start our analysis with the simplest possible linear extrapolation.

220. Linear Extrapolation

- In this case, we make our future predictions based only on two utility values: u_0 and u_{-1}.
- Since $u_0 < u_{-1}$, we get a linear decreasing function.
- Its values tend to $-\infty$ as the time t increases.
- So, when q is close to 1, the corresponding value

$$u_0 + q \cdot u_1 + \ldots \approx u_0 + u_1 + \ldots \text{ becomes negative.}$$

- This explains why in the second situation, the revolution is much more probable.

221. What About More Realistic Approximations?

- One may think that the above explanation is caused by our oversimplification of the extrapolation model.
- Of course, linear extrapolation is a very crude and oversimplified idea.
- What happens if we use higher degree polynomials for extrapolation?
- Let us assume that for extrapolation, we use polynomials of order d.
- The corresponding family of polynomials have d_1 parameters, so we can fit $d+1$ values $u_0, u_{-1}, \ldots, u_{-d}$.
- Let us find the polynomial $P(t)$ of degree d that fits these values: $P(0) = c_-$, $P(-1) = \ldots = P(-d) = c_+$.

222. Realistic Approximations (cont-d)

- For $Q(t) \stackrel{\text{def}}{=} P(t) - c_+$, we have $Q(-d) = \ldots = Q(-1) = 0$ and $Q(0) = c_- - c_+$.

- This polynomial of degree d has d roots $t = -1, \ldots, t = -d$, so $Q(t) = C \cdot (t+1) \cdot (t+2) \cdot \ldots \cdot (t+d)$, and
$$Q(t) = c_+ + (c_- - c_+) \cdot \frac{(t+1) \cdot (t+2) \cdot \ldots \cdot (t+d)}{1 \cdot 2 \cdot \ldots \cdot d}.$$

- Since $c_- < c_+$, this value is negative – and tends to $-\infty$ as the time t increases.

- In comparison with the linear extrapolation case, it tends to $-\infty$ even faster: as t^d.

- So, *the revolution phenomenon can be explained* no matter what degree of extrapolation we use.

223. Discussion

- We have explained the seemingly counterintuitive revolution phenomenon.
- Based on our analysis, we can make auxiliary conclusions (which also fit well with common sense).
- Revolutions only happen if people care about the future.
- If they don't, if $q \approx 0$, people are happy with their present-day level of living.
- The more into the past the people go in their analysis, the more probable it is that they will revolt.
- People who do not know their history are less prone to revolutions than people who do.

Part VIII
How Mathematics and Computing Can Help Fight the Pandemic: Two Pedagogical Examples

224. First Example: Need for Social Distancing

- This problem is related to the pandemic-related need to observe a social distance of at least 2 meters (6 feet).
- Two persons are on two sides of a narrow-walkway street, waiting for the green light.
- They start walking from both sides simultaneously.
- For simplicity, let us assume that they walk with the same speed.
- If they follow the shortest distance path – i.e., a straight line AB – they will meet in the middle.
- This is not good; so one of them should move somewhat to the left, another should move somewhat to the right.
- At all moments of time, they should be at least 2 meters away from each other.
- What is the fastest way for them to do it?

225. Formulating This Problem in Precise Terms

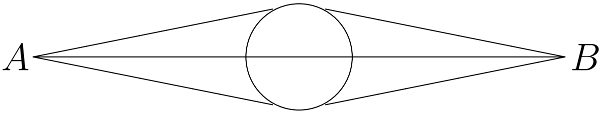
- The situation is absolutely symmetric with respect to the reflection in the midpoint M of the segment AB.
- So, it is reasonable to require that:
  - the trajectory of the 2nd person can be obtained from the trajectory of the 1st person
  - by this reflection.
- Thus, at any given moment of time, the midpoint M is the midpoint between the two persons.
- In these terms:
  - the requirement that they are separated by $\geq 2m$
  - means that each of them should always be at a distance at least 1 meter from the midpoint M.
- In other words, both trajectories should avoid the disk of radius 1 meter with a center at the midpoint M.

226. Formulating the Problem (cont-d)

- We want the fastest possible trajectory.
- The speed is assumed to be constant.
- So, they should follow the shortest possible trajectory.
- In other words, we need to find:
  - the shortest possible trajectory going from point A to point B
  - that avoids the disk centered at the midpoint M of the segment AB.

227. Solution

- To get the shortest path, outside the disk, the trajectory should be straight.
- Where it touches the circle, it should be smooth.
- Thus, the solution is as follows:
  - first, we follow a straight line until it touches the circle as a tangent,
  - then, we follow the circle,
  - and finally, we follow the straight line again – which again starts as a tangent to the circle:



228. Second Example: Need for Fast Testing

- One of the challenges related to the COVID-19 pandemic is that:
  - this disease has an unusually long incubation period,
  - about 2 weeks.
- As a result, people with no symptoms may be carrying the virus and infecting others.
- As of now, the only way to prevent such infection is to perform massive testing of the population.
- The problem is that there is not enough test kits to test everyone.

229. What Was Proposed

- To solve this problem, researchers proposed the following idea:
  - instead of testing everyone individually,
  - why not combine material from a group of several people, and
  - test each combined sample by using a single test kit.
- If no viruses are detected in the combined sample, this means that all the people from the corresponding group are virus-free.
- So there is no need to test them again.
- After this, we need to individually test only folks from the groups that showed the presence of the virus.

230. Resulting Problem

- Suppose that we need to test a large population of N people.
- Based on the previous testing, we know the proportion p of those who have the virus.
- In accordance with the above idea, we divide N people into groups.
- The question is: what should be the size s of each group?
- If the size is too small, we are still using too many test kits.

231. Resulting Problem (cont-d)

- If the size is too big, then:
  - every group, with a high probability, has a sick person,
  - so we are not dismissing any people after such testing, and thus, we are not saving any kits at all.
- So what is the optimal size of the group?
- Of course, this is a simplified formulation.
- It does not take into account that:
  - for large group sizes s, when each individual testing material is diluted too much,
  - tests may not be able to detect infected individuals.

232. Let Us Formulate This Problem in Precise Terms

- If we divide N people into groups of s persons each, we thus get N/s groups.
- The probability that a person is virus-free is $1 - p$.
- Thus, the probability that all s people from a group are virus-free is $(1 - p)^s$.
- So, out of N/s groups, the number of virus-free groups is $(1 - p)^s \cdot (N/s)$.
- Each of these groups has s people.
- So the overall number of tested people can be obtained by multiplying the number of virus-free groups by s.
- This results in $(1 - p)^s \cdot N$.

233. Formulation and Solution

- For the remaining $N - (1-p)^s \cdot N$ folks, we need individual testing.
- So, the overall number of needed test kits is

$$N_t = \frac{N}{s} + N - (1-p)^s \cdot N.$$

- We want to minimize the number of test kits.
- So, we want to find the group size s for which this number is the smallest possible.
- Equating the derivative to 0 and dividing both sides of this equation by N, we get:

$$-\frac{1}{s^2} - (1-p)^s \cdot \ln(1-p) = 0.$$

234. Solution (cont-d)

- For small p, we have $(1-p)^s \approx 1$ and $\ln(1-p) \approx -p$, so $-\dfrac{1}{s^2}+p \approx 0$, and $s \approx \dfrac{1}{\sqrt{p}}$.
- For example:
  - for $p = 1\%$, we have $s \approx 10$;
  - for $p = 0.1\%$, we get $s \approx 30$; and
  - for $p = 0.01\%$, we get $s \approx 100$.
- The resulting number of tests N_t can also be approximately estimated.
- When the group size s is described by the approximate formula, we have $\dfrac{N}{s} \approx \sqrt{p} \cdot N$.
- If we take into account that $(1-p)^s \approx 1 - p \cdot s$, then

$$N - (1-p)^s \cdot N \approx p \cdot s \cdot N \approx \sqrt{p} \cdot N.$$

235. Solution (cont-d)

- Thus, we get $N_t \approx \sqrt{p} \cdot N$.
- For example:
  - for $p = 1\%$, we need 10 times fewer test kits than for individual testing;
  - for $p = 0.1\%$, we need 30 times fewer test kits; and
  - for $p = 0.01\%$, we need 100 times fewer test kits.

- T. S. Perry, "Researchers are using algorithms to tackle the coronavirus test shortage: the scramble to develop new test kits that deliver faster results", *IEEE Spectrum*, 2020, Vol. 57, No. 6, p. 4.

- N. Shental, S. Levy, V. Wuvshet, S. Skorniakov, Y. Shemer-Avni, A. Porgador, and T. Hertz, *Efficient High Throughput SARS-CoV-2 Testing to Detect Asymptomatic Carriers*, medRxiv preprint https://doi.org/10.1101/2020.04.14.20064618, posted on April 20, 2020.

Part IX

Which Distributions (or Families of Distributions) Best Represent

Interval Uncertainty: Case of Permutation-Invariant Criteria

237. Interval Uncertainty Is Ubiquitous

- An engineering designs comes with numerical values of the corresponding quantities, be it:
  - the height of ceiling in civil engineering or
  - the resistance of a certain resistor in electrical engineering.
- Of course, in practice, it is not realistic to maintain the exact values of all these quantities.
- We can only maintain them with some tolerance.
- As a result, the engineers:
  - not only produce the desired ("nominal") value x of the corresponding quantity,
  - they also provide positive and negative tolerances $\varepsilon_+ > 0$ and $\varepsilon_- > 0$.

238. Interval Uncertainty Is Ubiquitous (cont-d)

- The actual value must be in the interval $\mathbf{x} = [\underline{x}, \overline{x}]$, where $\underline{x} \stackrel{\text{def}}{=} x - \varepsilon_-$ and $\overline{x} \stackrel{\text{def}}{=} x + \varepsilon_+$.
- All the manufacturers need to do is to follow these interval recommendations.
- There is no special restriction on probabilities of different values within these intervals.
- These probabilities depends on the manufacturer.
- Even for the same manufacturer, they may change when the manufacturing process changes.

239. Data Processing Under Interval Uncertainty Is Often Difficult

- Interval uncertainty is ubiquitous.
- So, many researchers have considered different data processing problems under this uncertainty.
- This research area is known as *interval computations*.
- The problem is that the corresponding computational problems are often very complex.
- They are much more complex than solving similar problems under *probabilistic* uncertainty:
  - when we know the probabilities of different values within the corresponding intervals,
  - we can use Monte-Carlo simulations to gauge the uncertainty of data processing results.

240. Interval Data Processing Is Difficult (cont-d)

- A similar problem for interval uncertainty:
  - is NP-hard already for the simplest nonlinear case
  - when the whole data processing means computing the value of a quadratic function.
- It is even NP-hard to find the range of variance when inputs are known with interval uncertainty.
- This complexity is easy to understand.
- Interval uncertainty means that we may have different probability distributions on the given interval.
- So, to get guaranteed estimates, we need, in effect, to consider all possible distributions.
- And this leads to very time-consuming computations.
- For some problems, this time can be sped up, but in general, the problems remain difficult.

241. It Is Desirable to Have a Family of Distributions Representing Interval Uncertainty

- Interval computation problems are NP-hard.
- In practical terms, this means that the corresponding computations will take forever.
- So, we cannot consider *all* possible distributions on the interval.
- A natural idea is to consider *some* typical distributions.
- This can be a finite-dimensional family of distributions.
- This can be even a finite set of distributions – or even a single distribution.
- For example, in measurements, practitioners often use uniform distributions on the corresponding interval.
- This selection is even incorporated in some international standards for processing measurement results.

242. Family of Distributions (cont-d)

- Of course, we need to be very careful which family we choose.
- By limiting the class of possible distributions, we introduce an artificial "knowledge".
- Thus, we modify the data processing results.
- So, we should select the family depending on what characteristic we want to estimate.
- We need to beware that:
  - a family that works perfectly well for one characteristic
  - may produce a completely misleading result when applied to some other desired characteristic.
- Examples of such misleading results are well known.

243. Continuous Vs. Discrete Distributions

- Usually, in statistics and in measurement theory:
  - when we say that the actual value x belongs to the interval $[a, b]$,
  - we assume that x can take any real value between a and b.
- However, in practice:
  - even with the best possible measuring instruments,
  - we can only measure the value of the physical quantity x with some uncertainty h.
- Thus, from the practical viewpoint, it does not make any sense to distinguish between a and $a + h$.
- Even with the best measuring instruments, we will not be able to detect this difference.

244. Continuous Vs. Discrete (cont-d)

- From the practical viewpoint, it makes sense to divide the interval $[a, b]$ into small subintervals
$$[a, a+h], [a+h, a+2h], \ldots$$
- Within each of them the values of x are practically indistinguishable.
- It is sufficient to find the probabilities $p_1, p_2, \ldots, p_n$ that the actual value x is in one of the subintervals:
  - the probability p_1 that x is in the first small subinterval $[a, a+h]$;
  - the probability p_2 that x is in the first small subinterval $[a+h, a+2h]$; etc.
- These probabilities should, of course, add up to 1:
$$\sum_{i=1}^{n} p_i = 1.$$

245. Continuous Vs. Discrete (cont-d)

- In the ideal case, we get more and more accurate measuring instruments – i.e., $h \to 0$.
- Then, the corresponding discrete probability distributions will tend to continuous ones.
- So, from this viewpoint:
  - selecting a probability distribution means selecting a tuple of values $p = (p_1, \ldots, p_n)$, and
  - selecting a family of probability distributions means selecting a family of such tuples.

246. Example: Estimating Maximum Entropy

- Whenever we have uncertainty, a natural idea is to provide a numerical estimate for this uncertainty.
- It is known that one of the natural measures of uncertainty is Shannon's entropy $-\sum\limits_{i=1}^{n} p_i \cdot \log_2(p_i)$.
- In the case of interval uncertainty, we can have several different tuples.
- In general, for different tuples, entropy is different.
- As a measure of uncertainty of the situation, it is reasonable to take the largest possible value.
- Indeed, Shannon's entropy can be defined as:
  - the average number of binary ("yes"-"no") questions
  - that are needed to uniquely determine the situation.

247. Maximum Entropy (cont-d)

- The larger this number, the larger the initial uncertainty.
- Thus, it is natural to take the largest number of such questions as a characteristic of interval uncertainty.
- For this characteristic, we want to select a distribution:
  - whose entropy is equal to
  - the largest possible entropy of all possible probability distributions on the interval.
- Selecting such a "most uncertain" distribution is known as the *Maximum Entropy approach*.
- This approach has been successfully used in many practical applications.

248. Maximum Entropy (cont-d)

- It is well known that:
  - out of all possible tuples with $\sum\limits_{i=1}^{n} p_i = 1$,
  - the entropy is the largest possible when all the probabilities are equal to each other, i.e., when

$$p_1 = \ldots = p_n = 1/n.$$

- In the limit $h \to 0$, such distributions tend to the uniform distribution on the interval $[a, b]$.
- This is one of the reasons why uniform distributions are recommended in some measurement standards.

249. Modification of This Example

- In addition to Shannon's entropy, there are other measures of uncertainty.
- They are usually called *generalized entropy*.
- For example, in many applications, practitioners use the quantity $-\sum\limits_{i=1}^{n} p_i^\alpha$ for some $\alpha \in (0,1)$.
- It is known that when $\alpha \to 0$, this quantity, in some reasonable sense, tends to Shannon's entropy.
- To be more precise:
  - the tuple at which the generalized entropy attains its maximum under different condition
  - tends to the tuple at which Shannon's entropy attains its maximum.
- The maximum of this characteristic is also attained when all the probabilities p_i are equal to each other.

250. Other Examples and Idea

- A recent paper analyzed how to estimate sensitivity of Bayesian networks under interval uncertainty.
- It also turned out that;
  - if we limit ourselves to a single distribution,
  - then the most adequate result also appears if we select a uniform distribution.
- The same uniform distribution appears in many different situations, under different optimality criteria.
- This makes us think that there must be a general reason for this distribution.
- In this talk, we indeed show that there is such a reason.

251. Beyond the Uniform Distribution

- For other characteristics, other possible distributions provide a better estimate. For example:
  - if we want to estimate the *smallest* possible value of the entropy,
  - then the corresponding optimal value 0 is attained for several different distributions.
- Specifically, there are n such distributions corresponding to different values $i_0 = 1, \ldots, n$.
- In each of these distributions, we have $p_{i_0} = 1$ and $p_i = 0$ for all $i \neq i_0$.
- In the continuous case $h \to 0$:
  - these probability distributions correspond to point-wise probability distributions
  - in which a certain value x_0 appears with probability 1.

252. Beyond the Uniform Distribution (cont-d)

- Similar distributions appear for several other optimality criteria.
- For example, when we minimize generalized entropy.
- How can we explain that these distributions appear as solutions to different optimization problems?
- Similar to the uniform case, there should also be a general explanation.
- A simple general explanation will indeed be provided in this talk.

253. Let Us Use Symmetries

- In general, our knowledge is based on *symmetries*, i.e., on the fact that some situations are similar.
- Indeed, if all the world's situations were completely different, we would not be able to make any predictions.
- Luckily, real-life situations have many features in common.
- So we can use the experience of previous situations to predict future ones.
- For example, when a person drops a pen, it starts falling down with the acceleration of 9.81 m/sec^2.
- If this person moves to a different location, he or she will get the exact same result.
- This means that the corresponding physics is invariant with respect to shifts in space.

254. Let Us Use Symmetries (cont-d)

- Similarly, if the person repeats this experiment in a year, the result will be the same.
- This means that the corresponding physics is invariant with respect to shifts in time.
- Alternatively, if the person turns around a little bit, the result will still be the same.
- This means that the underlying physics is also invariant with respect to rotations, etc.
- This is a very simple example, but such symmetries are invariances are actively used in modern physics.

255. Let Us Use Symmetries (cont-d)

- Moreover, many previously proposed fundamental physical theories can be derived from symmetries:
  - Maxwell's equations that describe electrodynamics,
  - Schroedinger's equations that describe quantum phenomena,
  - Einstein's General Relativity equation that describe gravity.
- Symmetries also help to explain many empirical phenomena in computing.
- From this viewpoint:
  - a natural way to look for what the two examples have in common
  - is to look for invariances that they have in common.

256. Permutations – Natural Symmetries in the Entropy Example

- We have n probabilities $p_1, \ldots, p_n$.
- What can we do with them that would preserve the entropy?
- The easiest possible transformations is when we do not change the values themselves, just swap them.
- Bingo! Under such swap, the value of the entropy does not change.
- Interestingly, the above-described generalized entropy is also permutation-invariant.
- Thus, we are ready to present our general results.

257. Definitions and Results

- We say that a function $f(p_1, \ldots, p_n)$ is *permutation-invariant* if for every permutation, we have
$$f(p_1, \ldots, p_n) = f(p_{\pi(1)}, \ldots, p_{\pi(n)}).$$
- By a *permutation-invariant optimization problem*, we mean a problem of optimizing:
  - a permutation-invariant function $f(p_1, \ldots, p_n)$
  - under constraints of the type $g_i(p_1, \ldots, p_n) = a_i$ or $h_j(p_1, \ldots, p_n) \geq b_j$
  - for permutation-invariant functions g_i and h_j.
- **Proposition.** *If a permutation-invariant optimization problem has only one solution, then for this solution:*
$$p_1 = \ldots = p_n.$$
- This explains why we get the uniform distribution in several cases (maximum entropy etc.)

258. Proof

- We will prove this result by contradiction.
- Suppose that the values p_i are not all equal.
- This means that there exist i and j for which $p_i \neq p_j$.
- Let us swap p_i and p_j, and denote the corresponding values by p'_i, i.e.:
  - we have $p'_i = p_j$,
  - we have $p'_j = p_i$, and
  - we have $p'_k = p_k$ for all other k.
- The values p_i satisfy all the constraints.
- All the constraints are permutation-invariant.
- So, the new values p'_i also satisfy all the constraints.
- Since the objective function is permutation-invariant, we have $f(p_1, \ldots, p_n) = f(p'_1, \ldots, p'_n)$.

259. Proof (cont-d)

- Since the values $(p_1, \ldots, p_n)$ were optimal, the values $(p'_1, \ldots, p'_n) \neq (p_1, \ldots, p_n)$ are thus also optimal.

- This contradicts to the assumption that the original problem has only one solution.

- This contradiction proves for the optimal tuple $(p_1, \ldots, p_n)$ that all the values p_i are indeed equal to each other.

- The proposition is proven.

260. Discussion

- What if the optimal solution is not unique?
- We can have a case when we have a small finite number of solutions.
- We can also have a case when we have a 1-parametric family of solutions – depending on one parameter.
- In our discretized formulation, each parameter has n values, so this means that we have n possible solutions.
- Similarly, a 2-parametric family means that we have n^2 possible solutions, etc.
- We say that a problem has a *small finite number of solutions* if it has $< n$ solutions.
- We say that a problem has a *d-parametric family of solutions* if it has $\le n^d$ solutions.

261. Second Result

- **Proposition.**
  - *If a permutation-invariant optimization problem has a small finite number of solutions,*
  - *then it has only one solution.*
- Due to Proposition 1, in this case, the only solution is the uniform distribution $p_1 = \ldots = p_n$.

262. Proof

- Since $\sum p_i = 1$:
  - there is only one possible solution for which
  $$p_1 = \ldots = p_n:$$
  - the solution for which
  $$p_1 = \ldots = p_n = 1/n.$$
- Thus, if the problem has more than one solution, some values p_i are different from others.
- In particular, some values are different from p_1.
- Let S denote the set of all j for which $p_j = p_1$.
- Let m denote the number of elements in this set.
- Since some values p_i are different from p_1, we have
$$1 \leq m \leq n - 1.$$

263. Proof (cont-d)

- Due to permutation-invariance, each permutation of this solution is also a solution.

- For each m-size subset of $\{1,\ldots,n\}$, we can have a permutation that transforms S into this set.

- Thus, it produces a new solution to the original problem.

- There are $\binom{n}{m}$ such subsets.

- For $0 < m < n$, the smallest value n of $\binom{n}{m}$ is attained when $m = 1$ or $m = n - 1$.

- Thus, if there is more than one solution, we have at least n different solutions.

- Since we assumed that we have fewer than n solutions, this means that we have only one. Q.E.D.

264. One More Result

- **Proposition.** *If a permutation-invariant optimization problem has a 1-parametric family of solutions, then:*
  - *this family of solutions is characterized by a real number* $c \le 1/(n-1)$, *for which*
  - *all these solutions have the following form:* $p_i = c$ *for* $i \ne i_0$ *and* $p_{i_0} = 1 - (n-1) \cdot c$.
- In particular, for $c = 0$:
  - we get the above-mentioned 1-parametric family of distributions for which
  - Shannon's entropy (or generalized entropy) attain the smallest possible value.

265. Proof

- We have shown that:
  - if in one of the solutions, for some value p_i we have m different indices j with this value,
  - then we will have at least $\binom{n}{m}$ different solutions.
- For all m from 2 to $n-2$, this number is at least as large as $\binom{n}{2} = \dfrac{n \cdot (n-1)}{2}$ and is, thus, larger than n.
- Since overall, we only have n solutions, this means that it is not possible to have $2 \le m \le n-2$.
- So, the only possible values of m are 1 and $n-1$.

266. Proof (cont-d)

- If there was no group with $n-1$ values:
  - this would means that all the groups must have $m = 1$,
  - i.e., consist of only one value.
- In other words, in this case, all n values p_i would be different.
- In this case, each of $n!$ permutations would lead to a different solution.
- So we would have $n! > n$ solutions, but there are only n solutions.
- Thus, this case is also impossible.
- So, we do have a group of $n-1$ values with the same p_i.
- Then we get exactly one of the solutions described in the formulation.

267. Conclusions

- Traditionally, in engineering, uncertainty is described by a probability distribution.
- In practice, we rarely know the exact distribution.
- In many practical situations:
  - the only information we know about a quantity
  - is the interval of possible values of this quantity.
- And we have no information about the probability of different values within this interval.
- Under such interval uncertainty, we cannot exclude any mathematically possible probability distribution; so:
  - to estimate the range of possible values of the desired uncertainty characteristic,
  - we must, in effect, consider all possible distributions.

268. Conclusions (cont-d)

- Not surprisingly, for many characteristics, the corresponding computational problem becomes NP-hard.

- For some characteristics, we can provide a reasonable estimate for their desired range if:
  - instead of all possible distributions,
  - we consider only distributions from some finite-dimensional family.

- For example:
  - to estimate the largest possible value of Shannon's entropy (or of its generalizations),
  - it is sufficient to consider only the uniform distribution.

269. Conclusions (cont-d)

- Similarly:
  - to estimate the smallest possible value of Shannon's entropy or of its generalizations,
  - it is sufficient to consider point-wise distributions.
- Different optimality criteria lead to the same distribution – or to the same family of distributions.
- This made us think that there should be a general reason for the appearance of these families.
- In this talk, we show that indeed:
  - the appearance of these distributions and these families can be explained
  - by the fact that all the corresponding optimization problems are permutation-invariant.

270. Conclusions (cont-d)

- Thus, in the future, if a reader encounters a permutation-invariant optimization problem:
  - for which it is known that there is a unique solution
  - or that there is only a 1-parametric family of solutions,
  - then there is no need to actually solve the corresponding problem.
- In such situations, it is possible to simply use our general symmetry-based results.
- Thus, we can find a distribution (or a family of distributions) that:
  - for the corresponding characteristic,
  - best represents interval uncertainty.

Part X

Expert Knowledge Makes Predictions More Accurate:

Theoretical Explanation of an Empirical Observation

271. Empirical Observation That Needs Explaining

- It is known that the use of expert knowledge makes predictions more accurate.
- For example, computer-based meteorological forecasts are regularly corrected by experts.
- A typical improvement is that the accuracy consistently improves by 10%.
- How can we explain this?

272. Towards an Explanation

- Use of expert knowledge means, in effect, that we combine:
  - an estimate produced by a computer model and
  - an expert estimate.
- Let σ_m and σ_e denote the standard deviations, correspondingly, of the model and of the expert estimate.
- In effect, the only information that we have about comparing the two accuracies is that
  - expert estimates are usually less accurate
  - than model results:

$$\sigma_m < \sigma_e.$$

- So, if we fix σ_e, then the only thing we know about σ_m is that σ_m is somewhere between 0 and σ_e.

273. Towards an Explanation (cont-d)

- We have no reason to assume that some values from the interval $[0, \sigma_e]$ are more probable than others.
- Thus, it makes sense to assume that all these values are equally probable.
- So, we have a uniform distribution on this interval.
- For this uniform distribution, the average value of σ_m is equal to $0.5 \cdot \sigma_e$.
- Thus, we have $\sigma_e = 2 \cdot \sigma_m$.
- In general:
  - if we combine two estimates x_m and x_e with accuracies σ_m and σ_e,
  - then the combined estimate x_c is obtained by minimizing the sum $\dfrac{(x_m - x_c)^2}{\sigma_m^2} + \dfrac{(x_e - x_c)^2}{\sigma_e^2}$.

274. Towards an Explanation (cont-d)

- The resulting estimate is $x_c = \dfrac{x_m \cdot \sigma_m^{-2} + x_e \cdot \sigma_e^{-2}}{\sigma_m^{-2} + \sigma_e^{-2}}$, with accuracy $\sigma_c^2 = \dfrac{1}{\sigma_m^{-2} + \sigma_e^{-2}}$.
- For $\sigma_e = 2\sigma_m$, we have $\sigma_e^{-2} = 0.25 \cdot \sigma_m^{-2}$, thus $\sigma_c^2 = \sigma_m^2 \cdot \dfrac{1}{1 + 0.25} = \sigma_m^2 \cdot \dfrac{1}{1.25} = 0.8 \cdot \sigma_m^2$, thus $\sigma_c \approx 0.9 \cdot \sigma_m$.
- So we indeed get a 10% increase in the resulting prediction.

275. Reference

- N. Silver, *The Signal and the Noise: Why So Many Decisions Fail – but Some Don't*, Penguin Press, New York, 2012.