

Kolmogorov-Arnold Networks: Why Piecewise Linear Activation Functions Work Best? Why Only Addition and Multiplication Work Well? Towards a Possible Explanation

Martine Ceberio¹, Christoph Lauter¹, Vladik Kreinovich¹, Olga Kosheleva²,
Christian Servin³, Jose David Diaz Roman⁴,

Boris Jesus Mederos Madrazo⁴, and Jose Manuel Mejia Muñoz⁴

^{1,2}Departments of ¹Computer Science and ²Teacher Education
University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA
{mceberio,cqlauter,vladik,olgak}@utep.edu

³Computer Science and Information Technology Systems Department
El Paso Community College, 919 Hunter, El Paso, TX 79915, USA,
cservin1@epcc.edu

⁴Universidad Autónoma de Ciudad Juárez (UACJ), Ciudad Juárez 32310, México
{david.roman,boris.mederos,jose.mejia}@uacj.mx

1. Kolmogorov-Arnold Networks (KAN): a Brief Reminder

- The last decades has shown spectacular success of neural networks.
- A neural network consists of *neurons*.
- A neuron is a computational device that transforms inputs $z_1, \dots, z_m$ into an output value $z = s(w_1 \cdot z_1 + \dots + w_m \cdot z_m + w_0)$.
- Here $s(x)$ is a function known as an *activation function*, and numbers w_i – called *weights* – are determined during training.
- Some neurons directly process the measurement results (and/or other input values) $x_1, \dots, x_n$.
- Other neurons use, as inputs, the outputs of some other neurons.
- Training means finding the smallest value of the error function.
- This value is obtained by using gradient descent – via a method known as *backpropagation*.
- The output of one of the neurons is then returns to the user as the computation result $f_{NN}(x_1, \dots, x_n)$.

2. Kolmogorov-Arnold Networks (KAN): a Brief Reminder (cont-d)

- In the usual neural networks, the weights are changed during training, but the activation function remains the same.
- In particular, most neurons in the current AI tools use the so-called Rectified Linear Units (ReLU), i.e., neurons with the activation function $s(z) = \max(0, z)$.
- It is known that in some applications, other activation functions lead to more effective results.
- To take this possibility into account, researchers have proposed to use:
 - instead of a single activation function $s(z)$,
 - a family of functions $s(z, C_1, \dots, C_n)$.
- So during training, we determine not only the weights w_i , but also the parameters C_i .

3. Kolmogorov-Arnold Networks (KAN): a Brief Reminder (cont-d)

- The values C_i are obtained by using the same backpropagation technique as finding the weights w_i .
- Such networks are known as *Kolmogorov-Arnold Networks* (KAN, for short).
- It turned out that for many practical problems, this indeed leads to more effective and/or efficient computations.

4. Computational Comment

- At first glance, introduction of additional parameters makes the problem more complicated.
- Indeed, now we have to find *both* the weights *and* the parameters describing the activation function.
- However, it is easy to show that this problem can be simplified.
- Namely, we can include weights into the list of parameters describing activation functions.
- Indeed, for most neurons, each inputs comes from some previous processing.
- This input has, thus, the form $z_i = s(t_i, C_1, \dots, C_n)$ for some t_i .
- So the contribution $Z_i \stackrel{\text{def}}{=} w_i \cdot z_i$ of this input to the linear combination z takes the form

$$Z_i = w_i \cdot s(t_i, C_1, \dots, C_n).$$

5. Computational Comment (cont-d)

- It is therefore reasonable to include w_i as one the parameters C_{n+1} , i.e., consider a family of activation functions $C_{n+1} \cdot s(z, C_1, \dots, C_n)$.
- We can similarly include the weight w_0 into one the inputs, e.g., by considering $Z_1 = w_1 \cdot s(t_i, C_1, \dots, C_n) + w_0$ and thus, the family

$$C_{n+1} \cdot s(z, C_1, \dots, C_n) + C_{n+2}.$$

- In this interpretation, each neuron simply applies its activation function to the sum $Z_1 + \dots + Z_m$ of all the inputs Z_i .

6. Biological Comment

- The original idea of artificial neurons came from the desire:
 - to simulate biological neurons and thus,
 - to borrow evolution-optimized nature's solution to the problem of data processing.
- From this viewpoint, it is worth mentioning that:
 - according to the latest biological studies,
 - some biological neurons can apply different activation functions to the their inputs,
 - exactly as Kolmogorov-Arnold neural networks.

7. Terminological Comment: Why the Name

- The name of these networks come from the theorem proved by A. N. Kolmogorov and V. Arnold.
- They proved that every continuous function of n variables can be represented as a composition of addition and functions of one variable.
- This is exactly what KAN computes – it:
 - either computes the sum,
 - or it applies some function of one variable – namely, the activation function.
- Specifically, Kolmogorov and Arnold proved an even stronger statement – that
 - every continuous function $f(x_1, \dots, x_n)$ on a bounded domain
 - can be represented as

$$f(x_1, \dots, x_n) = \sum_j f_j \left(\sum_{i=1}^n f_{i,j}(x_i) \right) \text{ for some continuous f-s } f_i(z) \text{ and } f_{i,j}(z).$$

8. First Remaining Challenge

- The original paper used splines – i.e., piecewise polynomial functions.
- It was recently shown that the effectiveness increases if we restrict ourselves to piecewise linear activation functions.
- Our own numerical experiments confirmed that:
 - piecewise linear activation functions
 - lead to more effective results than, e.g., piecewise quadratic ones.
- How can we explain this empirical fact?

9. Second Remaining Challenge

- Another interesting empirical observation is related to the fact that in the original Kolmogorov-Arnold Theorem:
 - instead of addition $a + b$,
 - we can use any commutative associative binary operation of the type $a \oplus_f b \stackrel{\text{def}}{=} f^{-1}(f(a) + f(b))$;
 - here, $f^{-1}(x)$ denotes the inverse function.
- Indeed, once we implement this operation:
 - we can compute addition as $a + b = f(f^{-1}(a) \oplus_f f^{-1}(b))$, i.e., as a superposition of this combination and functions of one variable,
 - and with addition, we can represent any continuous function.
- In particular, for $f(x) = \ln(x)$, we have $f^{-1}(x) = \exp(x)$ and thus $a \oplus_f b = \exp(\ln(a) + \ln(b)) = a \cdot b$.

10. Second Remaining Challenge (cont-d)

- If we use different functions $f(x)$, we get different operations.
- It turned out that out of all such binary operations, only multiplication has led to efficient results.
- How can we explain this?
- In this talk, we provide semi-heuristic ideas that may help to understand these empirical facts.

11. Piecewise Linear Activation Functions Work Best? Towards a Possible Explanation

- The whole idea of machine learning is as follows:
 - We know that the value of a quantity y is determined by the values of some other quantities $x_1, \dots, x_n$.
 - In mathematical terms, we know that that $y = f(x_1, \dots, x_n)$ for some function $f(x_1, \dots, x_n)$.
 - However, we do not know this function.
- All we know are the values $y^{(k)} = f(x_1^{(k)}, \dots, x_n^{(k)})$ of this function in finitely many cases $k = 1, \dots, K$.
- Based on this information, we want to find the values of this function for all other tuples $(x_1, \dots, x_n)$.
- In numerical mathematics, this is known as *interpolation*.

12. Piecewise Linear Activation Functions Work Best? Towards a Possible Explanation (cont-d)

- Similarly, based on finitely many examples, we cannot uniquely determine the family $s(z, C_1, \dots, C_n)$ of activation functions.
- So, we need to use interpolation.

13. Which Interpolation Should We Choose

- The usual criterion for accuracy in statistics and in machine learning is minimizing the mean squared deviation. So:
 - if we know the value of an activation function $s(z)$ at several points $z_1, \dots, z_K$, and
 - we want to interpolate it and come up with a function $s_0(z)$ that is defined for all z ,
- it is reasonable to select a function for which the expected value of the mean square difference is the smallest possible:

$$\int_z \int_s (s_0(z) - s(z))^2 ds dz \rightarrow \min .$$

- To make sense of this formula, we need to select some probability measure on the set of all continuous functions.

14. Which Probability Measure Should We Select

- The whole idea of KAN comes from the Kolmogorov-Arnold theorem, according to which:
 - any continuous function – in particular, any function that is not differentiable,
 - can be represented as a superposition of addition and functions of over variable.
- These functions cannot be all smooth (= differentiable) since:
 - superposition of differential functions is also differentiable (its derivative is described by the chain rule),
 - and thus, we will not be able to represent all continuous functions.
- So, we need to consider functions $s(z)$ that are not necessarily differentiable.
- What does it mean in statistical terms?

15. Which Probability Measure Should We Select (cont-d)

- Differentiable means, in effect, that for small Δz , the differences $s(z + \Delta z) - s(z)$ and $s(z) - s(z - \Delta z)$ are strongly correlated.
- Indeed, both of these differences are close to $s'(z) \cdot \Delta z$, where, as usual, s' means the derivative of the function.
- We do not want to limit ourselves to differentiable functions.
- So, it is reasonable to require that the differences $s(z + \Delta z) - s(z)$ and $s(z) - s(z - \Delta z)$ be statistically independent.
- Under this assumption, for any $a < b$ and for any large number N :
 - we can consider points $z_i = a + i \cdot (b - a)/N$ between a and b ,
 - so that $z_0 = a$ and $z_N = b$.
- So, we can conclude that the difference $s(b) - s(a)$ can be represented as a sum of a large number of small independent random variables:
$$s(b) - s(a) = (s(z_1) - s(z_0)) + (s(z_2) - s(z_1)) + \dots + (s(b) - s(z_{N-1})).$$

16. Which Probability Measure Should We Select (cont-d)

- In this case, according to the Central Limit Theorem, its distribution is Gaussian (normal).
- It is known that, in general:

- the distribution of a multi-D Gaussian random vector

$$\eta = (\eta_1, \dots, \eta_n)$$

- is uniquely determined by its means $m_i = E[\eta_i]$ and covariances

$$E[(\eta_i - m_i) \cdot (\eta_j - m_j)].$$

- In our case, we thus need to know the means values $E[s(a)]$ for all a and the corresponding covariances.
- Let us describe reasonable mean and covariance values.
- We are discussing the distribution that would be applicable to all possible practical situations.

17. Which Probability Measure Should We Select (cont-d)

- In many practical situations – e.g., when measuring time or temperature – the choice of the starting point is just a matter of convenience.
- We can select a different starting point which is z_0 values smaller.
- Then all the numerical values are increased by z_0 : $a \mapsto a + z_0$ and $b \mapsto b + z_0$.
- This change of the starting point does not change the physical situation.
- So, it makes sense to require that statistical characteristics also should not change under such a transformation.
- Thus, we should have $E[s(a)] = E[s(a + z_0)]$ for all a and z_0 .
- In particular, for any a and b , if we take $z_0 = b - a$, we conclude that $E[s(a)] = E[s(b)]$.
- Thus, for all a and b , we have $E[s(b) - s(a)] = E[s(b)] - E[s(a)] = 0$.

18. Which Probability Measure Should We Select (cont-d)

- The mean value of the difference $s(b) - s(a)$ is 0.
- So, the variance – that we will denote by $V(a, b)$ – of this difference is simply equal to the expected value of its square

$$V(a, b) = E [(s(b) - s(a))^2] .$$

- For the variance, the requirement that nothing should depend on the choice of the starting point means that:
 - for all a, b , and z_0 ,
 - we should have $V(a, b) = V(a + z_0, b + z_0)$.
- In particular, for $z_0 = -a$, we conclude that $V(a, b) = V(0, b - a)$, i.e., that $C(a, b) = v(b - a)$, where we denoted $v(z) \stackrel{\text{def}}{=} V(0, z)$.
- As we have mentioned earlier, for any $a > 0$ and $b > 0$, the differences $s(a) - s(0)$ and $s(a + b) - s(a)$ are independent.
- It is known that the variance of the sum of two independent random variables is equal to the sum of their variances.

19. Which Probability Measure Should We Select (cont-d)

- So, for the sum $s(a + b) - s(0)$ of the two differences, we get

$$V(a + b, 0) = V(a, 0) + V(a + b, a).$$

- In terms of the function $v(z)$, that $v(a + b) = v(a) + v(b)$.
- Variances are non-negative, and it is known that all non-negative solutions to this equation have the form $v(a) = C \cdot a$ for some $C \geq 0$.
- So, $V(a, b) = C \cdot (b - a)$.
- We get a Gaussian random process for which the differences $s(b) - s(a)$ for all $a < b$ have:
 - mean value 0 and
 - the variance proportional to $b - a$.
- This process is known as a *Wiener process* or a *random walk*.

20. So What Is the Optimal Interpolation?

- We have selected a reasonable probability distribution on the set of all continuous functions.
- Now, we can start answering the original question:
 - once we know the values $t_1 = s(z_1), \dots, t_k = s(z_K)$,
 - what is the optimal interpolation $s(z)$?
- The solution to this question is known.
- Indeed, the minimized mean square value is the sum of the values corresponding to different z .
- Thus, to minimize this mean square value, it is sufficient:
 - to minimize, for each z , the expected value $E[(s_0(z) - s(z))^2]$,
 - i.e., to find the value $s_0(z)$ for which this conditional mean square value is the smallest possible.
- Differentiating this expression with respect to the unknown $s_0(z)$ and equating the derivative to 0, we conclude that $2 \cdot (s_0(z) - E[s(z)]) = 0$.

21. So What Is the Optimal Interpolation (cont-d)

- So, $s_0(z)$ should be equal to the expected value of $s(z)$ under the condition that we have $s(z_k) = t_k$ for all k .
- Specifically, let us find out what is the optimal interpolated value $s(z)$ when z is between z_k and z_{k+1} .
- Knowing the values t_k is equivalent to knowing the differences $d_k = s(z_{k+1}) - s(z_k)$.
- Based on these differences, we want to reconstruct the difference $s(z) - s(z_k)$.
- Of course, the variables that are independent on the desired quantity cannot affect its value.
- The differences corresponding to non-intersecting intervals are independent.
- So, the desired difference is independent of all the differences $s(z_{i+1}) - s(z_i)$ except for one of them: $s(z_{k+1}) - s(z_k)$.

22. So What Is the Optimal Interpolation (cont-d)

- Thus, only this difference affects the optimal interpolation result.
- So, we want to find the conditional expected value $E[X|Y = c]$, where $X = s(z) - s(z_k)$, $Y = s(z_{k+1}) - s(z_k)$, and $c = d_k$.
- We can solve this problem by taking into account that:
 - the conditional value of X under the condition $Y = c$ is the same as the conditional value of the linear combination $X + \beta \cdot (Y - c)$,
 - for all possible values β .
- So, to use the independence argument, let us find the value β for which this combination is independent of Y .
- For Gaussian variables, once the covariance between them is 0, the variables are independent.
- So, let us find the value β for which the covariance $E[((X + \beta \cdot Y) \cdot Y)]$ is equal to 0.
- This covariance is equal to $E[X \cdot Y] + \beta \cdot E[Y^2]$.

23. So What Is the Optimal Interpolation (cont-d)

- Thus, this expression is equal to 0 when $\beta = -E[XY]/E[Y^2]$.
- For this β , the linear combination is independent of Y .
- Thus, its conditional expected value:
 - under the condition that Y has value c ,
 - is simply equal to the unconditional expected value of this linear combination.
- Since $E[X] = E[Y] = 0$, this unconditional expected value of the linear combination is equal to:

$$s_0(z) - s(z_k) = E[X|Y = c] = E[X + \beta \cdot (Y - c)] = -\beta \cdot c = \frac{E[XY]}{E[Y^2]} \cdot c.$$

- In our case, as we have mentioned earlier,

$$E[Y^2] = E[(s(z_{k+1}) - s(z_k))^2] = C \cdot (z_{k+1} - z_k), \text{ and}$$

$$E[X \cdot Y] = E[(s(z) - s(z_k)) \cdot (s(z_{k+1}) - s(z_k))].$$

24. So What Is the Optimal Interpolation (cont-d)

- Here, $s(z_{k+1}) - s(z_k) = (s(z_{k+1}) - s(z)) + (s(z) - s(z_k))$, so:

$$E[X \cdot Y] = E[(s(z) - s(z_k)) \cdot (s(z_{k+1}) - s(z))] + E[(s(z) - s(z_k))^2] .$$

- The first term in the right-hand side is 0, since the corresponding intervals do not intersect.
- The second term is equal to $C \cdot (z - z_k)$.
- Thus, the above leads to the following optimal interpolation value

$$s_0(z) - s(z_k) = \frac{C \cdot (z - z_k)}{C \cdot (z_{k+1} - z_k)} \cdot (s(z_{k+1}) - s(z_k)), \text{ i.e.};$$

$$s_0(z) = s(z_k) + \frac{z - z_k}{z_{k+1} - z_k} \cdot (s(z_{k+1}) - s(z_k)).$$

- This is a linear function of z .

25. So What Is the Optimal Interpolation (cont-d)

- So, indeed, the optimal interpolation leads to a function that is linear on each segment $[z_k, z_{k+1}]$ – i.e., to a piecewise linear function.
- This provides a possible explanation of why piecewise linear activation functions are so efficient.

26. Why Only Addition and Multiplication Have Worked Well? Towards a Possible Explanation

- Let us now provide some semi-heuristic ideas that may help to explain why
 - out of all possible combination operations,
 - only addition and multiplication work well.
- To come up with these ideas, let us consider reasonable requirements on combination operations.

27. First Requirement: We Want the Expression $a \oplus_f b$ to Be Differentiable

- One of the reasons why neural networks are so successful is that back-propagation is a very fast optimization procedure.
- We want to be able to use this procedure.
- For this purpose, we need to make sure that all our operations are differentiable.
- The operation $a \oplus b$ is based on some function $f(x)$.
- So, to make sure that this operation is differentiable, it is reasonable to require that function $f(x)$ should be differentiable.

28. Scale-Invariance: Another Natural Requirement

- In most computations, we process the values of the physical quantities.
- For example, when we predict weather:
 - we input temperature, wind speed, humidity at different locations,
 - and we predict the future values of these quantities at desired locations.
- In a computer, all these quantities are described by numbers.
- It is important to take into account that the numerical value of each quantity depends on the choice of a measuring unit.
- For example, for the same wind speed, we get different numerical results if we measure it in kilometers per hour or in miles per hour.

29. Scale-Invariance: Another Natural Requirement (cont-d)

- In general:
 - if we replace the original measuring unit by a new one which is λ times smaller,
 - then all numerical values get multiplied by λ : $x \mapsto \lambda \cdot x$.
- For example, if we replace meters with centimeters, then 1.7 m becomes $100 \cdot 1.7 = 170$ cm.
- It is reasonable to require that the computations remain the same whatever measuring units we use.
- It makes no sense to use one algorithm for meters and a completely different one for centimeters.

30. Scale-Invariance: Another Natural Requirement (cont-d)

- In particular, this means that:
 - if we re-scale a and b into $\lambda \cdot a$ and $\lambda \cdot b$,
 - then the value $(\lambda \cdot a) \oplus (\lambda \cdot b)$ should represent the same quantity as the original combination $a \oplus b$, but maybe in different units,
 - i.e., we should have $(\lambda \cdot a) \oplus (\lambda \cdot b) = \mu(\lambda) \cdot (a \oplus b)$ for some μ depending on λ .
- In other words, we want the operation $a \oplus b$ to be, in this sense, scale-invariant.
- For addition $a \oplus b = a + b$, we have $\lambda \cdot a + \lambda \cdot b = \lambda \cdot (a + b)$.
- So the scale-invariance property is true for $\mu(\lambda) = \lambda$.
- For multiplication $a \oplus b = a \cdot b$, we have $(\lambda \cdot a) \cdot (\lambda \cdot b) = \lambda^2 \cdot (a \cdot b)$.
- So the scale-invariance property holds for $\mu(\lambda) = \lambda^2$.

31. But Why Should We Require Scale-Invariance: Scale-Invariance Follow from Optimality

- But why should an optimal operation be scale-invariant?
- To make the above heuristic idea more convincing, let us recall what we mean by saying that some alternative is optimal.
- It usually means that we have a way to compare two different alternatives A and B and conclude that:
 - either A is better than or equivalent to B (with respect to the chosen optimality criterion); we will denote this case by $A \succeq B$;
 - or B is better than or equivalent to A : $B \succeq A$.
- In particular, it is possible that:
 - with respect to the chosen optimality criterion,
 - the alternatives A and B are equally good, i.e., that $A \succeq B$ and $B \succeq A$.
- We will denote this by $A \sim B$.

32. But Why Should We Require Scale-Invariance: Scale-Invariance Follow from Optimality (cont-d)

- For example, we can say that:
 - an algorithm A is better than or equivalent to B
 - if its average computation time is smaller or the same as the average computation time of the algorithm B .
- Of course, these preferences should be transitive:
 - if A is better than or of the same quality as B ,
 - and B is better than or of the same quality as C ,
 - then A should be better than or of the same quality as C .
- In mathematical terms, such transitive relations are known as *pre-orders*.
- Their difference from the more traditional order relations is that for pre-orders, we can have two different alternatives A and B for which

$$A \sim B.$$

33. But Why Should We Require Scale-Invariance: Scale-Invariance Follow from Optimality (cont-d)

- In these terms, the statement that the alternative A_{opt} is optimal means that for every other alternative A , we have $A_{\text{opt}} \succeq A$.
- It may happen that with respect to the chosen optimality criterion, several different alternatives are optimal.
- In this case, we can use the resulting non-uniqueness to optimize something else.
- For example:
 - among all algorithms with the fastest average computation time,
 - we can select the one with the fastest worst-case computation time.
- This means, in effect, that we are modifying our definition of which alternative is better.

34. But Why Should We Require Scale-Invariance: Scale-Invariance Follow from Optimality (cont-d)

- In the new definition, A is better than B if:
 - either A is better than B according to the original criterion.
 - or A and B are equivalent with respect to the original criterion but A is better than B in some new sense.
- It may be that with respect to this new optimality criterion, we still have several different optimal alternative.
- Then we can again use this non-uniqueness to optimize something else.
- If needed, we can similarly modify the optimality criterion again and again.
- This modification process stops when we reach:
 - the *final* optimality criterion,
 - i.e., a criterion for which there is only one optimal alternative.

35. But Why Should We Require Scale-Invariance: Scale-Invariance Follow from Optimality (cont-d)

- Let us describe all this in precise terms.
- Let S be a set. Its elements will be called *alternatives*.
- By an *optimality criterion*, we mean a transitive relation $\succeq$ on the set A for which for every A and B , either $A \succeq B$ or $B \succeq A$ or both.
- We say that an alternative A_{opt} is *optimal* with respect to $\succeq$ of $A_{\text{opt}} \succeq A$ for all $A \in S$.
- We say that the optimality criterion is *final* if there exists exactly one optimal alternative.
- Suppose now that we have some transformation T that does not change the physical situation – like scaling in our case.
- In this case, it make sense to assume that the relative quality of two alternatives should not change if we apply this transformation.

36. But Why Should We Require Scale-Invariance: Scale-Invariance Follow from Optimality (cont-d)

- For example:
 - if an algorithm A for processing lengths of different objects is better than a different algorithm B on examples when we measure all the lengths in meters,
 - then it is reasonable to expect that A will still be better than B if we process data measured in centimeters.
- It would be strange to find out that, e.g., one sorting algorithm works better for meters while a different one is better for centimeters.
- In such situations, we get the following result.
- Let $T : S \mapsto S$ be a 1-1-transformation.
- We say that the optimality criterion is T -invariant if for every A and B , $A \succeq B$ if and only if $T(A) \succeq T(B)$.

37. But Why Should We Require Scale-Invariance: Scale-Invariance Follow from Optimality (cont-d)

- **Proposition.** *For every final T -invariant optimality criterion, the optimal alternative A_{opt} is also T -invariant, i.e., $T(A_{\text{opt}}) = A_{\text{opt}}$.*
- **Proof.** Let A_{opt} be the optimal alternative.
- By definition, this means that $A_{\text{opt}} \succeq A$ for all $A \in S$.
- In particular, for each A , we have $A_{\text{opt}} \succeq T^{-1}(A)$, where $T^{-1}(A)$ denotes the inverse transformation.
- Since the optimality criterion is T -invariant, we conclude that for every A , we have $T(A_{\text{opt}}) \succeq T(T^{-1}(A)) = A$.
- So, $T(A_{\text{opt}}) \succeq A$ for $A \in S$.
- By definition of an optimal alternative, this means that the alternative $T(A_{\text{opt}})$ is also optimal.
- But since the optimality criterion is final, we can have only one optimal alternative.

38. But Why Should We Require Scale-Invariance: Scale-Invariance Follow from Optimality (cont-d)

- So, $T(A_{\text{opt}}) = A_{\text{opt}}$.
- The proposition is proven.
- Let us apply this result to scaling.
- In general, an algorithm $c = F(a, b) = a \oplus_f b$ may transform numerical values a and b into a numerical value of some quantity c .
- This resulting quantity c may be expressed in the same units – as in the case of addition.
- The quantity c may also be expressed in terms of other units.
- For example, for multiplication, the product is described in square meters or square centimeters.
- In general, when we re-scale a and b into $\lambda \cdot a$ and $\lambda \cdot b$, we will have $c \mapsto \mu(\lambda) \cdot c$ for some μ depending on λ .

39. But Why Should We Require Scale-Invariance: Scale-Invariance Follow from Optimality (cont-d)

- We do not want to impose any prior restrictions on the function $\mu(\lambda)$ – that describes how this new unit changes.
- So, instead of the functions $F(a, b)$ themselves, we will consider *families* of function $\{C \cdot F(a, b)\}_C$.
- Here, the function $F(a, b)$ is fixed, and C can take any value.
- For such families, it makes sense to require that:
 - once we change the measuring unit, the relative quality of the corresponding KAN-type network will not change,
 - i.e., that the optimality criterion on the set of all such families should be invariant with respect to scaling.
- By the above Proposition, it means that the optimal family should also be invariant with respect to scaling:

$$\{C \cdot F(\lambda \cdot a, \lambda \cdot b)\}_C = \{C \cdot F(a, b)\}_C.$$

40. But Why Should We Require Scale-Invariance: Scale-Invariance Follow from Optimality (cont-d)

- In particular, the function $F(\lambda \cdot a, \lambda \cdot b)$ from the left-hand set must belong to the right-hand set.
- So we must have $F(\lambda \cdot a, \lambda \cdot b) = C(\lambda) \cdot F(a, b)$ for some $C(\lambda)$.
- This is exactly the invariance condition, with $\mu(\lambda) = C(\lambda)$.
- So, we provided arguments why this condition should be satisfied for the optimal combination operation.

41. Definitions and the Main Result of This Section

- Now we are ready to formulate (and prove) the main result of this section.
- Let $f(x)$ be an invertible differentiable function.
- Then, we denote $a \oplus_f b \stackrel{\text{def}}{=} f^{-1}(f(a) + f(b))$.
- We say that the operation $a \oplus_f b$ is *scale-invariant* if:
 - for every $\lambda > 0$ there exists a $\mu(\lambda) > 0$ for which,
 - for all a and b , we have $(\lambda \cdot a) \oplus_f (\lambda \cdot b) = \mu(\lambda) \cdot (a \oplus_f b)$.
- **Proposition.** *The only scale-invariant differentiable operations $\oplus_f$ are:*
 - addition $a \oplus_f b = a + b$ and
 - multiplication $a \oplus_f b \stackrel{\text{def}}{=} C \cdot a \cdot b$ for some constant C .
- This result may help to explain why only addition and multiplication lead to effective results.

42. Proof: Part 1

- Let us first use scale-invariance to derive an equality in terms of the derivative of the desired function $f(a)$.
- Scale-invariance implies that if $a \oplus_f b = (a + \Delta a) \oplus_f (b + \Delta b)$, then

$$(\lambda \cdot a) \oplus_f (\lambda \cdot b) = ((\lambda \cdot (a + \Delta a)) \oplus_f (\lambda \cdot (b + \Delta b))).$$

- By definition of the operation $\oplus_f$, this means that if $f(a) + f(b) = f(a + \Delta a) + f(b + \Delta b)$, then

$$f(\lambda \cdot a) + f(\lambda \cdot b) = f(\lambda \cdot (a + \Delta a)) + f(\lambda \cdot (b + \Delta b)).$$

- By moving all the a -terms to one side and all the b -terms to the other side, we conclude that

$$-(f(a + \Delta a) - f(a)) = f(b + \Delta b) - f(b).$$

- For small Δa and Δb , we get $f(a + \Delta a) = f(a) + f'(a) \cdot \Delta a + o(\Delta a)$.
- So, the above takes the form

$$-f'(a) \cdot \Delta a = f'(b) \cdot \Delta b + o(\Delta a).$$

43. Proof: Part 1 (cont-d)

- To get the desired equality, we need to select Δb and Δa for which the following equality holds:

$$\frac{\Delta b}{\Delta a} = -\frac{f'(b)}{f'(a)} + o(1).$$

- For these same Δa and Δb , the scaled equality similarly leads to

$$-(f(\lambda \cdot a + \lambda \cdot \Delta a) - f(\lambda \cdot a)) = f(\lambda \cdot b + \lambda \cdot \Delta b) - f(\lambda \cdot b).$$

- Thus, it leads to:

$$\frac{\lambda \Delta b}{\lambda \Delta a} = \frac{\Delta b}{\Delta a} = -\frac{f'(\lambda \cdot b)}{f'(\lambda \cdot a)} + o(1).$$

- When $\Delta a \rightarrow 0$, then we can conclude that for all a , b , and λ , we have:

$$\frac{f'(b)}{f'(a)} = \frac{f'(\lambda \cdot b)}{f'(\lambda \cdot a)}.$$

44. Proof: Part 2

- Let us now derive a functional equation for $f'(a)$.
- For this purpose, let us separate the variables in the formula

$$\frac{f'(b)}{f'(a)} = \frac{f'(\lambda \cdot b)}{f'(\lambda \cdot a)}.$$

- We can do it if we multiply both sides by $f'(\lambda \cdot a)$ and divide both sides by $f'(b)$.
- Then, we get the following equality:

$$\frac{f'(\lambda \cdot a)}{f'(a)} = \frac{f'(\lambda \cdot b)}{f'(b)}.$$

- The fact that this equality holds for all a and b means that:
 - the ratio on each side of the equality does not depend on a ,
 - it depends only on λ .

45. Proof: Part 2 (cont-d)

- Let us denote this ratio by $r(\lambda) \stackrel{\text{def}}{=} \frac{f'(\lambda \cdot a)}{f'(a)}$.
- Multiplying both sides by $f'(a)$, we conclude that for all a and $\lambda > 0$, we have: $f'(\lambda \cdot a) = r(\lambda) \cdot f'(a)$.

46. Proof: Part 3

- Let us now find possible functions $f(a)$ by solving the resulting functional equation for $f'(a)$: $f'(\lambda \cdot a) = r(\lambda) \cdot f'(a)$.
- The function $f(a)$ is differentiable, its derivative $f'(a)$ is measurable.
- Thus the ratio $r(\lambda)$ is also measurable.
- It is known that every measurable solution to the above functional equation has the form $f'(a) = c_1 \cdot x^d$ for some c_1 and d .
- Here, $c_1 \neq 0$ – since otherwise we would have $f'(a) = 0$, the function $f(a)$ will be constant, but we assumed that this function is invertible.

47. Proof: Part 3 (cont-d)

- Integrating both sides of the formula that describes the derivative $f'(a)$, we conclude that:
 - for $d \neq -1$, we have $f(a) = c_2 \cdot a^\alpha + c_0$, where $c_2 = c_1/(d+1)$, $\alpha = d+1$, and c_0 is an integration constant;
 - for $d = -1$, we have $f(a) = c_1 \cdot \ln(a) + c_0$, where c_0 is an integration constant.
- After some preliminary analysis, let us consider these two cases one by one.

48. Proof: Part 4

- Let us first perform some preliminary analysis of the scale-invariance condition.
- By definition, the condition $a \oplus_f b = c$ means that

$$f^{-1}(f(a) + f(b)) = c.$$

- If we apply the function $f(z)$ to both sides of this equality, we conclude that $f(a) + f(b) = f(c)$.
- Similarly, the scaled condition means that $f(\lambda \cdot a) + f(\lambda \cdot b) = f(\lambda \cdot c)$.
- Thus, scale-invariance means that for all a , b , and $\lambda > 0$, this formula follows from $f(a) + f(b) = f(c)$.

49. Proof: Part 5

- Let us now consider the case when $d \neq -1$ and $f(a)$ is described by the power law formula.
- For this formula, the original $f(a) + f(b) = f(c)$ means that:

$$c_2 \cdot a^\alpha + c_0 + c_2 \cdot b^\alpha + c_0 = c_2 \cdot c^\alpha + c_0.$$

- This is equivalent to:

$$c_2 \cdot a^\alpha + c_2 \cdot b^\alpha - c_2 \cdot c^\alpha = -c_0.$$

- The re-scaled equality takes the form:

$$c_2 \cdot (\lambda \cdot a)^\alpha + c_0 + c_2 \cdot (\lambda \cdot b)^\alpha + c_0 = c_2 \cdot (\lambda \cdot c)^\alpha + c_0.$$

- This is equivalent to:

$$c_2 \cdot \lambda^\alpha \cdot a^\alpha + c_2 \cdot \lambda^\alpha \cdot b^\alpha - c_2 \cdot \lambda^\alpha \cdot c^\alpha = -c_0.$$

- Multiplying both sides of the original equality by λ^α , we conclude that

$$c_2 \cdot \lambda^\alpha \cdot a^\alpha + c_2 \cdot \lambda^\alpha \cdot b^\alpha - c_2 \cdot \lambda^\alpha \cdot c^\alpha = -\lambda^\alpha \cdot c_0.$$

50. Proof: Part 5 (cont-d)

- The left-hand sides of this formula and of the scaled formula are the same.
- So the right-hand sides must be equal, and we must have

$$-c_0 = -\lambda^\alpha \cdot c_0.$$

- For $\lambda \neq 1$, this implies that $c_0 = 0$.
- So, the above formula takes the form

$$c_2 \cdot a^\alpha + c_2 \cdot b^\alpha = c_2 \cdot c^\alpha.$$

- If we divide both sides by $c_2 \neq 0$, then we get $c^\alpha = a^\alpha + b^\alpha$, i.e.,

$$a \oplus_f b = (a^\alpha + b^\alpha)^{1/\alpha}.$$

- Let us now take into account that the function $a \oplus_f b$ must be differentiable.

51. Proof: Part 5 (cont-d)

- If we use the chain rule to differentiate the expression for $a \oplus_f b$ with respect to a , we conclude that

$$\frac{\partial}{\partial a} \left[(a^\alpha + b^\alpha)^{1/\alpha} \right] = \frac{1}{\alpha} \cdot (a^\alpha + b^\alpha)^{1/\alpha-1} \cdot \alpha \cdot a^{\alpha-1}.$$

- Let us simplify this formula.
- We can represent $a^\alpha + b^\alpha$ as

$$a^\alpha + b^\alpha = a^\alpha \cdot \left(1 + \left(\frac{b}{a} \right)^\alpha \right), \text{ thus:}$$
$$(a^\alpha + b^\alpha)^{1/\alpha-1} = (a^\alpha)^{1/\alpha-1} \cdot \left(1 + \left(\frac{b}{a} \right)^\alpha \right)^{1/\alpha-1}. \quad (40)$$

- Here,

$$(a^\alpha)^{1/\alpha-1} = a^{\alpha \cdot (1/\alpha-1)} = a^{1-\alpha}, \text{ so:}$$
$$(a^\alpha + b^\alpha)^{1/\alpha-1} = a^{1-\alpha} \cdot \left(1 + \left(\frac{b}{a} \right)^\alpha \right)^{1/\alpha-1}.$$

52. Proof: Part 5 (cont-d)

- Let us substitute this expression into the formula for the derivative and taking into account that the factors $1/\alpha$ and α cancel each other.
- Then, we conclude that:

$$\frac{\partial}{\partial a} \left[(a^\alpha + b^\alpha)^{1/\alpha} \right] = a^{1-\alpha} \cdot \left(1 + \left(\frac{b}{a} \right)^\alpha \right)^{1/\alpha-1} \cdot a^{\alpha-1}.$$

- Here, $a^{1-\alpha} \cdot a^{\alpha-1} = 1$, so

$$\frac{\partial}{\partial a} \left[(a^\alpha + b^\alpha)^{1/\alpha} \right] = \left(1 + \left(\frac{b}{a} \right)^\alpha \right)^{1/\alpha-1}.$$

- When a and b are close to 0, we should get close to the derivative of the function $a \oplus_f b$ for $a = b = 0$.
- However, when $\alpha \neq 1$, we have $1/\alpha - 1 \neq 0$ and thus, we will have different expressions for different ratios b/a .

53. Proof: Part 5 (cont-d)

- So, the only case when the operation $a \oplus_f b$ is differentiable everywhere – in particular, for $a = b = 0$ – is when $\alpha = 1$.
- In this case, the formula for $a \oplus_f b$ turns into $c = a + b$.

54. Proof: Part 6

- Finally, let us consider the case when $d = -1$, and the function $f(a)$ is described by the log formula.
- In this case, $c = a \oplus_f b$ means that:

$$c_1 \cdot \ln(a) + c_0 + c_1 \cdot \ln(b) = c_1 \cdot \ln(c) + c_0.$$

- If we divide both sides by $c_1 \neq 0$ and move the term c_0/c_1 into the left-hand side, we conclude that

$$c = \ln(a) + \ln(b) + c_0/c_2.$$

- Applying $\exp(z)$ to both sides, we conclude that $c = C \cdot a \cdot b$, where we denoted $C \stackrel{\text{def}}{=} \exp(c_0/c_1)$.
- The proposition is proven.

55. Acknowledgments

This work was supported by:

- the AT&T Fellowship in Information Technology,
- the Institute for Risk and Reliability, Leibniz Universitaet Hannover, Germany,
- the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Focus Program SPP 100+ 2388, Grant Nr. 501624329,
- the European Union under the project ROBOPROX (No. CZ.02.01.01/00/22 008/0004590),
- the Center of Excellence in Econometrics, Faculty of Economics, Chiang Mai University, Thailand,
- the Ho Chi Minh City University of Banking, Vietnam, and
- Thang Long University, Hanoi, Vietnam.