# Why Exp-Minus-Log (EML): an explanation based on optimization and symmetry

Vladik Kreinovich[1], Leobardo Valera[2], Igor Atamanyuk[3,4], and Yuriy Kondratenko[5]
Department of Computer Science
University of Texas at El Paso
500 W. University
El Paso, Texas 79968, USA
vladik@utep.edu
[2]Department of Science
Doña Ana Community College,
New Mexico State University
Las Cruces, New Mexico 88003, USA
leobardovalera@gmail.com
[3]Department of Applied Mathematics
Warsaw University of Life Sciences
Warsaw, Poland, ihor_atamaniuk@sggw.edu.pl
[4]Mykolaiv National Agrarian University, Mykolaiv, Ukraine
[5]Intelligent Information Systems Department
Petro Mohyla Black Sea National University
10, 68th Desantnykiv Str.
Mykolaiv 54003 Ukraine, y_kondrat2002@yahoo.com

**Abstract**

It has been recently shown that, similarly to the fact that every Boolean operation can be performed if we only have gates of a single type – NAND or NOR – many elementary functions can be obtained by using a constant 1 and a single function $\text{eml}(x, y) = \exp(x) - \ln(y)$. This function is known as exp-minus-log, eml, for short. There are several other functions with the same property, but among them, eml is the simplest. In this paper, we show that, in some reasonable sense, eml is optimal. To prove this, we use symmetry-based approach and its relation to optimization.

# 1 Formulation of the problem

**In propositional logic, it is, in principle, sufficient to have gates of only one type.** A recent result [7] – that we describe below – was motivated by the following fact. There are many different operations in Boolean logic. Some of them can be described in terms of others; e.g.:

- "and" can be described in terms of "or" and "not" as $x \,\&\, y = \neg(\neg x \vee \neg y)$;
- "or" can be described in terms of "and" and "not" as $x \vee y = \neg(\neg x \,\&\, \neg y)$; etc.

It is known that we can compute any Boolean expression if we have only one operation – this can be either $\text{NAND}(x, y) \stackrel{\text{def}}{=} \neg(x \,\&\, y)$ or $\text{NOR}(x, y) \stackrel{\text{def}}{=} \neg(x \vee y)$.

This result is used in the design of computer chips, where NAND gates are actively used.

**A related recent result.** It has been recently shown [7] that all arithmetic operations (addition, subtraction, multiplication, division), several basic elementary functions ($\exp(x)$, $\ln(x)$, $x^y$, trigonometric functions, etc.), and several constants such as $\mathrm{i} = \sqrt{-1}$ and $e$ can be obtained if we only have a single function $\text{eml}(x, y) = \exp(x) - \ln(y)$ (and a constant 1). This function is known as exp-minus-log, or eml, for short.

For example, $e^x = eml(x, 1)$, $\ln(x) = eml(1, eml(eml(1, x), 1))$, etc. Similar to how the result about NAND and NOR gates is used in computing Boolean expressions, this new result can be potentially useful in computing functions of several variables.

The paper [7] lists several other functions with the same universality property. Out of all these functions, eml seems to be the simplest.

**Natural question.** A natural question is: how can we explain the effectiveness of eml?

**What we do in this paper.** In this paper, we show that the eml function is indeed optimal in some reasonable sense; namely, we show that it is optimal based on some natural symmetry (= invariance) properties.

**The structure of this paper.** We start the paper by recalling, in Section 2, why we need data processing in the first place. Based on this analysis, in Section 3, we explain why symmetries naturally appear in such situations. In Section 4, we consider the two most natural transformations with respect to which data processing techniques should be invariant. We also emphasize that, depending on the practical situations, we need to distinguish between two types of invariances: invariance proper and covariance; this is done in Section 5. In Section 6, we explain the general relation between invariance and optimality. This relation leads to a description of optimal data processing functions of one variable in Section 7. In Sections 8, we use invariances to find the optimal binary operations for combining functions of one variable – which leads us, in Section 9, to a class containing the eml function.

## 2 Why we need data processing in the first place

The main purpose of elementary functions and their combinations is to process data, i.e., to process values of different physical quantities. So, to describe which functions are more adequate, it is important to recall why we need data processing in the first place.

What do we humans want? We want to know the current state of the world, we want to know the future state of the world, and we want to know what we can do to make the future state as beneficial for us as possible. The state of the world is characterized by the values of physical quantities. Some of these values we can measure directly: e.g., we can simply measure the distance between the two city locations. However, for many quantities $y$, it is difficult to measure them directly. For example, it is difficult to directly measure the distance from the Earth to a farway star. Since we cannot *directly* measure the value $y$, we need to measure it *indirectly*:

- we find some easier-to-measure quantities $x_1, \ldots, x_n$ that are related to $y$ by a known dependence $y = f(x_1, \ldots, x_n)$;
- we measure these quantities and use the measurement results $\widetilde{x}_1, \ldots, \widetilde{x}_n$ to provide the following estimate for $y$:

$$\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n). \tag{1}$$

Computing this estimate is one of the cases of data processing.

Future values we cannot measure directly at all – we want to know these values now, before they can be measured. For these values, the only way to estimate a future value $y$ is:

- to find some current (and past) values $x_1, \ldots, x_n$ that are related to $y$ by a known dependence $y = f(x_1, \ldots, x_n)$,
- to measure these quantities, and then
- use the measurement results $\widetilde{x}_i$ to estimate the desired future value $y$.

Similarly, to determine the parameters $y_k$ of the best control or of the design strategy, we need to know how each of these parameters depends on the values $x_i$ that describe the current state of the world – and then use data processing $y_k = f_k(x_1, \ldots, x_n)$ to estimate $y$.

## 3 Why symmetries naturally appear

How do we get all the knowledge? For example, how did our ancestors know that the Sun will rise in the evening? Because they observed it again and again. The seasons changed, the behavior changed, the location may have changed, but the Sun continued to rise. So, they concluded that whether the Sun will rise does not depend on the location on the Earth, does not depend on the season, does not depend on their behavior. In mathematical terms, they concluded that the Sun-rising event is invariant with respect to changing the location, with respect to moving ahead in time, etc. This enabled them to predict that the Sun will continue rising in the future, no matter where they are.

How do we know that if we drop a pen, it will start falling down with the acceleration of 9.81 m/sec$^2$? Because people tried this in different locations on Earth, at different orientations, at different moments of time – and always got the same result. So, they concluded that this phenomenon is invariant with respect to shifts in space and in time, and this enabled them to predict that this will always happen anywehere on Earth.

How do we know that Ohm's law $V = I \cdot R$ will be valid on Mars? Because people had many experiments at different locations, at different moments of time – and it always worked. So we can conclude that this law is valid everywhere.

This is the background of all our knowledge: we have some phenomenon that repeats in somewhat different situations, so we conclude that this phenomenon is invariant with respect to the corresponding changes – and thus, predict that it will remain the same under new similar changes – e.g., after time shift, i.e., in the future moments of time. So, invariances – that physicists call *symmetries* – indeed naturally appear as a foundation of all our knowledge.

# 4 What are natural transformations

A large part of data processed by computers are values of different physical quantities. From this viewpoint, it is important to remember that the numerical value of a quantity depends on the choice of a measuring unit – and, for some quantities, on the selecting of the starting value. For different measuring units and for different starting values, the same quantity is described by different numbers.

For example, the 2 meters length has a numerical value $2 \cdot 100 = 200$ when described in centimeters. The temperature of 35°C has a numerical value of $35 + 273.15 = 308.15$ when described in Kelvins – a scale that has a different starting point.

In general, if we replace the original measuring unit by a different unit which is $\lambda$ times smaller, then each numerical value $x$ is multiplied by $\lambda$: $x \to x' = \lambda \cdot x$. In particular, for a transition from meters to centimeters, $\lambda = 100$.

Similarly, if we replace the original starting point with a new starting point which is $c$ before the original one, then $c$ is added to all the numerical values: $x \to x' = x + c$. In particular, for a transition from Celsius to Kelvin temperature scales, $c = 273.15$.

# 5 Invariance vs. covariance

In the previous text, when we talked about the fact that some physical processes do not change under appropriate transformations, we call such situations *invariances*. However, strictly speaking, we need to distinguish between two types of such situations:

- There are invariances in the narrow sense of this word, meaning that the values of some physical quantities do not change when we apply the corresponding transformation. For example, the acceleration of the free falling object does not change if we move this object to a nearby location on Earth.
- There are also what is called *covariances*, when the physical process remains the same, but the values of the corresponding quantities need to be adjusted. For example, when we test a scaled model of an airplane in a wind tunnel – to analyze how a proposed airplane design will work – the airflows of the model and of the airplane are (largely) similar, but the aerodynamic forces acting on the model need to be appropriately re-scaled to get the forces acting on the actual airplane.

As of now, this difference may sound purely terminological, but it is actually useful: we will use it in our explanation of the eml.

# 6 How are invariances related to optimality: a simple general result

**What does "optimal" mean: discussion.** In many cases, "optimal" means that we have some objective function $f(x)$ that we want to maximize (or minimize), and an alternative $a_{\rm opt}$ is optimal if $f(a_{\rm opt}) \ge f(a)$ for all other alternatives $a$ (or, in case of minimization, when $f(a_{\rm opt}) \le f(a)$ for all other alternatives $a$).

However, this is not the only way to describe optimality. Sometimes, for a given objective functions $f(x)$, there are several different alternatives that are all optimal. For example, if we are looking for the best sorting algorithm, and our objective function is the average computation time, then several sorting algorithms are optimal: quicksort, mergesort, heapsort, etc.; see, e.g., [1]. If we are looking for the best faculty candidate for a position, we may have several candidates whose research records are equally good.

In such situations, it is reasonable to use this non-uniqueness to optimize something else. For example, we may want to select, among all pre-selected sorting algorithms, a one with the smallest worst-case computation time. Among all pre-selected candidates, we may want to select a one who is the best teacher, etc. In this case, the optimality criterion is no longer described by a single objective function. In this case, an alternative $a$ is better than the alternative $b$ is either $f(a) > f(b)$ or $(f(a) = f(b)$ and $g(a) > g(b))$ for some other function $g(x)$.

If after this clarification, we still have several equally good alternatives, this means that our optimality criterion is not final: we can use the remaining non-uniqueness to optimize something else. When only one optimal alternative is left, this means that we have reached the final optimality criterion.

On all the stages, what is important is to have some way to compare two alternatives $a$ and $b$ and decide whether $a$ is better, or $b$ is better, or that these alternatives are of equal quality. Let us denote "$a$ is better than $b$" by $a \succ b$, and "$a$ and $b$ are of equal quality" by $a \sim b$. These two relations have to be consistent: e.g., if $a$ is better than $b$ and $b$ is better than $c$, then $a$ should be better than $c$.

**What does "optimal" mean: definitions.** Thus, we arrive at the following definitions.

**Definition 1.** *Let $A$ be a set. Its elements will be called* alternatives*. By an* optimality criterion*, we mean a pair of binary relations $(\succ, \sim)$ that satisfy the following properties for all $a$, $b$, and $c$:*

- *if $a \succ b$ and $b \succ c$, then $a \succ c$;*
- *if $a \succ b$ and $b \sim c$, then $a \succ c$;*
- *if $a \sim b$ and $b \succ c$, then $a \succ c$;*
- *if $a \sim b$ and $b \sim c$, then $a \sim c$;*
- *if $a \sim b$, then $b \sim a$;*
- *if $a \succ b$, then we cannot have $a \sim b$;*
- *$a \sim a$.*

**Definition 2.** *We say that an alternative $a_{\text{opt}}$ is* optimal *if for every $a \in A$, we have either $a_{\text{opt}} \succ a$ or $a_{\text{opt}} \sim a$.*

**Definition 3.** *We say that the optimality criterion is* final *if there is exactly one optimal alternative.*

**Invariances.** In many practical situations, there are some transformations that do not change which alternative is better. For example, when we are looking for investment alternatives and we want to select the most profitable one, it does not matter whether we measure the profit in US dollars, in Euros, on in Yen. Changing from one monetary unit to another one means multiplying all numerical values by an appropriate constant $\lambda$: $x \mapsto \lambda \cdot x$. In this case, these transformations are invariances. As we have mentioned, in physics, such invariances are called *symmetries*; see, e.g., [2, 8].

If the comparison does not change when we apply a transformation $g_1$, and does not change when we apply another transformation $g_2$, then, of course, it will not change if we first apply $g_1$ and then apply $g_2$, i.e., equivalently, if we apply the composition $g_2(g_1)$. Similarly, the comparison will not change if we apply the inverse transformation $g^{-1}$. Thus, the class of such transformations must be closed under composition and under taking an inverse. In mathematics, such classes are known as *transformation groups*. So, we arrive at the following definition.

**Definition 4.** *Let $G$ be a transformation group on the set $A$. We say that an optimality criterion $(\succ, \sim)$ is* $G$-invariant *for all $a, b \in A$ and for all $g \in G$, the following two properties are satisfied:*

- *if $a \succ b$ then $g(a) \succ g(b)$, and*
- *if $a \sim b$ then $g(a) \sim g(b)$.*

**How is invariance related to optimality.** This relation is described by the following simple result:

**Proposition 1.** [6] *For every $G$-invariant final optimality criterion, the optimal alternative $a_{\rm opt}$ is also $G$-invariant, i.e., $g(a_{\rm opt}) = a_{\rm opt}$ for all $g \in G$.*

**Proof.** By definition of optimality, for every $a \in A$, we have $a_{\rm opt} \succ a$ or $s_{\rm opt} \sim a$. In particular, for every $g \in G$, we have $a_{\rm opt} \succ g^{-1}(a)$ or $s_{\rm opt} \sim g^{-1}(a)$. Since the optimality criterion is $G$-invariant, this implies that $g(a_{\rm opt}) \succ a$ or $g(a_{\rm opt}) \sim a$. By definition of an optimal alternative, this means that $g(a_{\rm opt})$ is optimal. But since the optimality criterion is final, there is only one optimal alternative. So indeed, $g(a_{\rm opt}) = a_{\rm opt}$. The proposition is proven.

**Conclusion.** For such criteria, to find an optimal alternative, it is sufficient to look for invariant alternatives.

# 7 Optimal functions of one variable

**What we do in this section.** In this section, we use the above-described invariance ideas to come up with the optimal functions $y = f(x)$ of one variable.

**Analysis of the problem.** In one of the previous sections, we mentioned two natural transformations; scaling (that corresponds to changing the measuring unit) and shift (that corresponds to changing the starting point). If we consider the possibility to change the measuring unit, then the same dependence $y = f(x)$ expressed in some $y$-units will take the form $C \cdot f(x)$ if we use a different measuring unit – which is $C$ times smaller. Since these two functions describe the exact same physical dependence, it makes sense to consider not individual functions $f(x)$, but whole families $\{C \cdot f(x)\}_{C>0}$ in which the function $f(x)$ is fixed but the coefficient $C$ can take any positive value.

Similarly, if we consider the possibility to change the starting point, then the same dependence $y = f(x)$ expressed in some starting point for $y$-measurements will take the form $f(x)+C$ if we use a different starting point – which is $C$ units earlier. Since these two functions describe the exact same physical dependence, it makes sense to consider not individual functions $f(x)$, but whole families $\{f(x) + C\}_C$ in which the function $f(x)$ is fixed but the coefficient $C$ can take any real value.

For simplicity, let us assume that in both cases, the dependence $f(x)$ is smooth, i.e., differentiable. We can similarly consider scaling and shift of the quantity $x$. Thus, we arrive at the following formulations – that we will list together with results.

**Proposition 2.** *In the formulas below, we assume that the functions $f(x)$ are differentiable.*

- *For every scale-invariant final optimality criterion on the set of all classes $\{C \cdot f(x)\}_{C>0}$, the optimal class corresponds to $f(x) = x^\alpha$ for some $\alpha$.*
- *For every shift-invariant final optimality criterion on the set of all classes $\{C \cdot f(x)\}_{C>0}$, the optimal class corresponds to $f(x) = \exp(a \cdot x)$ for some $a$.*
- *For every scale-invariant final optimality criterion on the set of all classes $\{f(x) + C\}_C$, the optimal class corresponds to $f(x) = a \cdot \ln(x)$ for some $a$.*
- *For every shift-invariant final optimality criterion on the set of all classes $\{f(x) + C\}_C$, the optimal class corresponds to $f(x) = a \cdot x$ for some $a$.*

*Comment.* In the last three cases, the simplest classes correspond to $a = 1$, i.e., to the functions $\exp(x)$, $\ln(x)$, and $x$.

**Proof of Proposition 2.** For all four results, Proposition 1 implies that the optimal family is itself invariant.

$1^\circ$. For the first case, this means that $\{C \cdot f(\lambda \cdot x)\}_{C>0} = \{C \cdot f(x)\}_{C>0}$. In particular, this implies that the function $f(\lambda \cdot x)$ from the first family belongs to the second family, i.e., that

$$f(\lambda \cdot x) = C(\lambda) \cdot f(x) \qquad (2)$$

for some $C(\lambda)$. Both the expressions $f(\lambda \cdot x)$ and $f(x)$ are differentiable, so their ratio $C(\lambda)$ is also differentiable. So, we can differentiate both side of the formula (2) with respect to $\lambda$ and get $x \cdot f'(\lambda \cdot x) = C'(\lambda) \cdot f(x)$. In particular, for $\lambda = 1$, we get $x \cdot f'(x) = \alpha \cdot f(x)$, where we denoted $\alpha \stackrel{\text{def}}{=} C'(1)$, i.e.:

$$x \cdot \frac{df}{dx} = \alpha \cdot f.$$

We can separate the variables if we divide both sides by $x$, by $f$, and multiply both sides by $dx$. Then, we get

$$\frac{df}{f} = \alpha \cdot \frac{dx}{x}.$$

Integrating both sides, we get $\ln(f) = \alpha \cdot \ln(x) + C_1$, where $C_1$ is the integration constant. Applying $\exp(x)$ to both sides, we get:

$$f(x) = e^{\alpha \cdot \ln(x) + C_1} = e^{\alpha \cdot \ln(x)} \cdot e^{C_1} = \left(e^{\ln(x)}\right)^\alpha \cdot e^{C_1} = C \cdot x^\alpha,$$

where we denoted $C \stackrel{\text{def}}{=} e^{C_1}$. This part of the proposition is thus proven.

2°. For the second case, this means that $\{C \cdot f(x+c)\}_{C>0} = \{C \cdot f(x)\}_{C>0}$. In particular, this implies that the function $f(x+c)$ from the first family belongs to the second family, i.e., that

$$f(x+c) = C(c) \cdot f(x) \tag{3}$$

for some $C(c)$. Both the expressions $f(x+c)$ and $f(x)$ are differentiable, so their ratio $C(c)$ is also differentiable. So, we can differentiate both side of the formula (3) with respect to $c$ and get $f'(x+c) = C'(c) \cdot f(x)$. In particular, for $c = 0$, we get $f'(x) = a \cdot f(x)$, where we denoted $a \stackrel{\text{def}}{=} C'(0)$, i.e.,

$$\frac{df}{dx} = a \cdot f.$$

We can separate the variables if we divide both sides by $f$ and multiply both sides by $dx$. Then, we get

$$\frac{df}{f} = a \cdot dx.$$

Integrating both sides, we get $\ln(f) = a \cdot x + C_1$, where $C_1$ is the integration constant. Applying $\exp(x)$ to both sides, we get

$$f(x) = \exp(a \cdot x + C_1) = C \cdot \exp(a \cdot x),$$

where we denoted $C \stackrel{\text{def}}{=} \exp(C_1)$. The second part of the proposition is also proven.

3°. For the third case, this means that $\{f(\lambda \cdot x) + C\}_C = \{f(x) + C\}_C$. In particular, this implies that the function $f(\lambda \cdot x)$ from the first family belongs to the second family, i.e., that

$$f(\lambda \cdot x) = f(x) + C(\lambda) \tag{4}$$

for some $C(\lambda)$. Both the expressions $f(\lambda \cdot x)$ and $f(x)$ are differentiable, so their difference $C(\lambda)$ is also differentiable. So, we can differentiate both side of the formula (4) with respect to $\lambda$ and get $x \cdot f'(\lambda \cdot x) = C'(\lambda)$. In particular, for $\lambda = 1$, we get $x \cdot f'(x) = c_0$, where we denoted $c_0 \stackrel{\text{def}}{=} C'(1)$, i.e.,

$$x \cdot \frac{df}{dx} = c_0.$$

We can separate the variables if we divide both sides by $x$ and multiply both sides by $dx$. Then, we get

$$df = c_0 \cdot \frac{dx}{x}.$$

Integrating both sides, we get $f(x) = a \cdot \ln(x) + C_1$, where we denoted $a \stackrel{\text{def}}{=} c_0$ and $C_1$ is the integration constant. So, this part of the proposition is also proven.

4°. For the fourth case, this means that $\{f(x+c)+C\}_C = \{f(x)+C\}_C$. In particular, this implies that the function $f(x+c)$ from the first family belongs to the second family, i.e., that

$$f(x+c) = f(x) + C(c) \tag{5}$$

for some $C(c)$. Both the expressions $f(x+c)$ and $f(x)$ are differentiable, so their difference $C(c)$ is also differentiable. So, we can differentiate both side of the formula (5) with respect to $c$ and get $f'(x+c) = C'(c)$. In particular, for $c = 0$, we get $f'(x) = a$, where we denoted $a \stackrel{\text{def}}{=} C'(0)$.

Integrating both sides, we get $f(x) = a \cdot x + C_1$, where $C_1$ is the integration constant. The last part of the proposition is thus proven, and so is the proposition.

# 8 Optimal binary operations

**What we do in this section: a reminder.** We would like to have a function of two variables which is, in some reasonable sense, the best candidate for the desired universal property – that many basic functions can be represented by superposition of such functions. This property does not seem to depend on the choice of a measuring unit or on a choice of the starting point.

To construct such a function, it is reasonable to apply an optimal binary operation to optimal functions of one variable – the ones that we described in the previous section. As planned, to find the optimal binary operations, we will use the above result relating optimality and invariance.

**Let us find invariant binary operations.** To describe what we mean by invariant binary operations, let us recall the difference between invariance proper and covariance.

**Definition 5.** *By a* binary operation *(or simply* operation*, for short), we mean a differentiable function of two variables $f(y_1, y_2)$.*

- *We say that an operation $f(y_1, y_2)$ is* shift-invariant *if for all $C$ and for all $y_1$ and $y_2$, we have $f(y_1 + C, y_2 + C) = f(y_1, y_2)$.*
- *We say that an operation $f(y_1, y_2)$ is* shift-covariant *if for all $C$ and for all $y_1$ and $y_2$, we have $f(y_1 + C, y_2 + C) = f(y_1, y_2) + C$.*
- *We say that an operation $f(y_1, y_2)$ is* scale-invariant *if for all $C > 0$ and for all $y_1$ and $y_2$, we have $f(C \cdot y_1, C \cdot y_2) = f(y_1, y_2)$.*
- *We say that an operation $f(y_1, y_2)$ is* scale-covariant *if for all $C > 0$ and for all $y_1$ and $y_2$, we have $f(C \cdot y_1, C \cdot y_2) = C \cdot f(y_1, y_2)$.*

**Discussion.** We want our binary operation $f(y_1, y_2)$ to be both shift- and scale-invariant. According to Definition 5, each of these two invariances can be interpreted in two different ways. So, we get $2 \cdot 2 = 4$ possible interpretations of the need to have both invariances. The following proposition describes what happens if we combine for all interpretations.

**Proposition 3.**

- *The only shift-invariant and scale-invariant binary operations are constants* $f(y_1, y_2) = \text{const}$.
- *The only shift-invariant and scale-covariant binary operations are*

$$f(y_1, y_2) = C \cdot (y_1 - y_2)$$

*for some constant* $C$.

- *No binary operation is shift-covariant and scale-invariant.*
- *The only shift-covariant and scale-covariant binary operations are*

$$f(y_1, y_2) = C \cdot y_1 + (1 - C) \cdot y_2,$$

*for some constant* $C$.

*Comment.* It is worth mentioning that for the values $C$ between 0 and 1, the last binary operation – known as *Hurwicz combination* after the Nobel prize winner who invented it – is exactly what decision theory recommends to use as a utility of an alternative if all we know is that the actual utility is locates somewhere between $y_1$ and $y_2$; see, e.g., [3, 4, 5].

**Proof of Proposition 3.**

$1^\circ$. Let us first consider the case of shift-invaraince and scale-invarience. It is easy to check that constants are both shift-invariant and scale-invariant. So, to prove this part of the proposition, we need to prove that every shift-invariant and scale-invariance operation is a constant.

Indeed, shift-invariance means that we always have $f(y_1 + C, y_2 + C) = f(y_1, y_2)$. In particular, for $C = -y_2$, we conclude that $f(y_1, y_2) = f(y_1 - y_2, 0)$, i.e., that $f(y_1, y_2) = F(y_1 - y_2)$, where we denoted $F(y) \stackrel{\text{def}}{=} f(y, 0)$. Now, for $y_2 = 0$, scale-invariance means that $f(C \cdot y_1, 0) = f(y_1, 0)$, i.e., that $F(C \cdot y_1) = F(y_1)$. Now:

- for $y_1 > 0$, we can take $C = 1/y_1$ and thus conclude that $F(y_1) = F(1)$;
- for $y_1 < 0$, we can take $C = -1/y_1$ and thus conclude that $F(y_1) = F(-1)$.

The function $F(y_1)$ is differentiable hence continuous. So:

- when we take $y_1 > 0$ and $y_1 \to 0$, we conclude that, in the limit, that $F(0) = F(1)$;
- when we take $y_1 < 0$ and $y_1 \to 0$, we conclude that, in the limit, that $F(0) = F(-1)$.

Since both $F(1)$ and $F(-1)$ are equal to the same number $F(0)$, they are thus equal, so $F(y_1) = F(1)$ for all $y_1$. So, the value $f(y_1, y_2) = F(y_1 - y_2)$ is also always equal to the constant $F(1)$. The first part of Proposition 3 is proven.

2°. Let us now consider the case of shift-invaraince and scale-covarience. It is easy to check that each operation $C \cdot (y_1 - y_2)$ is both shift-invariant and scale-covariant. So, to prove this part of the proposition, we need to prove that every shift-invariant and scale-covariant operation has this form.

Similarly to Part 1 of this proof, from shift-invariance, we conclude that $f(y_1, y_2) = F(y_1 - y_2)$, where $F(y) \stackrel{\text{def}}{=} f(y, 0)$. For $y_2 = 0$, scale-covariance means that $f(C \cdot y_1, 0) = C \cdot f(y_1, 0)$, i.e., that $F(C \cdot y_1) = C \cdot F(y_1)$. Now:

- for $y_1 > 0$, we can take $C = 1/y_1$ and thus conclude that $F(1) = (1/y_1) \cdot F(y_1)$, hence $F(y_1) = F(1) \cdot y_1$;
- for $y_1 < 0$, we can take $C = -1/y_1$ and thus conclude that $F(-1) = (-1/y_1) \cdot F(-1)$, hence $F(y_1) = -F(-1) \cdot y_1$.

The function $F(y_1)$ is differentiable hence continuous, so in the limit, we get $F(0) = 0$.

In particular, the function $F(y_1)$ is differentiable for $y_1 = 0$, i.e., we have a limit

$$F'(0) = \lim_{y_1 \to 0} \frac{F(y_1) - F(0)}{y_1}.$$

- when we take $y_1 > 0$ and $y_1 \to 0$, we conclude that, in the limit, that $F'(0) = F(1)$;
- when we take $y_1 < 0$ and $y_1 \to 0$, we conclude that, in the limit, that $F'(0) = -F(-1)$.

Since both $F(1)$ and $-F(-1)$ are equal to the same number $F'(0)$, they are thus equal, so $F(y_1) = F(1) \cdot y_1$ for all $y_1$. So, the value $f(y_1, y_2) = F(y_1 - y_2) = F(1) \cdot (y - y_2)$ has the desired form for $C = F(1)$. The second part of Proposition 3 is proven.

3°. Let us prove, by contradiction, that no operation can be both shift-covariant and scale-invariant.

Indeed, let $f(y_1, y_2)$ be such an operation. For $C = -y_2$, shift-covariance implies that $f(y_1 - y_2, 0) = f(y_1, y_2) - y_2$, i.e., that $f(y_1, y_2) = f(y_1 - y_2, 0) + y_2$. In particular, $f(C \cdot y_1, C \cdot y_2) = f(C \cdot (y_1 - y_2), 0) + C \cdot y_2$. Thus, scale-invariance means that $f(C \cdot (y_1 - y_2), 0) + C \cdot y_2 = f(y_1 - y_2, 0) + y_2$. In particular, for $y_1 = y_2 + 1$, we get $y_1 - y_2 = 1$, so $f(C, 0) + C \cdot y_2 = f(1, 0) + y_2$, i.e., equivalently, $f(C, 0) - f(1, 0) = (1 - C) \cdot y_2$. For $C = 0.5$, this implies that $f(0.5, 0) - f(1, 0) = 0.5 \cdot y_2$, i.e., $y_2 = 2 \cdot (f(0.5, 0) - f(1, 0))$. This must be true fo all $y_2$, but if we take a value $y_2$ which is different from the right-hand side, this equality is no longer true. So, the assumption that such an operation is possible leads to a contradiction. The third part of the proposition is also proven.

4°. Finally, let us consider the case of shift-covaraince and scale-covarience. It is easy to check that all the operations $C \cdot y_1 + (1 - C) \cdot y_2$ are both shift-covariant and scale-covariant. So, to prove this part of the proposition, we need to prove that every shift-covariant and scale-covariant operation has this form.

Indeed, similarly to Part 3 of this proof, shift-covariance implies that

$$f(y_1, y_2) = f(y_1 - y_2, 0) + y_2.$$

For this expression, scale-covariance leads to $f(C \cdot (y_1 - y_2), 0) + C \cdot y_2 = C \cdot f(y_1 - y_2, 0) + C \cdot y_2$. Subtracting $C \cdot y_2$ from both sides, we conclude that $F(C \cdot (y_1 - y_2)) = C \cdot F(y_1 - y_2)$, where we denoted $F(y) \stackrel{\text{def}}{=} f(y, 0)$. From Part 2 of this proof, we know that this implies that $F(y) = C \cdot y$ for some constant $C$. Thus,

$$f(y_1 - y_2) = F(y_1 - y_2) + y_2 = C \cdot (y_1 - y_2) + y_2 = C \cdot y_1 + (1 - C) \cdot y_2.$$

This part is also proven, and so is the proposition.

# 9 Let us apply optimal binary operations to optimal functions of one variable: this explains the success of Exp-Minus-Log

We want to get an optimal function of two variables by applying an optimal binary operation $f(y_1, y_2)$ to two optimal functions $f_1(x_1)$ and $f_2(x_2)$, resulting in $g(x_1, x_2) = f(f_1(x_1), f_2(x_2))$. Which operation should we apply and to which functions of one variable?

Applying a constant operation does not make any sense: we just get a constant function $g(x_1, x_2) = \text{const}$ based on which we cannot compute anything. So, we are left with subtraction and Hurwicz combination. Maybe Hurwicz combination will also work, but clearly, subtraction is, from the computational viewpoint, the simplest of the two.

To which of the optimal functions of one variables should we apply our binary operation? In line with the above result relating optimality with invariance, optimality implies that we want a function that is both scale- and shift-invariant. In this requirement both scale- and shift-invariances appear in a similar way, so it makes sense to select a pair of optimal functions of one variable in such a way that this pair will be invariant with respect to swapping scale- and shift-invariance. One can see that there are only two such pairs:

- if we pick one of the functions to be shift-invariant under shifts of the input, the second function should be scale-invariant with respect to scaling of the input; for this pair, we need to combine $x$ and $x^{\alpha}$;

- similarly, if we pick one of the functions to be scale-invariant under shift of the input, the second function should be shift-invariant with respect to scaling of the input; for this pair, we need to combine $\exp(x)$ and $\ln(x)$.

From the commonsense viewpoint, the second options looks more promising, since:

- in the first pair, one of the function is the identity $f_1(x_1) = x_1$ (or a simple linear function), while
- in the second pair, both functions are non-trivial.

We can apply each of the two optimal binary operations – difference and Hurwicz combination – to each of these two pairs. We thus get $2 \cdot 2 = 4$ possible functions of two variables. Out of these four, if we apply the simplest optimal binary operation to the most promising pair of optimal functions of one variable, we get exactly the desired function $\exp(x) - \ln(y)$. So, we have indeed provided some explanation of this function's success.

## Acknowledgments

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

## References

[1] Th. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT Press, Cambridge, Massachusetts, 2022.

[2] R. Feynman, R. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Addison Wesley, Boston, Massachusetts, 2005.

[3] L. Hurwicz, *Optimality Criteria for Decision Making Under Ignorance*, Cowles Commission Discussion Paper, Statistics, No. 370, 1951.

[4] V. Kreinovich, "Decision making under interval uncertainty (and beyond)", In: P. Guo and W. Pedrycz (eds.), *Human-Centric Decision-Making Models for Social Sciences*, Springer Verlag, 2014, pp. 163–193.

[5] R. D. Luce and R. Raiffa, *Games and Decisions: Introduction and Critical Survey*, Dover, New York, 1989.

[6] H. T. Nguyen and V. Kreinovich, *Applications of Continuous Mathematics to Computer Science*, Kluwer, Dordrecht, 1997.

[7] A. Odrzywolek, *All elementary functions from a single operator*, arXiv:2603.21852v2, April 2026.

[8] K. S. Thorne and R. D. Blandford, *Modern Classical Physics: Optics, Fluids, Plasmas, Elasticity, Relativity, and Statistical Physics*, Princeton University Press, Princeton, New Jersey, 2021.