Why Derivative: Invariance-Based Explanation

Julio Urenda[1,2], Olga Kosheleva[3], and Vladik Kreinovich[2]
[1]Department of Mathematical Science
[2]Department of Computer Science
[3]Department of Teacher Education
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
jcurenda@utep.edu, olgak@utep.edu, vladik@utep.edu

Abstract

To many students, the notion of a derivative seems unrelated to any previous mathematics – and is, thus, difficult to study and to understand. In this paper, we show that this notion can be naturally derived from a more intuitive notion of invariance.

1 Formulation of the Problem

To a student studying mathematics, the notion of the derivative seems to appear out of nowhere, without any explanation and without any reasonable relation with previously studied mathematical notions. This un-relatedness may be one of the reasons why calculus is so difficult for many students, even for those who have successfully studied previous mathematical subjects.

In this paper, we show that the notion of the derivative can be explained by the natural ideas of invariance. We hope that this explanation will make this notion more natural – and thus, easier to learn.

2 Natural Notions of Invariance

Natural transformations. Many numbers used in mathematical computations represent values of physical quantities such as time, coordinate, distance, etc. The corresponding numbers, however, depend on what unit we choose to measure the corresponding quantity, and on what starting point we choose.

For example, we can measure time in minutes or in seconds. If we started with time in minutes, then, to get time in seconds, we need to multiply all numerical values by 60:

- 1 minute becomes 60 seconds,

- 2 minutes becomes $2 \cdot 60 = 120$ seconds, etc.

In general, if we replace original measuring unit by a new unit which is a times smaller, then all numerical values of the corresponding physical quantity will be multiplied by a, i.e., instead of the original numerical value x, we will have a new numerical value $\widetilde{x} = a \cdot x$.

Similarly, if we start measuring time not from the original starting point, but from a moment which is b seconds earlier, then we need to add b to all the numerical values: $x \to \widetilde{x} = x + b$.

In general, if we change both the measuring unit and the starting point, we then replace the original numerical values x with new values $\widetilde{x} = a \cdot x + b$.

Another natural transformation is the change in direction. For example, we can point the coordinate axis in an opposite direction, in which case each value x is changed to $-x$. Another example of such sign-invariance is electricity, where which charges we call positive and which negative is simply a matter of convenience: nothing changes if we simply rename positive into negative and vice versa.

Invariance: general idea. The physics does not change if we simply change the measuring unit or change the starting point. Thus, it is often reasonable to require that the corresponding mathematical model also not change.

Comment. In general, invariance – including invariance with respect to more complex transformations – is one of the main concepts in modern physics; see, e.g., [2, 3].

Which dependencies are invariant. From this viewpoint, let us consider which dependencies $y = f(x)$ between two physical quantities x and y are invariant. Of course, since these properties are related, if we change the unit for measuring x, we may need to also change the unit for measuring y. For example, if we change the unit for length (e.g., from meters to centimeters), then, to preserve the formulas describing areas and volumes, we need to correspondingly change units for area and volume: from square meters to square centimeters and from cubic meters to cubic centimeters.

Similarly, if we change the starting point for the quantity x, we may need to correspondingly change the starting point for y.

A natural case to consider is the dependence $y = f(x)$ for which, for each re-scaling of the x-scale, there is a corresponding re-scaling of the y-scale in which the dependence looks exactly the same. In other words, for every a_x and b_x there exist such values a_y and b_y that for each x and y, $y = f(x)$ implies that $\widetilde{y} = f(\widetilde{x})$, where $\widetilde{x} = a_x \cdot x + b_x$ and $\widetilde{y} = a_y \cdot y + b_y$.

It turns out that among all continuous dependencies – or, even more generally, among all the functions $f(x)$ which are, in some reasonable sense, definable – the only functions $f(x)$ satisfying this invariance property are linear functions

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

For linear functions, invariance is easy to prove. Indeed, suppose that $y = a \cdot x + b$. Multiplying both sides by a_x, we conclude that $a_x \cdot y = a \cdot (a_x \cdot x) + a_x \cdot b$.

Here, $a_x \cdot x = \widetilde{x} - b_x$, so we get $a_x \cdot y = a_x \cdot \widetilde{x} + a_x \cdot b - a \cdot b_x$. If we add a constant $c = b - (a_x \cdot b - a \cdot b_x)$ to both sides of this equality, we conclude that $a_x \cdot y + c = a \cdot \widetilde{x} + b$, i.e., that $\widetilde{y} = a \cdot \widetilde{x} + b$, where the coefficients in the expression $\widetilde{y} = a_y \cdot y + b_y$ are equal to $a_y = a_x$ and $b_y = c$.

That only linear functions have this property is more difficult to prove; see, e.g., [1].

3 Invariance Naturally Leads to the Derivative

Let us start the construction. Now, we are ready to show that the natural notions of invariance indeed lead to the expression for the derivative. We will do it step-by-step, adding more invariance requirements as we go.

We have a function $y = f(x)$. Based on the values of this function, we want to build a new auxiliary function $g(x)$. Let us consider the simplest case when at each point x, the value of the new function $g(x)$ will depend only on two values of the original function $f(x)$. In other words, we consider the case when

$$g(x) = F(f(p_1(x)), f(p_2(x)), \tag{2}$$

where:

- $p_1(x)$ and $p_2(x)$ describe how these two points depend on x, and
- $F(y, z)$ is an algorithm that transforms the corresponding two values of the function $f(x)$ into the value $g(x)$ of the new function.

First invariance requirement: invariance with respect to x-shifts. The first natural invariance requirement that we will impose is *x-shift-invariance*: if we use a different starting point for measuring x, the expressions for the corresponding dependencies $p_1(x)$ and $p_2(x)$ should not change. Let us describe this requirement in precise terms.

Each expression $p_i(x)$ describes how the value of the corresponding point x_i in the original x-scale depends on the value of the parameter x in the same scale. If we change the starting point, then each original value x will take the new form $\widetilde{x} = x + b$, so that $x = \widetilde{x} - b$, and the point $x_i = p_i(x)$ at which we should compute $f(x)$ will take a new form $\widetilde{x}_i = p_i(x) + b$. Substituting the expression $x = \widetilde{x} - b$ into this formula, we conclude that in the new scale, the dependence of the corresponding point $\widetilde{x}_i$ on $\widetilde{x}$ should take the form $\widetilde{x}_i = p_i(\widetilde{x} - b) + b$. Invariance means that this dependence should be expressed by the same formula as in the original scale, i.e., we should have $\widetilde{x}_i = p_i(\widetilde{x})$.

Comparing these two expressions, we conclude that $p_i(\widetilde{x} - b) + b = p_i(\widetilde{x})$ for all b and $\widetilde{x}_i$. In particular, for $b = \widetilde{x}$, we conclude that $p_i(\widetilde{x}) = \widetilde{x} + p_i(0)$. Thus, due to this invariance requirements, each function $p_i(x)$ has the form $p_i(x) = x + \text{const}$. Let us denote the corresponding constant by c_i. Then, we have $p_i(x) = x + c_i$, and the formula (2) takes the form

$$g(x) = F(f(x + c_1), f(x + c_2)). \tag{3}$$

This expression is simpler than the original expression (2):

- in the original expression (2), we had three unknown functions $F(y_1, y_2)$, $p_1(x)$, and $p_2(x)$, while
- now, we have only one unknown function $F(y_1, y_2)$ and two unknown numbers c_1 and c_2.

Second invariance requirement: invariance with respect to y-shifts. Another reasonable requirement is that the values $g(x)$ should not change if we simply change the starting point for measuring y. As we have mentioned earlier, this change simply adds the same constant b to all the y-values – i.e., in our case, to both values of the function $f(x)$. Thus, instead of the original value $F(f(x+c_1), f(x+c_2))$, we will have a new value $F(f(x+c_1)+b, f(x+c_2)+b)$. Invariance means that these two values must coincide, i.e., that we should have

$$g(x) = F(f(x+c_1), f(x+c_2)) = F(f(x+c_1)+b, f(x+c_2)+b)$$

for all x and b. In particular, for $b = -f(x+c_2)$, we have

$$g(x) = F(f(x+c_1), f(x+c_2)) = F(f(x+c_1) - f(x+c_2), 0),$$

i.e., equivalently, that

$$g(x) = G(f(x+c_1) - f(x+c_2)), \tag{4}$$

where we denoted $G(y) \stackrel{\text{def}}{=} F(y, 0)$. This expression is even simpler than the expression (3):

- in the expression (3), we had an unknown function $F(y_1, y_2)$ of two variables, while
- now, we have only an unknown function $G(y)$ of one variable.

Next invariance requirement: invariance with respect to y-scaling. Another natural invariance requirement is that the dependence (4) should not change if we change the unit in which we measure y-values like $f(x)$ or $g(x)$. In other words, if we have the expression (4) and we replace $f(x+c_i)$ with $\widetilde{f}(x+c_i) = a \cdot f(x+c_i)$ and $g(x)$ with $\widetilde{g}(x) = c \cdot g(x)$, then we should have the same relation between the re-scaled values, i.e., we should have

$$\widetilde{g}(x) = G\left(\widetilde{f}(x+c_1) - \widetilde{f}(x+c_2)\right).$$

In other words, we should have

$$G(\lambda \cdot f(x+c_1) - \lambda \cdot f(x+c_2)) = \lambda \cdot g(x) = \lambda \cdot G(f(x+c_1) - f(x+c_2)),$$

i.e.,

$$G(\lambda \cdot (f(x+c_1) - f(x+c_2))) = \lambda \cdot G(f(x+c_1) - f(x+c_2)).$$

Since the difference $z \stackrel{\text{def}}{=} f(x+c_1) - f(x+c_2)$ can take any possible real value, we thus have

$$G(\lambda \cdot z) = \lambda \cdot G(z).$$

In particular, for $z = 1$, we conclude that $G(\lambda) = \lambda \cdot G(1)$, i.e., that $G(\lambda) = K \cdot \lambda$, where we denoted $K \stackrel{\text{def}}{=} G(1)$. For this function $G(z)$, the formula (4) takes an even simpler form

$$g(x) = K \cdot (f(x+c_1) - f(x+c_2)). \tag{5}$$

We can have different expressions like that, for different values c_1 and c_2. In general, the coefficient K may depend on which values c_1 and c_2 we select, so we get

$$g(x) = K(c_1, c_2) \cdot (f(x+c_1) - f(x+c_2)). \tag{5}$$

Which values $K(c_1, c_2)$ should we choose? In general, the value of the expression (5) changes when we change the values c_1 and c_2. In particular, this is true even if we consider the invariant dependencies $f(x)$ – which, as we have shown in the previous section, correspond to linear functions $f(x) = a \cdot x + b$.

For a linear function $f(x) = a \cdot x + b$, the expression (5) takes the form

$$g(x) = K(c_1, c_2) \cdot ((a \cdot (x+c_1) + b) - (a \cdot (x+c_2) + b)) = K(c_1, c_2) \cdot a \cdot (c_1 - c_2) = a \cdot c,$$

where we denoted $c \stackrel{\text{def}}{=} K(c_1, c_2) \cdot (c_1 - c_2)$.

Thus, it is possible to select the coefficient $K(c_1, c_2)$ in such a way that for linear functions, the resulting value $g(x)$ will not depend on the selection of c_1 and c_2. Namely, to make sure the product $c = K(c_1, c_2) \cdot (c_1 - c_2)$ remains the same for all c_1 and c_2, we should select $K(c_1, c_2) = \dfrac{c}{c_1 - c_2}$. In this case, the expression (5) takes the form

$$g(x) = c \cdot \frac{f(x+c_1) - f(x+c_2)}{c_1 - c_2}. \tag{6}$$

Which values c_1 and c_2 should we choose? A reasonable idea is to consider *local* characteristics, i.e., characteristics $g(x)$ that depend only the values of the original function $f(x)$ in a small vicinity of the point x: e.g., in the ε-vicinity of all the points which are ε-close to x. Thus, we consider cases when the values c_1 and c_2 are small: e.g., $|c_i| \le \varepsilon$ for $i = 1, 2$.

As we consider the smaller and smaller neighborhoods, the values c_i tend to 0 and thus, we get the value

$$g(x) = c \cdot \lim_{c_1, c_2 \to 0} \frac{f(x+c_1) - f(x+c_2)}{c_1 - c_2}. \tag{7}$$

Modulo a multiplicative constant c, this is exactly the derivative – i.e., exactly the expression that we wanted to explain.

How can we describe the above expression for the derivative in a more standard form. While the expression (7) is equal to the derivative (modulo c), it is *different* from the standard definitions of the derivative. We can make it closer if we impose an additional invariance requirement: that the formula (6) (and thus, the formula (7)) should not change if we replace c with $\widetilde{x} = -x$. In this case, we have $x = -\widetilde{x}$, so, instead of the original function $f(x)$, we get a new function $\widetilde{f}\,(\widetilde{x}) \stackrel{\text{def}}{=} f\,(-\widetilde{x})$. If we apply the formula (6) to this new function, we get the expression

$$\widetilde{g}(\widetilde{x}) = c \cdot \frac{\widetilde{f}\,(\widetilde{x} + c_1) - \widetilde{f}\,(\widetilde{x} + c_2)}{c_1 - c_2}.$$

Invariance means that, when we substitute the formulas for $\widetilde{f}(z)$ and for $\widetilde{x} = -x$ into this expression, we should get the same formula (6). Here,

$$\widetilde{f}\,(\widetilde{x} + c_i) = f\,(-\,(\widetilde{x} + c_i)) = f(-(-x + c_i)) = f(x - c_i),$$

thus the desired equality takes the form:

$$\frac{f(x - c_1) - f(x - c_2)}{c_1 - c_2} = \frac{f(x + c_1) - f(x + c_2)}{c_1 - c_2}.$$

This equality should be satisfied for all possible functions $f(x)$. Thus, the left-hand side should use the values of the function $f(x)$ at exactly the same two points as the right-hand side. The only two possible options for this equality are:

- the case when $c_1 = -c_1$ and $c_2 = -c_2$, and
- the case when $c_1 = -c_2$ and $c_2 = -c_1$.

In the first case, we get $c_1 = c_2 = 0$ and thus, $g(x)$ is always equal to 0. The only non-trivial case is the second case, in which case (6) takes the form

$$g(x) = c \cdot \frac{f(x + h) - f(x - h)}{2h}, \tag{8}$$

where we denoted $h \stackrel{\text{def}}{=} c_1$. In this case, the limit expression (7) turns into one of the often-used versions of the standard definition of the derivative:

$$g(x) = c \cdot \lim_{h \to 0} \frac{f(x + h) - f(x - h)}{2h}. \tag{9}$$

Acknowledgments

This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science) and HRD-1242122 (Cyber-ShARE Center of Excellence).

References

[1] J. Aczel and J. Dhombres, *Functional Equations in Several Variables*, Cambridge University Press, Cambridge, UK, 1989.

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

[3] K. S. Thorne and R. D. Blandford, *Modern Classical Physics: Optics, fluids, Plasmas, Elasticity, Relativity, and Statistical Physics*, Princeton University Press, Princeton, New Jersey, 2017.