Mathematical Structures and Modeling
2017. N. 3(43). PP. 43–49

UDC 519.652
DOI: 10.25513/2222-8772.2017.3.43-49

WHY LINEAR INTERPOLATION?

Andrzej Pownuk
Ph.D. (Phys.-Math.), Instructor, e-mail: ampownuk@utep.edu
Vladik Kreinovich
Ph.D. (Phys.-Math.), Professor, e-mail: vladik@utep.edu

University of Texas at El Paso, El Paso, Texas 79968, USA

Abstract. Linear interpolation is the computationally simplest of all possible interpolation techniques. Interestingly, it works reasonably well in many practical situations, even in situations when the corresponding computational models are rather complex. In this paper, we explain this empirical fact by showing that linear interpolation is the only interpolation procedure that satisfies several reasonable properties such as consistency and scale-invariance.

Keywords: linear interpolation, scale-invariance.

1. Formulation of the Problem

Need for interpolation. In many practical situations, we know that the value of a quantity y is uniquely determined by the value of some other quantity x, but we do not know the exact form of the corresponding dependence $y = f(x)$.

To find this dependence, we measure the values of x and y in different situations. As a result, we get the values $y_i = f(x_i)$ of the unknown function $f(x)$ for several values $x_1, \ldots, x_n$. Based on this information, we would like to predict the value $f(x)$ for all other values x. When x is between the smallest and the largest of the values x_i, this prediction is known as the *interpolation*; for values x smaller than the smallest of x_i or larger than the largest of x_i, this prediction is known as *extrapolation*; see, e.g., [1].

Simplest possible case of interpolation. The simplest possible case of interpolation is when we only know the values $y_1 = f(x_1)$ and $y_2 = f(x_2)$ of the function $f(x)$ at two points $x_1 < x_2$, and we would like to predict the value $f(x)$ at points $x \in (x_1, x_2)$.

In many cases, linear interpolations works well: why? One of the most well-known interpolation techniques is based on the assumption that the function $f(x)$ is linear on the interval $[x_1, x_2]$. Under this assumption, we get the following formula for $f(x)$:

$$f(x) = \frac{x - x_1}{x_2 - x_1} \cdot f(x_2) + \frac{x_2 - x}{x_2 - x_1} \cdot f(x_1).$$

This formula is known as *linear interpolation*.

The usual motivation for linear interpolation is simplicity: linear functions are the easiest to compute, and this explains why we use linear interpolation.

An interesting empirical fact is that in many practical situations, linear interpolation works reasonably well. We know that in computational science, often very complex computations are needed, so we cannot claim that nature prefers simple functions. There should be another reason for the empirical fact that linear interpolation often works well.

What we do. In this paper, we show that linear interpolation can indeed be derived from fundamental principles.

2. Analysis of the Problem: What Are Reasonable Properties of an Interpolation

What is interpolation. We want to be able, given values y_1 and y_2 of the unknown function at points x_1 and x_2, and a point $x \in (x_1, x_2)$, to provide an estimate for $f(x)$. In other words, we need a function that, given the values x_1, y_1, x_2, y_2, and x, generates the estimate for $f(x)$. We will denote this function by $I(x_1, y_1, x_2, y_2, x)$.

What are the reasonable properties of this function?

Conservativeness. If both observed values $y_i = f(x_i)$ are smaller than or equal to some threshold value y, it is reasonable to expect that all intermediate values of $f(x)$ should also be smaller than or equal to y. Thus, if $y_1 \leqslant y$ and $y_2 \leqslant y$, then we should have $I(x_1, y_1, x_2, y_2, x) \leqslant y$.

In particular, for $y = \max(y_1, y_2)$, we conclude that

$$I(x_1, y_1, x_2, y_2, x) \leqslant \max(y_1, y_2).$$

Similarly, if both observed values $y_i = f(x_i)$ are greater than or equal to some threshold value y, it is reasonable to expect that all intermediate values of $f(x)$ should also be greater than or equal to y. Thus, if $y \leqslant y_1$ and $y \leqslant y_2$, then we should have $y \leqslant I(x_1, y_1, x_2, y_2, x)$.

In particular, for $y = \min(y_1, y_2)$, we conclude that

$$\min(y_1, y_2) \leqslant I(x_1, y_1, x_2, y_2, x)$$

These two requirements can be combined into a single double inequality

$$\min(y_1, y_2) \leqslant I(x_1, y_1, x_2, y_2, x) \leqslant \max(y_1, y_2).$$

We will call this property *conservativeness*.

x-scale-invariance. The numerical value of a physical quantity depends on the choice of the measuring unit and on the starting point. If we change the starting point to the one which is b units smaller, then b is added to all the numerical values. Similarly, if we replace a measuring unit by a one which is $a > 0$ times smaller, then all the numerical values are multiplied by a. If we perform both changes, then each original value x is replaced by the new value $x' = a \cdot x + b$.

For example, if we know the temperature x in Celsius, then the temperature x' in Fahrenheit can be obtained as $x' = 1.8 \cdot x + 32$.

It is reasonable to require that the interpolation procedure should not change if we simply change the measuring unit and the starting point — without changing the actual physical quantities. In other words, it is reasonable to require that

$$I(a \cdot x_1 + b, y_1, a \cdot x_2 + b, y_2, a \cdot x + b) = I(x_1, y_1, x_2, y_2, x).$$

y-scale-invariance. Similarly, we can consider different units for y. The interpolation result should not change if we simply change the starting point and the measuring unit. So, if we replace y_1 with $a \cdot y_1 + b$ and y_2 with $a \cdot y_2 + b$, then the result of interpolation should be obtained by a similar transformation from the previous result: $I \to a \cdot I + b$. Thus, we require that

$$I(x_1, a \cdot y_1 + b, x_2, a \cdot y_2 + b, x) = a \cdot I(x_1, y_1, x_2, y_2, x) + b.$$

Consistency. Let us assume that we have $x_1 \leqslant x_1' \leqslant x \leqslant x_2' \leqslant x_2$. Then, the value $f(x)$ can be estimated in two different ways:

- we can interpolate directly from the values $y_1 = f(x_1)$ and $y_2 = f(x_2)$, getting $I(x_1, y_1, x_2, y_2, x)$, or
- we can first use interpolation to estimate the values $f(x_1') = I(x_1, y_1, x_2, y_2, x_1')$ and $f(x_2') = I(x_1, y_1, x_2, y_2, x_2')$, and then use these two estimates to estimate $f(x)$ as

$$I(x_1, f(x_1'), x_2, f(x_2'), x) =$$

$$= I(x_1', I(x_1, y_1, x_2, y_2, x_1'), x_2', I(x_1, y_1, x_2, y_2, x_2'), x).$$

It is reasonable to require that these two ways lead to the same estimate for $f(x)$:

$$I(x_1, y_1, x_2, y_2, x) = I(x_1', I(x_1, y_1, x_2, y_2, x_1'), x_2', I(x_1, y_1, x_2, y_2, x_2'), x).$$

Continuity. Most physical dependencies are continuous. Thus, when the two value x and x' are close, we expect the estimates for $f(x)$ and $f(x')$ to be also close. Thus, it is reasonable to require that the interpolation function $I(x_1, y_1, x_2, y_2, x)$ is continuous in x — and that for both $i = 1, 2$ the value $I(x_1, y_1, x_2, y_2, x)$ converges to $f(x_i)$ when $x \to x_i$.

Now, we are ready to formulate our main result.

3. Main Result

Definition 1. *By an* interpolation function*, we mean a function* $I(x_1, y_1, x_2, y_2, x)$ *which is defined for all* $x_1 < x < x_2$ *and which has the following properties:*

- *conservativeness:*

$$\min(y_1, y_2) \leqslant I(x_1, y_1, x_2, y_2, x) \leqslant \max(y_1, y_2)$$

  for all x_i*,* y_i*, and* x*;*
- x*-scale-invariance:* $I(a \cdot x_1 + b, y_1, a \cdot x_2 + b, y_2, a \cdot x + b) = I(x_1, y_1, x_2, y_2, x)$ *for all* x_i*,* y_i*,* x*,* $a > 0$*, and* b*;*
- y*-scale invariance:* $I(x_1, a \cdot y_1 + b, x_2, a \cdot y_2 + b, x) = a \cdot I(x_1, y_1, x_2, y_2, x) + b$ *for all* x_i*,* y_i*,* x*,* $a > 0$*, and* b*;*
- *consistency:*

$$I(x_1, y_1, x_2, y_2, x) = I(x_1', I(x_1, y_1, x_2, y_2, x_1'), x_2', I(x_1, y_1, x_2, y_2, x_2'), x)$$

  for all x_i*,* x_i'*,* y_i*, and* x*; and*
- *continuity: the expression* $I(x_1, y_1, x_2, y_2, x)$ *is a continuous function of* x*,* $I(x_1, y_1, x_2, y_2, x) \to y_1$ *when* $x \to x_1$ *and* $I(x_1, y_1, x_2, y_2, x) \to y_2$ *when* $x \to x_2$*.*

Proposition. *The only interpolation function satisfying all the properties from Definition 1 is the linear interpolation*

$$I(x_1, y_1, x_2, y_2, x) = \frac{x - x_1}{x_2 - x_1} \cdot y_2 + \frac{x_2 - x}{x_2 - x_1} \cdot y_1. \tag{1}$$

Discussion. Thus, we have indeed explained that linear interpolation follows from the fundamental principles – which may explain its practical efficiency.

Proof.

1°. When $y_1 = y_2$, the conservativeness property implies that $I(x_1, y_1, x_2, y_1, x) = y_1$. Thus, to complete the proof, it is sufficient to consider two remaining cases: when $y_1 < y_2$ and when $y_2 < y_1$.

We will consider the case when $y_1 < y_2$. The case when $y_2 < y_1$ is considered similarly. So, in the following text, without losing generality, we assume that $y_1 < y_2$.

2°. When $y_1 < y_2$, then we can get these two values y_1 and y_2 as $y_1 = a \cdot 0 + b$ and $y_2 = a \cdot 1 + b$ for $a = y_2 - y_1$ and $b = y_1$. Thus, the y-scale-invariance implies that

$$I(x_1, y_1, x_2, y_2, x) = (y_2 - y_1) \cdot I(x_1, 0, x_2, 1, x) + y_1. \tag{2}$$

If we denote $J(x_1, x_2, x) \stackrel{\text{def}}{=} I(x_1, 0, x_2, 1, x)$, then we get

$$I(x_1, y_1, x_2, y_2, x) = (y_2 - y_1) \cdot J(x_1, x_2, x) + y_1 =$$

$$= J(x_1, x_2, x) \cdot y_2 + (1 - J(x_1, x_2, x)) \cdot y_1. \tag{3}$$

3°. Since $x_1 < x_2$, we can similarly get these two values x_1 and x_2 as $x_1 = a \cdot 0 + b$ and $x_2 = a \cdot 1 + b$, for $a = x_2 - x_1$ and $b = x_1$. Here, $x = a \cdot r + b$, where

$$r = \frac{x - b}{a} = \frac{x - x_1}{x_2 - x_1}.$$

Thus, the x-scale invariance implies that

$$J(x_1, x_2, x) = J\left(0, 1, \frac{x - x_1}{x_2 - x_1}\right).$$

So, if we denote $w(r) \stackrel{\text{def}}{=} J(0, 1, r)$, we then conclude that

$$J(x_1, x_2, x) = w\left(\frac{x - x_1}{x_2 - x_1}\right),$$

and thus, the above expression (3) for $I(x_1, y_1, x_2, y_2, x)$ in terms of $J(x_1, x_2, x)$ takes the following simplified form:

$$I(x_1, y_1, x_2, y_2, x) = w\left(\frac{x - x_1}{x_2 - x_1}\right) \cdot y_2 + \left(1 - w\left(\frac{x - x_1}{x_2 - x_1}\right) \cdot y_2\right) \cdot y_1. \qquad (4)$$

To complete our proof, we need to show that $w(r) = r$ for all $r \in (0, 1)$.

4°. Let us now use consistency.

Let us take $x_1 = y_1 = 0$ and $x_2 = y_2 = 1$, then

$$I(0, 0, 1, 1, x) = w(x) \cdot 1 + (1 - w(x)) \cdot 0 = w(x).$$

Let us denote $\alpha \stackrel{\text{def}}{=} w(0.5)$.

By consistency, for $x = 0.25 = \dfrac{0 + 0.5}{2}$, the value $w(0.25)$ can be obtained if we apply the same interpolation procedure to $w(0) = 0$ and to $w(0.5) = \alpha$. Thus, we get

$$w(0.25) = \alpha \cdot w(0.5) + (1 - \alpha) \cdot w(0) = \alpha^2.$$

Similarly, for $x = 0.75 = \dfrac{0.5 + 1}{2}$, the value $w(0.75)$ can be obtained if we apply the same interpolation procedure to $w(0.5) = \alpha$ and to $w(1) = 1$. Thus, we get

$$w(0.75) = \alpha \cdot w(1) + (1 - \alpha) \cdot w(0.5) = \alpha \cdot 1 + (1 - \alpha) \cdot \alpha = 2\alpha - \alpha^2.$$

Finally, for $x = 0.5 = \dfrac{0.25 + 75}{2}$, the value $w(0.5)$ can be obtained if we apply the same interpolation procedure to $w(0.25) = \alpha^2$ and to $w(0.75) = 2\alpha - \alpha^2$. Thus, we get

$$w(0.5) = \alpha \cdot w(0.75) + (1 - \alpha) \cdot w(0.25) =$$
$$= \alpha \cdot (2\alpha - \alpha^2) + (1 - \alpha) \cdot \alpha^2 = 3\alpha^2 - 2\alpha^3.$$

By consistency, this estimate should be equal to our original estimate $w(0.5) = \alpha$, i.e., we must have

$$3\alpha^2 - 2\alpha^3 = \alpha. \tag{5}$$

5°. One possible solution is to have $\alpha = 0$. In this case, we have $w(0.5) = 0$. Then, we have

$$w(0.75) = \alpha \cdot w(1) + (1 - \alpha) \cdot w(0.5) = 0,$$

and by induction, we can show that in this case, $w(1 - 2^{-n}) = 0$ for each n. In this case, $1 - 2^{-n} \to 1$, but $w(1 - 2^{-n}) \to 0$, which contradicts to the continuity requirement, according to which $w(1 - 2^{-n}) \to w(1) = 1$.

Thus, the value $\alpha = 0$ is impossible, so $\alpha \neq 0$, and we can divide both sides of the above equality (5) by α.

As a result, we get a quadratic equation

$$3\alpha - 2\alpha^2 = 1,$$

which has two solutions: $\alpha = 1$ and $\alpha = 0.5$.

6°. When $\alpha = 1$, we have $w(0.5) = 1$. Then, we have

$$w(0.25) = \alpha \cdot w(0.5) + (1 - \alpha) \cdot w(0) = 1,$$

and by induction, we can show that in this case, $w(2^{-n}) = 1$ for each n. In this case, $2^{-n} \to 0$, but $w(2^{-n}) \to 1$, which contradicts to the continuity requirement, according to which $w(2^{-n}) \to w(0) = 0$.

Thus, the value $\alpha = 1$ is impossible, so $\alpha = 0.5$.

7°. For $\alpha = 0.5$, we have $w(0) = 0$, $w(0.5) = 0.5$, and $w(1) = 1$. Let us prove, by induction over q, that for every binary-rational number $r = \dfrac{p}{2^q} \in [0, 1]$, we have $w(r) = r$.

Indeed, the base case $q = 1$ is proven. Let us assume that we have proven it for $q - 1$, let us prove it for q. If p is even, i.e., if $p = 2k$, then $\dfrac{2k}{2^q} = \dfrac{k}{2^{q-1}}$, so the desired equality comes from the induction assumption. If $p = 2k + 1$, then

$$r = \frac{p}{2^q} = \frac{2k+1}{2^q} = 0.5 \cdot \frac{2k}{2^q} + 0.5 \cdot \frac{2 \cdot (k+1)}{2^q} = 0.5 \cdot \frac{k}{2^{q-1}} + 0.5 \cdot \frac{k+1}{2^{q-1}}.$$

By consistency, we thus have

$$w(r) = 0.5 \cdot w\left(\frac{k}{2^{q-1}}\right) + 0.5 \cdot w\left(\frac{k+1}{2^{q-1}}\right).$$

By induction assumption, we have

$$w\left(\frac{k}{2^{q-1}}\right) = \frac{k}{2^{q-1}} \text{ and } w\left(\frac{k+1}{2^{q-1}}\right) = \frac{k+1}{2^{q-1}}.$$

So, the above formula takes the form

$$w(r) = 0.5 \cdot \frac{k}{2^{q-1}} + 0.5 \cdot \frac{k+1}{2^{q-1}},$$

hence $w(r) = \dfrac{2k+1}{2^q} = r$.

The statement is proven.

8°. The equality $w(r) = r$ is true for all binary-rational numbers. Any real number x from the interval $[0, 1]$ is a limit of such numbers — namely, truncates of its infinite binary expansion. Thus, by continuity, we have $w(x) = x$ for all x.

Substituting $w(x) = x$ into the above formula (4) for $I(x_1, y_1, x_2, y_2, x)$ leads exactly to linear interpolation.

The proposition is proven.

Acknowledgments

This work was supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721, and by an award "UTEP and Prudential Actuarial Science Academy and Pipeline Initiative" from Prudential Foundation.

References

1. Burden R.L., Faires J.D., Burden A.M. Numerical Analysis. Cengage Learning, Boston, Massachusetts, 2015.

ПОЧЕМУ ЛИНЕЙНАЯ ИНТЕРПОЛЯЦИЯ?

А. Повнук
к.ф.-м.н., ст. преподаватель, e-mail: ampownuk@utep.edu
В. Крейнович
к.ф.-м.н., профессор, e-mail: vladik@utep.edu

Техасский университет в Эль Пасо, США

Аннотация. Линейная интерполяция — это простейший в вычислительном отношении из всех возможных методов интерполяции. Интересно, что он работает достаточно хорошо во многих практических ситуациях, даже в ситуациях, когда соответствующие вычислительные модели довольно сложны. В этой статье мы объясняем этот эмпирический факт, показывая, что линейная интерполяция является единственной процедурой интерполяции, которая удовлетворяет нескольким разумным свойствам, таким как согласованность и масштабная инвариантность.

Ключевые слова: линейная интерполяция, масштабная инвариантность.

Дата поступления в редакцию: 09.04.2017