ISSN 1592-7415 **eISSN 2282-8214**

Volume n. 36 – 2019

RATIO MATHEMATICA

Journal of Mathematics, Statistics, and Applications

Publisher
A.P.A.V.
Accademia Piceno – Aprutina dei Velati in Teramo
www.eiris.it – www.apav.it
apavsegreteria@gmail.com

RATIO MATHEMATICA

Journal of Mathematics, Statistics, and Applications

ISSN 1592-7415 - eISSN 2282-8214

Ratio Mathematica is an **International, double peer-reviewed, open access journal, published every six months** (June-December).

The main topics of interest for Ratio Mathematica are:

Foundations of Mathematics: Epistemology of Mathematics, Critique of the Foundations of Mathematics, Elementary Mathematics from a higher point of view, Elementary Theory of Numbers, Foundations of Mathematics of Uncertain; Criticism of the Foundations of Computer Science.

Applications of Mathematics: Applications to Engineering and Economics, Applications for Management and Business Administration, Decision making in conditions of uncertainty, Theory of Games, Fuzzy Logic and Applications, Probability, Statistics and Applications.

New theories and applications: Algebraic hyperstructures and Applications, Discrete Mathematics and Applications, Fuzzy structures.

New theories and practices for dissemination and communication of mathematics: Communication of History and Foundations, Communication of Discrete Mathematics, Communication of Probability and Statistics, Communication with Media.

Note on Peer-Review
All manuscripts are subjected to a double-blind review process. The reviewers are selected from the editorial board, but they also can be external subjects. The journal's policies are described at: http://eiris.it/ojs/index.php/ratiomathematica/about/submissions#authorGuidelines

Publisher: **APAV** - Accademia Piceno-Aprutina dei Velati in Teramo
Tax Code 92036140678, **Vat Id. Number** 02184450688
Registered Office Address: via del Concilio, 24 - 65121 Pescara
Operational Office: via Chiarini, 191 - 65126 Pescara

Ratio Mathematica Volume 36, 2019, pp. 5-26

How to define and test explanations in populations

Peter J. Veazie*

Abstract

Solving applied social, economic, psychological, health care and public health problems can require an understanding of facts or phenomena related to populations of interest. Therefore, it can be useful to test whether an explanation of a phenomenon holds in a population. However, different definitions for the phrase "explain in a population" lead to different interpretations and methods of testing. In this paper, I present two definitions: The first is based on the number of members in the population that conform to the explanation's implications; the second is based on the total magnitude of explanation-consistent effects in the population. I show that claims based on either definition can be tested using random coefficient models, but claims based on the second definition can also be tested using the more common, and simpler, population-level regression models. Consequently, this paper provides an understanding of the type of explanatory claims these common methods can test.

Keywords: Explanation, statistical testing, population regression models, random coefficient models, mixture models

2010 AMS subject classification: 62A01; 62F03.†

* University of Rochester, Rochester NY, USA; peter_veazie@urmc.rochester.edu.

† Received on February 12th, 2019. Accepted on May 3rd, 2019. Published on June 30th, 2019. doi: 10.23755/rm.v36i1.463. ISSN: 1592-7415. eISSN: 2282-8214.

1. Introduction

Science provides explanations for facts, phenomena, and other explanations. In applied research that draws on theories from disciplines such as Economics, Psychology, Sociology, and Organizational Science, among others, this can require testing whether a proposed explanation explains a given fact, phenomenon, and other explanation in a specified population. For example, one might wish to test whether a proposed explanation based on Psychology's Regulatory Focus Theory [1, 2] explains physician risk tolerance in treatment choice (the phenomenon) among primary care physicians in the United States (the population). However, what is meant by the phrase *explains in a population*? Is it that the proposed explanation accounts for the behavior of every member of the population? This is a high bar: one member of the population for whom the explanation does not hold falsifies the claim. Is it that the proposed explanation accounts for the behavior of at least one member? This is equally extreme: only one member of a population for whom the explanation holds warrants the claim. The claim is ambiguous. Specific definitions are required if such claims are to be understood and tested.

This paper provides definitions and identifies methods for testing corresponding explanatory claims. These definitions and the identification of corresponding methods are new contributions that provide conceptual and methodological guidance for researchers who seek to test explanations in populations. The methods themselves, however, are in common use: random coefficient models and population-level regression models. Therefore, whereas a goal of this paper is to show which methods can be used to test specific explanatory claims, I do not present the implementation of the methods: there are many textbooks and articles that provide this information [e.g. 3, 4]. For simplicity of presentation, I only reference phenomena as the target of explanation rather than also facts and other explanations; however, any of these are applicable throughout.

2. Defining *explain*

Before providing the required definitions, I will clarify what I mean by *to explain* and by *an explanation*. For this paper, to explain something is to provide a way of understanding it through a conceptual structure that accounts, at least in part, for that which is being explained [5, Ch. 9]. The conceptual structure is the explanation. One might imagine there is a single explanation for any given phenomenon. However, for macro-level phenomena, such as organization and human behaviors, there may be multiple ways of understanding them. For example, a human behavioral phenomenon may have sociological explanations, psychological explanations, physiological explanations, and more. Any one of the explanations could be referred to as

an explanation, and no one of them referred to as exclusively *the explanation*. Moreover, an explanation need not be complete. There may be many causal factors or mechanisms that contribute to the phenomenon; however, an explanation might focus only on a subset.

An explanation can be intended to provide an understanding of a phenomenon as it is [6, Ch. 4], a *de re* explanation; or, it can be intended to provide an understanding that, nonetheless, contains explicitly presumed falsehoods [7, 8], a *de ficta* explanation. All terms of a *de re* explanation refer to presumed real objects, qualities, characteristics, and relationships. Designation as a *de re* explanation does not guarantee truth, nor does it imply the researcher believes it is true; indeed, if the researcher believed the explanation was in fact true, there is no need for further inquiry [9]. Moreover, it is common to expect even a well-established theory-based explanation to be incorrect in some unknown way. It is the ontological commitments (the presumption that explanatory terms intend to have real referents) of the explanation's terms that qualify it as a *de re* explanation. However, a *de ficta* explanation contains at least one identified term that is presumed to be false. These are often explanations that contain idealizations (e.g. the discrete energy levels in the Bohr model of the atom [10-12], and the rationality of the rational choice model in classic microeconomics [13, 14]) or analogies (e.g. the computer analogy or corporate analogy of information processing in cognitive science [15]). Given there need only be a single presumed false term to warrant designation as a *de ficta* explanation, the remaining terms have substantive ontological commitments. Such *de ficta* explanations are presumed to be partially true [7]. Although these definitions do not restrict explanations to those that are amenable to empirical investigation, this paper is written to provide guidance for empirical researchers. Consequently, the focus of the discussion herein is on scientific explanations that have empirical implications.

In the applied sciences, the goal of both *de re* and *de ficta* explanations is to guide interventions, actions, or policy. The pursuit and use of a *de re* explanation are based on the belief that understanding the world as it is provides assurance that consequent interventions, actions, and policies are more likely to work and generalize, and the causes for their failure are more likely to be identified. The *de ficta* explanation does not carry as great an assurance in these regards as it includes identified false claims. However, the *de ficta* explanation can be simpler, easier to develop and understand, and easier to apply. Both types of explanation are usefully employed.

Explanations are often assessed in terms of explanatory power. Explanatory power characterizes explanations in terms of explanatory virtues such as generality, coherence, accuracy, and predictive ability, among others [8, 16]. It has been qualitatively defined in terms of the scope of questions it

can address [16], and it has been the basis for formal probability-based measures [17-20]. However, for the purposes of applied science another aspect of power can be useful: effective power.

Applied researchers often focus on the ability to influence specific outcomes and therefore seek explanations to inform actions that can produce specific effects. For example, researchers may seek to reduce systolic blood pressure, decrease expected expenditures, or expand social networks rather than seek to account for variation. To achieve such goals, it can be important to assess a phenomenon's responsiveness to an explanation, its effective power. Effective power is different from accuracy and predictive power (the abilities to account for and predict phenomena and behavior). Consider an explanation of the relationship between behavior Y and explanatory factor X for two individuals w and v. Suppose the effect of the explanation on Y can be modeled as a simple linear function of X with a positive coefficient, in which variable X completely determines Y for individual w and only partially determines Y for individual v:

$$Y_w = \beta_w \cdot X_w$$

and

$$Y_v = \beta_v \cdot X_v + E_v.$$

The predictive power for w is greater than that for v; indeed, the predictive power for w is perfect, whereas it is only partial for v, due to the additional term E_v. However, if $\beta_w = \beta_v$, then variable X has the same relationship with behavior Y for both and thereby having the same effective power: a difference in X corresponds to the same difference in Y for both w and v. If $\beta_v > \beta_w$, then the explanation has greater effective power for v, even though it has greater predictive power for w. Effective power represents the responsiveness to the explanation whereas accuracy and predictive power represents the extent of Y accounted for by the explanation. As an analogy, consider a regression analysis, in the above example effective power is analogous to β and predictive power is analogous to the coefficient of determination (commonly termed R-square) or an out-of-sample prediction metric. Like Schupbach and Sprenger's [18] definition of explanatory power, effective power can be negative for a proposed explanation, if the response is counter to that implied by the explanation: for example, the case in which the β's in the preceding example were in fact negative, contrary to the explanatory implication of positive β's.

We can understand a population-level *de re* or *de ficta* explanatory claim as a reductive explanation: an explanation that applies to a population in virtue of an aggregation of the explanation's application to its members. This is kin to what Strevens terms an aggregative explanation [8]. For example, where I

may seek to explain physician risk tolerance in treatment choice among primary care physicians in the United States, the proposed explanation is regarding its members' relevant behaviors (the behaviors of individual physicians). So, regardless of the number of members in the population, which can be as few as one, our definition of the phrase *a potential explanation explains a given phenomenon in a population* represents an aggregation of an individual-level explanation across the members of the population.

As stated in the introduction, definitions that require explanation of either every member or only one member of a population are extreme. Appropriate definitions are likely somewhere in between. This paper focuses on two:

Definition 1. An explanation explains a phenomenon in a population if, and only if, it has positive effective power for most members of the population.

Definition 2. An explanation explains a phenomenon in a population if, and only if, its cumulative magnitudes of effective power among the members of the population for whom the explanation holds exceeds its cumulative magnitudes of effective power among the members of the population for whom the explanation does not hold.

These definitions are based on minimal criteria. In the first case, it would be difficult to support an explanatory claim regarding scope if the possible explanation only applied to a minority of population members. In the second case, it would be difficult to support an explanatory claim regarding cumulative power if the possible explanation was associated with less cumulative power than the counter-explanation in a population. However, this is arbitrary, and we need not take the minimal stance. We can generalize the definitions to vary with a definitional parameter q:

General Definition 1. An explanation explains a phenomenon in a population if, and only if, it has effective power for at least q percent of the members of the population.

General Definition 2. An explanation explains a phenomenon in a population if, and only if, its cumulative magnitudes of effective power among the members of the population for whom the explanation holds exceeds q times its cumulative magnitudes of effective power among the members of the population for whom the explanation does not hold.

The remaining sections focus on the minimal definitions, however the general testing method in Section 4.1 can be used to test these general definitions as well.

3. Defining Testable Implications

To test claims based on the preceding definitions, we required corresponding operational definitions in terms of testable implications:

Operational Definition 1. If an explanation explains a phenomenon in a population, then the implications of the explanation hold for most of the members of the population. And, under reasonable presumption (i.e. credible alternative explanations are accounted for), if the implications of the explanation hold for most of the members of the population, then an explanation explains a phenomenon in a population.

Operational Definition 2. If an explanation explains a phenomenon in a population, then the cumulative strength of the explanation's implications among the members of the population for whom the explanation holds exceeds the cumulative strength of the counter-implications among the members of the population for whom the explanation does not hold. And, under reasonable presumption (i.e. credible alternative explanations are accounted for), if the cumulative strength of the explanation's implications among the members of the population for whom the explanation holds exceeds the cumulative strength of the counter-implications among the members of the population for whom the explanation does not hold, then an explanation explains a phenomenon in a population.

The first conditional in each operational definition allows evidence against each consequent (the testable implications) to provide evidence against the explanatory claim. The second conditional allows evidence for each antecedent (the testable implications) to provide evidence for the explanatory claim. The first conditionals are typically derived from the explanation. The second conditionals draw more upon the weaker condition of presumption-based reasoning [21], which is grounded in current background knowledge and is thereby defeasible: future changes in scientific understanding can negate the conditional. A strong reasonable presumption for the second conditionals is achieved if there are no credible alternative explanations for the testable implications.

Regarding operational definition 1, we might say, for example, that a Regulatory-Focus-Theory-based explanation explains physician risk tolerance in treatment choice among primary care physicians in the United States if a higher promotion focus (a term in Regulatory Focus Theory [1, 22]) leads physicians to have higher risk tolerance (the explanation's implication) for more than half of the physicians, accounting for alternative explanations. Regarding operational definition 2, we might say that a Regulatory-Focus-Theory-based explanation explains physician risk tolerance in treatment choice among primary care physicians in the United States if the cumulative

magnitudes of effect of promotion focus on risk tolerance among physicians for whom a higher promotion focus leads the physician to have higher risk tolerance exceeds the cumulative magnitudes of effect of promotion focus on risk tolerance among physicians for whom a higher promotion focus leads the physician to have lower risk tolerance (or no relationship).

We can generalize the operational definitions, as we did with the original definitions, to vary with a definitional parameter *q*:

General Operational Definition 1. If an explanation explains a phenomenon in a population, then the implications of the explanation hold for *q* percent of the members of the population. And, under reasonable presumption (i.e. credible alternative explanations are accounted for), if the implications of the explanation hold for *q* percent of the members of the population, then an explanation explains a phenomenon in a population

General Operational Definition 2. If an explanation explains a phenomenon in a population, then the cumulative strength of the explanation's implications among the members of the population for whom the explanation holds exceeds *q* times the cumulative strength of the counter-implications among the members of the population for whom the explanation does not hold. And, under reasonable presumption (i.e. credible alternative explanations are accounted for), if the cumulative strength of the explanation's implications among the members of the population for whom the explanation holds exceeds *q* times the cumulative strength of the counter-implications among the members of the population for whom the explanation does not hold, then an explanation explains a phenomenon in a population.

To test claims based on the preceding definitions, we start by identifying the proposed explanation's implications. Specifically, we presume an explanation-implied relationships *g* between variables *Y* and *X* (as defined in the context of the phenomenon and explanation), with parameter θ:

$$y = g(x;\theta) \text{ , such that } \frac{\partial g(x;\theta)}{\partial x} \in \mathbb{D}_e \text{ , } \forall x \in \mathbb{R}_X . \quad (1)$$

This is to say that we have a proposed explanation *e* of a phenomenon that implies variables *X* and *Y* are related by some, perhaps unknown, function *g* such that for all values *x* in range $\mathbb{R}_X$ the derivative of *g* with respect to *x* (or the difference quotient if $\mathbb{R}_X$ is a discrete set) is in the set $\mathbb{D}_e$. Note that the implications can be more general: The $\frac{\partial g(x;\theta)}{\partial x}$ term can be a vector of derivatives across multiple *X* variables. And, the implications for any given derivative can be multi-part, having different ranges for the derivative across

different *x*-values. However, for ease of presentation this paper focuses on single-part implications.

A simple example is *g* specified as a linear relationship, $y = \alpha + \beta \cdot x$, such that the proposed explanation *e* implies $dy/dx > 0$, i.e. $\mathbb{D}_e = (0,\infty)$, for all positive values of *X*, i.e. $\mathbb{R}_X = (0,\infty)$. Applying this equation to all members of Ω, we can say that if β is positive for most members of a population Ω, then *e* explains by definition 1. If the sum of the magnitude of β's across all members of Ω for whom $\beta>0$ exceeds the sum of the magnitudes of β's across all members for whom $\beta\leq 0$, then *e* explains by definition 2.

To formalize the concept of *explain*, consider the following variable Δ defined for $w \in \Omega$ and $x \in \mathbb{R}_X$:

$$\Delta(w,x) = h\left(\frac{\partial g(x;\Theta(w))}{\partial x}\right). \tag{2}$$

The function *h* provides the relevant interpretation for *explain*. The two functions considered in this paper for *h* provide interpretations for *explain* as the scope of the explanation (definition 1 above) and as the power of the explanation (definition 2 above). These are detailed below.

We can use two functions to separate the Δ's into groups. The first picks out Δ for the explanation-implied range of values for $\partial g/\partial x$, and the second picks out Δ for the range of values outside of the explanation-implied range—the counter-explanation range:

$$\Delta^{+}(w,x) = \begin{cases} \Delta(w,x) & \text{if } \dfrac{\partial g(x;\Theta(w))}{\partial x} \in \mathbb{D}_e \\ 0 & \text{Otherwise} \end{cases} \tag{3}$$

and

$$\Delta^{-}(w,x) = \begin{cases} \Delta(w,x) & \text{if } \dfrac{\partial g(x;\Theta(w))}{\partial x} \notin \mathbb{D}_e \\ 0 & \text{Otherwise} \end{cases}. \tag{4}$$

The sum of the magnitudes of Δ^+ across population Ω at value *x* reflects the extent of the proposed explanation's implications in the population at *x* (the interpretation depending on *h*). The sum of the magnitudes of Δ^- across population Ω at value *x* reflects the extent of counter-explanation implications in the population at *x*.

For both specifications of *h* discussed below, a useful formalization of *explain* is to say that the proposed explanation explains a phenomenon in a population if the accumulated magnitudes of Δ is larger in the explanation-

implied region than in the counter-explanation region for all points in a specified set *B* of *x*-values. For arbitrary value *x* in *B*, this implies for both definitions 1 and 2 that

$$\sum_{w\in\{w:X(w)=x\}}\left(\left|\Delta^{+}(w,x)\right|\right) > \sum_{w\in\{w:X(w)=x\}}\left(\left|\Delta^{-}(w,x)\right|\right). \tag{5}$$

For the generalized definitions this is

$$\sum_{w\in\{w:X(w)=x\}}\left(\left|\Delta^{+}(w,x)\right|\right) > q^{\circ}\cdot\sum_{w\in\{w:X(w)=x\}}\left(\left|\Delta^{-}(w,x)\right|\right), \tag{6}$$

where $q^{\circ} = q/(100 - q)$ for generalized definition 1, and $q^{\circ} = q$ for generalized definition 2.

Denoting the statement *e explains p in* Ω *on set B* as $E(e, p, \Omega, B)$, the corresponding claims are $E(e, p, \Omega, B) = True$ and $E(e, p, \Omega, B) = False$. The claim that the proposed explanation holds (i.e. $E(e, p, \Omega, B) = True$) is asserted if for all points *x* in the set *B* the proposed explanation's implication exceeds that for the counter-explanation implication. The claim that the proposed explanation does not hold (i.e. $E(e, p, \Omega, B) = False$) is asserted if there exists at least one point in *B* for which the counter-explanation implication exceeds the proposed explanation's implication.

It is useful to take *B* to be one of two sets: either a singleton $\{x\}$ or the phenomenologically-relevant range $\mathbb{R}_X$. Claims $E(e, p, \Omega, \mathbb{R}_X)$ are what we may consider when testing whether a proposed explanation explains, whereas point-wise claims $E(e, p, \Omega, \{x\})$ are useful in understanding where in the range of *x*-values the claims $E(e, p, \Omega, \mathbb{R}_X)$ fail, if indeed they fail, or at which points of *X* is the underlying proposed explanation is either least or most powerful. There are occasions, however, when $E(e, p, \Omega, \mathbb{R}_X)$ is too strict: do we really want to say a proposed explanation does not explain in a population because it doesn't hold at a single point *x*? For example, suppose economic demand follows the predicted relationship with price at all prices except at $1, do we say the price-demand theory does not hold in the population because of this singular exception? Perhaps we should account for how important it is that the explanation hold at $1, or account for how many people face a price of $1 for the good being considered. We can address these concerns by taking a weighted average of *x*-specific effects across the range of *x*-values in $\mathbb{R}_X$ using a probability distribution for *X* conditional on $x \in \mathbb{R}_X$. Denoting this general explanatory claim as $E(e, p, \Omega)$, it requires the weighted sum across all *x*-values being considered and thereby can balance non-explanatory points of $\mathbb{R}_X$ with other strongly explanatory points. Its interpretation depends on

the definition of the probability for X [23]. For example, it can be helpful to consider claims regarding $E(e, p, \Omega)$ in terms of random variables defined on population Ω, with equal probabilities assigned to each member of Ω. Using Ω as its domain, the variable X provides the value x that each member is facing. The probability distribution of X therefore represents the actual normalized frequency of X in the population, and consequently $E(e, p, \Omega)$ is based on the corresponding weighted average across this distribution.

Figure 1 presents an example in which the explanation implies negative derivatives of g with respect to x, i.e. $\mathbb{D}_e = (-\infty, 0)$ for all values of x in $\mathbb{R}_X$, but for which the actual g is as shown. It is clear, regarding the point-wise explanations, that the claim $E(e, p, \Omega, \{x\}) = True$ holds true only for x less than x^*, but $E(e, p, \Omega, \{x\}) = False$ for all x greater than x^*. Consequently, due to the existing values of X for which the explanatory implications do not hold (i.e. for $x > x^*$), the overall claim is therefore $E(e, p, \Omega, \mathbb{R}_X) = False$. On the other hand, for $f(x)$ denoting the density of X based on $P(x \mid x \in \mathbb{R}_X)$, the general claim weighted by this probability is $E(e, p, \Omega) = True$ as there is little probability associated with x-values in the contra-explanatory range of derivatives.

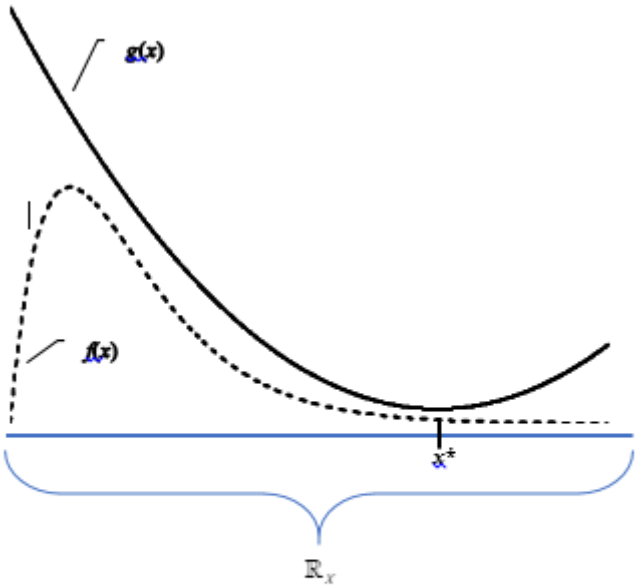

Figure 1. Example of $E(e, p, \Omega, \mathbb{R}_X) = No$ because dg/dx > 0 for some x in $\mathbb{R}_X$ (i.e. for $x > x^*$), and $E(e, p, \Omega) = Yes$ because the density f weights dg/dx heavily in the explanation-consistent region (i.e. where dg/dx < 0) and trivially in the non-explanation consistent region (i.e. where dg/dx > 0).

As mentioned above, two specifications for *h* are considered here. The first, for definition 1, specifies *h* as a constant function with value 1:

$$\Delta(w, x) = 1, \text{ for all } w \text{ and } x. \tag{7}$$

This leads to

$$\Delta^{+}(w, x) = \begin{cases} 1 \text{ if } \dfrac{\partial g(x; \Theta(w))}{\partial x} \in \mathbb{D}_e \\ 0 \quad \text{Otherwise} \end{cases} \tag{8}$$

and

$$\Delta^{-}(w, x) = \begin{cases} 1 \text{ if } \dfrac{\partial g(x; \Theta(w))}{\partial x} \notin \mathbb{D}_e \\ 0 \quad \text{Otherwise} \end{cases}. \tag{9}$$

By this definition, the sum of the absolute values of Δ^{+} is the number of people whose *X* and *Y* relationship follows the proposed explanation's prediction at specified *x*-values. The sum of absolute value of Δ^{-} is the number of people whose *X* and *Y* relationship do not follow the proposed explanation's prediction. A proposed explanation explains at *x*, by equation 5, if more people in the population follow the prediction than do not when $X = x$.

The second specification, which is used for definition 2, is to define *h* as the identity function, and therefore Δ is

$$\Delta(w, x) = \frac{\partial g(x; \Theta(w))}{\partial x}. \tag{10}$$

This leads to

$$\Delta^{+}(w, x) = \begin{cases} \dfrac{\partial g(x; \Theta(w))}{\partial x} \text{ if } \dfrac{\partial g(x; \Theta(w))}{\partial x} \in D_e \\ 0 \quad \text{Otherwise} \end{cases} \tag{11}$$

and

$$\Delta^{-}(w, x) = \begin{cases} \dfrac{\partial g(x; \Theta(w))}{\partial x} \text{ if } \dfrac{\partial g(x; \Theta(w))}{\partial x} \notin D_e \\ 0 \quad \text{Otherwise} \end{cases}. \tag{12}$$

The corresponding definition for explain compares the accumulated magnitudes of Δ between the explanation-implied region and the counter-

explanation region, which reflects the cumulative effective power of the explanation in the population.

The difference between these two corresponding specifications for *h* is that the first claim, E_1, focuses on the scope (the number or proportion of the population consistent with the explanation), whereas the second claim, E_2, focuses on the cumulative power of the explanation. It is possible for an explanation to apply to a minority of people in the population, but it does so with greater strength in the magnitude of Δ among this minority than is the magnitude of Δ for the majority, who are not in the implied region. In this case the explanation would be considered as explaining in terms of E_2, which uses the identity function for *h*, but not in terms of E_1, which uses the constant function for *h*. On the other hand, in the case where a majority has only a tiny magnitude of Δ in the implied region but a minority has a large magnitude of Δ in the non-implied region, the explanation would be considered as explaining in terms of E_1 but not in terms of E_2. This is analogous to considering the importance of whether a treatment has a larger total positive effect among those that benefit relative to the total negative affect among those who do not benefit (E_2), or whether the treatment simply positively affects a greater proportion of people regardless of how small the effect (E_1). Which definition is appropriate depends on the research goal.

These definitions are population-specific. Consequently, it is possible for a proposed explanation to explain in one population but not another. Moreover, it is possible to not explain in a population but to explain in one of its subpopulations, and vice versa. Consider a population Ω made up of two subpopulations Ω_1 and Ω_2: it is possible for $E(e, p, \Omega, \mathbb{R}_X) = False$, and yet $E(e, p, \Omega_1, \mathbb{R}_X) = True$. This is often the advantage of doing subgroup analysis, to determine if a proposed explanation holds better in one group than another. Indeed, the primary scientific aim of a study may be to identify for which population the proposed explanation holds.

4. Testing explanations

4.1 General tests using random coefficient models

How do we empirically test a hypothesis of the form $E(e, p, \Omega, \mathbb{R}_X) = True$ or $E(e, p, \Omega, \mathbb{R}_X) = False$? A general approach is conceptually straightforward, albeit empirically challenging. This approach is based on the idea that if we can estimate the distribution of Δ, we can estimate the conditions for $E(e, p, \Omega, \mathbb{R}_X) = True$ and $E(e, p, \Omega, \mathbb{R}_X) = False$. To estimate

the distribution of Δ, assuming our data generating process can support it, we can use a random coefficient model [3].

Suppose we define random variables (or random vectors) *Y*, *X*, Θ, and $\mathcal{E}$ on the population Ω, representing a population model such that

$$Y(w) = g(X(w); \Theta(w)) + \mathcal{E}(w), \text{ for } w \in \Omega. \tag{13}$$

If we have a data generating process with *N* observations, $i \in \{1, \ldots N\}$, we can consider the mixture model for the regression of *Y* on *X*:

$$E(Y_i \mid x_i) = \int E(Y_i \mid x_i, \theta_i) \cdot dF(\theta_i \mid x_i). \tag{14}$$

Substituting equation 13 for Y_i on the right-hand side of equation 14, yields

$$E(Y_i \mid x_i) = \int g(x_i, \theta_i) \cdot dF(\theta_i \mid x_i) + \int E(\mathcal{E}_i \mid x_i, \theta_i) \cdot dF(\theta_i \mid x_i), \tag{15}$$

which is the expected value of g plus the expected value of $\mathcal{E}$, each conditioned on $X = x$:

$$E(Y_i \mid x_i) = \int g(x_i, \theta_i) \cdot dF(\theta_i \mid x_i) + E(\mathcal{E}_i \mid x_i). \tag{16}$$

Under the assumption that the expected value of the error terms is 0 for all values of *X*, the regression is

$$E(Y_i \mid x_i) = \int g(x_i, \theta_i) \cdot dF(\theta_i \mid x_i). \tag{17}$$

The derivative of *g* and the estimated distribution for *F* can be used to obtain a distribution for Δ and thereby estimate the conditions for the explanation to hold. Notice, however, from equation 17 the function *g* must be the expected value of *Y* conditional on values of *X* and Θ, i.e. equation 14. Consequently, if a statistically adequate model [24] for $E(Y_i \mid x_i, \theta_i)$ can be empirically determined, an explicit a priori specification for *g* is not required, only hypotheses regarding implications (e.g. derivatives or difference quotients) are required a priori.

Estimation can be achieved using a mixture model, or random parameters model, if the study design and context allow for estimation of such a model. It is best to use a non-parametric estimator for $F(\theta \mid x)$ since results in this case are likely to be very sensitive to the distribution (we are integrating under different regions of the distribution, rather than merely estimating parameters of the distribution). For example, we may consider using Fox et al's non-parametric estimator for the distribution of random effects [25, 26].

Suppose we can assume the error term is independent of *X* and that we have a relationship such that

$$g(x_i, \theta_i) = e^{x_i \cdot \theta_i}, \tag{18}$$

which has the derivative

$$\frac{dg(x_i, \theta_i)}{dx} = \theta_i \cdot e^{x_i \cdot \theta_i}. \tag{19}$$

The expected value of *Y* conditional on *X* is

$$E(Y_i \mid x_i) = \int e^{(x_i \cdot \theta_i)} \cdot dF(\theta_i \mid x_i). \tag{20}$$

With an estimator for *F*, denoted as $\hat{F}$, we can estimate, using numeric integration, the population proportion of those whose derivative falls in the explanation-implied range for any *x*,

$$\hat{p}(x) = \int I(\theta \cdot e^{\theta \cdot x} > 0) \cdot d\hat{F}(\theta \mid x), \tag{21}$$

in which $I(\cdot)$ is an indicator function returning 1 if its argument is true, 0 otherwise. Equation 21 can be used to test *E1*.

For the general *E1*, based on the population distribution for *X* and representative sampling, we would average estimates from equation 21 for each observation in the data to obtain

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} \hat{p}(x_i). \tag{22}$$

In this case, because $e^{x \cdot \theta}$ is always positive, the sign of the derivative is determined by the sign of θ. Therefore, we can estimate $\hat{p}$ based solely on an indicator of $\theta > 0$:

$$\hat{p}(x) = \int I(\theta > 0) \cdot d\hat{F}(\theta \mid x). \tag{23}$$

If we can assume the distribution *F* is independent of *x*, i.e. $F(\theta \mid x) = F(\theta)$ for all *x*, then $\hat{p}$ is not a function of *x*, and $\hat{p}(x)$ is the same for all *x*; therefore

$$\hat{p} = \int I(\theta > 0) \cdot d\hat{F}(\theta) = 1 - \hat{F}(0). \tag{24}$$

In this case we can base our test on $1 - \hat{F}(0)$. Using a bootstrap distribution for $\hat{p}$ (for either equation 23 or equation 24), if a legitimate bootstrap method applies [27], we can test whether *E1* is the case using the p-value $P(p \geq \hat{p} \mid p = 0.5)$ if $\hat{p} \geq 0.5$, and p-value $P(p \leq \hat{p} \mid p = 0.5)$ if $\hat{p} \leq 0.5$ [28].

For testing *E2* at specific *x*-values we calculate

$$c(x) = \int \left[\left(I(\theta > 0) \cdot \left| \theta \cdot e^{\theta \cdot x} \right| \right) - \left(I(\theta \leq 0) \cdot \left| \theta \cdot e^{\theta \cdot x} \right| \right) \right] \cdot d\hat{F}(\theta). \qquad (25)$$

For testing the general E_2 we average $c(x)$ across the data. Again, we can use the bootstrap distribution for F to obtain p-values $P(c \geq \hat{c} \mid c = 0)$ or $P(c \leq \hat{c} \mid c = 0)$.

4.2 Testing E_2 using population-level regression models

The preceding method, which uses random coefficient models and numeric integration, is complicated—particularly for E_2, which represents definition 2. We can greatly simplify our method for testing E_2, if the explanation's implications are regarding positive vs non-positive (or negative vs non-negative) derivatives. In this case, with an additional statistical assumption, we can use population-level regression models to test the explanation. The argument is as follows: As above, we say that *e* explains phenomenon *p* at *x* if inequality 5 holds. Under the definition for E_2, in the case of $\mathbb{D}_e$ being either positive, negative, non-positive or non-negative, the absolute values can be moved outside of the summations,

$$\left| \sum_{w \in (w:X(w)=x\}} \left(\Delta^{+}(w,x) \right) \right| > \left| \sum_{w \in (w:X(w)=x\}} \left(\Delta^{-}(w,x) \right) \right|. \qquad (26)$$

Consider $\mathbb{D}_e = (0,\infty)$, i.e. the explanation implies positive derivatives. In this case, for the left-hand side of inequality 26 the summation of the Δ^+ across the population with $X = x$ is the same as the summation of the product of each Δ-value and its frequency for Δ-values greater than 0:

$$\sum_{w \in (w:X(w)=x\}} \left(\Delta^{+}(w,x) \right) = \sum_{\Delta > 0} \Delta \cdot Freq(\Delta \mid x). \qquad (27)$$

Similarly, regarding Δ^-,

$$\sum_{w \in (w:X(w)=x\}} \left(\Delta^{-}(w,x) \right) = \sum_{\Delta \leq 0} \Delta \cdot Freq(\Delta \mid x). \qquad (28)$$

Therefore, to determine E_2 we can consider whether

$$\left| \sum_{\Delta > 0} \Delta \cdot Freq(\Delta \mid x) \right| > \left| \sum_{\Delta \leq 0} \Delta \cdot Freq(\Delta \mid x) \right|. \qquad (29)$$

However, the inequality remains true if both sides are multiplied by the same positive constant. So, if we multiply by $1/N_x$, denoting the inverse of the population size with value $X = x$, then

$$\left|\sum_{\Delta>0}\Delta\cdot\frac{Freq(\Delta\mid x)}{N_x}\right|>\left|\sum_{\Delta\le 0}\Delta\cdot\frac{Freq(\Delta\mid x)}{N_x}\right|, \tag{30}$$

which is

$$\left|\sum_{\Delta>0}\Delta\cdot f(\Delta\mid x)\right|>\left|\sum_{\Delta\le 0}\Delta\cdot f(\Delta\mid x)\right| \tag{31}$$

for f denoting a probability mass function (however, the above logic and derivation also applies to Δ as a continuous variable in which f is a density, and the summation is replaced with an integral).

Multiplying the left side of inequality 31 by 1 written as

$$\frac{P(\Delta>0\mid x)}{P(\Delta>0\mid x)},$$

and multiplying the right side by 1 written as

$$\frac{P(\Delta\le 0\mid x)}{P(\Delta\le 0\mid x)},$$

yields

$$\left|\sum_{\Delta>0}\Delta\cdot f(\Delta\mid x)\cdot\frac{P(\Delta>0\mid x)}{P(\Delta>0\mid x)}\right|>\left|\sum_{\Delta\le 0}\Delta\cdot f(\Delta\mid x)\cdot\frac{P(\Delta\le 0\mid x)}{P(\Delta\le 0\mid x)}\right|. \tag{32}$$

Because on the left side of this inequality

$$\frac{f(\Delta\mid x)}{P(\Delta>0\mid x)}=f(\Delta\mid\Delta>0,x), \tag{33}$$

and on the right side of the inequality

$$\frac{f(\Delta\mid x)}{P(\Delta\le 0\mid x)}=f(\Delta\mid\Delta\le 0,x), \tag{34}$$

the inequality can be rewritten as

$$\left|\sum_{\Delta>0}\Delta\cdot f(\Delta\mid\Delta>0,x)\cdot P(\Delta>0\mid x)\right|>\left|\sum_{\Delta\le 0}\Delta\cdot f(\Delta\mid\Delta\le 0,x)\cdot P(\Delta\le 0\mid x)\right|. \tag{35}$$

Note that on the left side of inequality 35

$$\sum_{\Delta>0}\Delta\cdot f(\Delta\mid\Delta>0,x)=E(\Delta\mid\Delta>0,x), \tag{36}$$

and on the right side of the inequality

$$\sum_{\Delta \le 0} \Delta \cdot f(\Delta \mid \Delta \le 0, x) = E(\Delta \mid \Delta \le 0, x). \quad (37)$$

By substitution into equation 35, this yields

$$\left|E(\Delta \mid \Delta > 0, x) \cdot P(\Delta > 0 \mid x)\right| > \left|E(\Delta \mid \Delta \le 0, x) \cdot P(\Delta \le 0 \mid x)\right|. \quad (38)$$

Subtracting the right side of inequality 38 from both sides yields

$$\underbrace{\left|E(\Delta \mid \Delta > 0, x) \cdot P(\Delta > 0 \mid x)\right|}_{\text{Part A}} - \underbrace{\left|E(\Delta \mid \Delta \le 0, x) \cdot P(\Delta \le 0 \mid x)\right|}_{\text{Part B}} > 0. \quad (39)$$

Since Part A of inequality 39 is the absolute value of a positive number (note we are conditioning on $\Delta > 0$), the absolute value function can be dropped. Similarly, since Part B is the absolute value of a non-positive number (note we are conditioning on $\Delta \le 0$), its subtraction from *A* is just the addition of the non-positive number. The absolute value operation can be dropped as well, if we add the components rather than subtract them. This yields

$$E(\Delta \mid \Delta > 0, x) \cdot P(\Delta > 0 \mid x) + E(\Delta \mid \Delta \le 0, x) \cdot P(\Delta \le 0 \mid x) > 0. \quad (40)$$

However, the left-hand side of this inequality is the expected value of Δ conditional on *x*. Therefore, explanation E_2 implies that

$$E(\Delta \mid x) > 0 \quad \forall x \in \mathbb{R}_X. \quad (41)$$

Since $\Delta = \partial g / \partial x$ and derivatives are linear operators (and assuming we can interchange the derivative and integral operations), we have

$$E(\Delta \mid x) = E\left(\frac{\partial g(x)}{\partial x}\middle| x\right) = \frac{dE(g(x) \mid x)}{dx}, \quad (42)$$

and therefore, the implication of the explanation we seek to test is the direction of the derivative of the expected value of *g*:

$$\frac{dE(g(x) \mid x)}{dx} > 0 \quad \forall x \in \mathbb{R}_X. \quad (43)$$

Unfortunately, whereas we are likely able to empirically evaluate E($Y \mid x$) in a regression analysis, we are not likely able to directly evaluate E($g \mid x$). This is okay, if we can we use E($Y \mid x$) to evaluate E($g \mid x$). When can we do this? The requirements are identified by taking the derivative of equation 16 with respect to *x*:

$$\frac{dE(Y \mid x)}{dx} = \int \frac{\partial g(x;\theta)}{\partial x} \cdot f(\theta \mid x) \cdot d\theta + \underbrace{\int g(x;\theta) \cdot \frac{\partial f(\theta \mid x)}{\partial x} \cdot d\theta}_{\text{Part A}} + \underbrace{\frac{\partial E(\mathcal{E} \mid x)}{\partial x}}_{\text{Part B}}. \quad (44)$$

If the distribution of parameter Θ is independent of X (which, in econometrics, is often considered as there is no selection on the gains [29]), then df/dx = 0 and consequently Part A of equation 44 is zero. If the error is mean independent of X, then Part B is zero (which in econometrics, is often considered as there is no selection on the outcome [29]). Under these conditions we have

$$\frac{dE(Y \mid x)}{dx} = \int \frac{\partial g(x;\theta)}{\partial x} \cdot f(\theta) \cdot d\theta \,. \tag{45}$$

But, the right-hand side of equation 45 is the E($\Delta \mid x$), which is what we seek to evaluate for our test. Consequently, our empirical claim regarding $E(e, p, \Omega, \mathbb{R}_X) = True$ for E_2 is

$$\frac{dE(Y \mid x)}{dx} \in \mathbb{D}_e, \quad \forall x \in \mathbb{R}_X \,. \tag{46}$$

Given the independence assumptions required for parts A and B to equal 0 in equation 44, we can test our proposed explanation E_2 by evaluating the derivative of a population-level regression function (the left-hand side of equation 45). If an empirically identified statistically adequate regression function can be used, an explicit functional form for g need not be specified a priori.

5. Conclusion

Knowing how to test a proposed explanation in a population requires having a definition for what is meant by explaining in a population. In this paper I gave definitions in terms of the scope of an explanation and in terms of the power of an explanation. I provided a general method for testing proposed explanations using random parameters models, and I showed when population-level regression models can be used to test proposed explanations in terms of effective power.

Although the tests were presented in terms of the minimal definitions, the tests can be extended to generalized definitions as described above. Using the random parameters method, we can define our explanations in terms of the explanation-implied region being a multiple of that for the non-implied region. For example, the proposed explanation explains if it applies to at least 90 percent of the population (rather than at least 50 percent as used in the minimal definitions).

I focused on defining and testing proposed explanations; however, in practice the requirements for such a test to provide evidence must be kept in mind. Specifically, a proposed explanation's testable empirical implications need to be specified such that alternative potential explanations for empirical

implications are accounted for or ruled out, typically by statistical or experimental control. The extent of evidence provided by the test depends on the confidence we have that alternative explanations for empirical findings are indeed ruled out: the less confident we are, the less evidence is provided by the test. This concern is addressed by calibrating our interpretation accordingly.

This paper addressed defining and testing explanations in populations. However, it should be noted that the general definition can be the basis for addressing estimation goals as well as testing goals. Using the random coefficients method the proportion of a population that conforms to the explanation's implications or the effective power can be estimated along with corresponding bootstrapped confidence intervals.

References

[1] E.T. Higgins, Promotion and prevention: Regulatory focus as a motivational principle, in: M.P. Zanna (Ed.) Adv Exp Soc Psychol, Academic Press, New York, 1998, pp. 1-46.

[2] P.J. Veazie, S. McIntosh, B. Chapman, J.G. Dolan, Regulatory focus affects physician risk tolerance, Health Psychology Research, 2 (2014) 85-88.

[3] E. Demidenko, Mixed models : theory and applications, Wiley-Interscience, Hoboken, N.J., 2004.

[4] R.B. Darlington, A.F. Hayes, Regression analysis and linear models : concepts, applications, and implementation, Guilford Press, New York, 2017.

[5] M. Bunge, Philosophy of science: From Explanation to Justification, Rev. ed., Transaction Publishers, New Brunswick, N.J., 1998.

[6] T. Sider, Writing the book of the world, Clarenson Press ; Oxford University Press, Oxford, New York, 2011.

[7] N.C.A. da Costa, S. French, Science and partial truth : a unitary approach to models and scientific reasoning, Oxford University Press, Oxford ; New York, 2003.

[8] M. Strevens, Depth : an account of scientific explanation, Harvard University Press, Cambridge, Mass., 2008.

[9] P.J. Veazie, Understanding Scientific Inquiry, Science and Philosophy, 6 (2018) 3-14.

[10] N. Bohr, On the Constitution of Atoms and Molecules, Philos Mag, 26 (1913) 857-875.

[11] N. Bohr, On the Constitution of Atoms and Molecules, Philos Mag, 26 (1913) 476-502.

[12] N. Bohr, On the Constitution of Atoms and Molecules, Philos Mag, 26 (1913) 1-25.

[13] S. DellaVigna, Psychology and Economics: Evidence from the Field, J. Econ. Lit., 47 (2009) 315-372.

[14] M. Rabin, A perspective on psychology and economics, Eur Econ Rev, 46 (2002) 657-685.

[15] E.F. Loftus, J.W. Schooler, Information-Processing Conceptualizations of Human Cognition: Past, present, and future, in: G.D. Ruben (Ed.) Informatoin and Behavior, Transaction Books, New Brunswick, NJ, 1985, pp. 225-250.

[16] P. Ylikoski, J. Kuorikoski, Dissecting explanatory power, Philos Stud, 148 (2010) 201-219.

[17] M.P. Cohen, On Three Measures of Explanatory Power with Axiomatic Representations, Brit J Philos Sci, 67 (2016) 1077-1089.

[18] J.N. Schupbach, J. Sprenger, The Logic of Explanatory Power, Philosophy of Science, 78 (2011) 105-127.

[19] J.N. Schupbach, Comparing Probabilistic Measures of Explanatory Power, Philosophy of Science, 78 (2011) 813-829.

[20] V. Crupi, K. Tentori, A Second Look at the Logic of Explanatory Power (with Two Novel Representation Theorems), Philosophy of Science, 79 (2012) 365-385.

[21] J.B. Freeman, Acceptable premises : an epistemic approach to an informal logic problem, Cambridge University Press, Cambridge, UK ; New York, 2005.

[22] E.T. Higgins, Beyond pleasure and pain, Am. Psychol., 52 (1997) 1280-1300.

[23] P. Veazie, What makes variables random : probability for the applied researcher, CRC Press, Taylor & Francis Group, Boca Raton, 2017.

[24] A. Spanos, Revisiting Haavelmo's structural econometrics: bridging the gap between theory and data, Journal of Economic Methodology, 22 (2015) 171-196.

[25] J.T. Fox, K.I. Kim, C.Y. Yang, A simple nonparametric approach to estimating the distribution of random coefficients in structural models, Journal of Econometrics, 195 (2016) 236-254.

[26] J.T. Fox, K.I. Kim, S.P. Ryan, P. Bajari, A simple estimator for the distribution of random coefficients, Quant Econ, 2 (2011) 381-418.

[27] G. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, New York, 1993.

[28] P.J. Veazie, Understanding Statistical Testing, Sage Open, 5 (2015).

[29] J.J. Heckman, E. Vytlacil, Econometric Evaluation of Social Programs, Part 1: Causal models, structural models and econometric policy evaluation, in: J. Heckman, E. Leamer (Eds.) Handbook of Econometrics, Elsevier, Amsterdam, 2007, pp. 4779-4874.

Ratio Mathematica Volume 36, 2019, pp. 27-42

Solution of two-point fuzzy boundary value problems by fuzzy neural networks

Mazin Hashim Suhhiem*

Basim Nasih Abood+

Mohammed Hadi Lafta++

Abstract

In this work, we have introduced a modified method for solving second-order fuzzy differential equations. This method based on the fully fuzzy neural network to find the numerical solution of the two-point fuzzy boundary value problems for the ordinary differential equations. The fuzzy trial solution of the two-point fuzzy boundary value problems is written based on the concepts of the fully fuzzy feed-forward neural networks which containing fuzzy adjustable parameters. In comparison with other numerical methods, the proposed method provides numerical solutions with high accuracy.

Keywords: Two-point fuzzy boundary value problem; fully fuzzy neural network; fuzzy trial solution; minimized error function; hyperbolic tangent activation function.

* Department of Statistics, University of Sumer, Alrifaee, Iraq; mazin.suhhiem@yahoo.com
+ Department of Mathematics, University of Wasit, Alkut, Iraq; basimabood@yahoo.com
++ Department of Statistics, University of Sumer, Alrifaee, Iraq; mohamnedhadi@yahoo.com
Received on March 5th, 2019. Accepted on April 29rd, 2019. Published on June 30th, 2019.
doi:10.23755/rm.v36i1.455. ISSN: 1592-7415. eISSN: 2282-8214.

1. Introduction

Many methods have been developed so far for solving fuzzy differential equations (FDEs) since it is utilized widely for the purpose of modelling problems in science and engineering. Most of the practical problems require the solution of the FDE which satisfies fuzzy initial conditions or fuzzy boundary conditions, therefore, the FDE must be solved. Many FDE could not be solved exactly, thus considering their approximate solutions is becoming more important.

The theory of FDE was first formulated by Kaleva and Seikkala. Kaleva was formulated FDE in terms of the Hukuhara derivative (H-derivative). Buckley and Feuring have given a very general formulation of a first order fuzzy initial value problem. They first find the crisp solution, make it fuzzy and then check if it satisfies the FDE.

In 1990 researchers began using the artificial neural network (ANN) for solving ordinary differential equation (ODE) and partial differential equation (PDE) such as: Lee and Kang in [1]; Meade and Fernandez in [2,3]; Lagaris and Likas in [4]; Liu and Jammes in [5]; Tawfiq in [6]; Malek and Shekari in [7]; Pattanaik and Mishra in [8]; Baymani and Kerayechian in [9]; and other researchers.

In 2010 researchers began using ANN for solving a fuzzy differential equation such as: Effati and Pakdaman in [10]; Mosleh and Otadi in [11]; Ezadi and Parandin in [12].

In 2012 researchers began using partially (non-fully) fuzzy artificial neural network(FANN) for solving a fuzzy differential equation such as Mosleh and Otadi in [13,14,15]. In (2016) Suhhiem [16] developed and used partially FANN for solving fuzzy and non-fuzzy differential equations.

In this work, we have used fully feed forward fuzzy neural network to find the numerical solution of the two-point fuzzy boundary value problems for the ordinary differential equations. The fuzzy trial solution of the fuzzy boundary value problem is written as a sum of two parts. The first part satisfies the fuzzy boundary condition, it contains no fuzzy adjustable parameters. The second part involves fully fuzzy feed-forward neural networks which containing fuzzy adjustable parameters.

2 Basic definitions

In this section, the basic notations which are used in fuzzy calculus are introduced

Definition(1),[16]: The r - level (or r - cut) set of a fuzzy set $\widetilde{A}$ labeled by A_r is the crisp set of all x in X (universal set) such that : $\mu_{\widetilde{A}}(x) \geq r$; i. e.

$$A_r = \{x \in X : \mu_{\widetilde{A}}(x) \geq r , r \in [0,1] \} . \quad (1)$$

Definition(2), Fuzzy Number[16]: A fuzzy number $\tilde{u}$ is completely determined by an ordered pair of functions $\left(\underline{u}(r) , \overline{u}(r)\right)$, $0 \leq r \leq 1$, which satisfy the following requirements:

1) $\underline{u}(r)$ is a bounded left continuous and non-decreasing function on $[0,1]$.

2) $\overline{u}(r)$ is a bounded left continuous and non-increasing function on $[0,1]$.

3) $\underline{u}(r) \leq \overline{u}(r)$, $0 \leq r \leq 1$. (2)

The crisp number (a) is simply represented by:

$\underline{u}(r) = \overline{u}(r) = a$, $0 \leq r \leq 1$.

The set of all the fuzzy numbers is denoted by E^1.

Remark(1),[10]: For arbitrary $\tilde{u} = \left(\underline{u} , \overline{u}\right)$, $\tilde{v} = \left(\underline{v} , \overline{v}\right)$ and $K \in R$, the addition and multiplication by K For all $r \in [0,1]$ can be defined as:

1) $\underline{(u + v)}\ (r) = \underline{u}(r) + \underline{v}(r)$.

2) $\overline{(u + v)}\ (r) = \overline{u}(r) + \overline{v}(r)$.

3) $\underline{(Ku)}\ (r) = K\,\underline{u}(r)$, $\overline{(Ku)}\ (r) = K\,\overline{u}(r)$, if $K \geq 0$.

4) $\underline{(Ku)}\ (r) = K\,\overline{u}(r)$, $\overline{(Ku)}\ (r) = K\,\underline{u}(r)$, if $K < 0$. (3)

Remark(2),[16]: The distance between two arbitrary fuzzy numbers $\tilde{u} = \left(\underline{u} , \overline{u}\right)$ and $\tilde{v} = \left(\underline{v} , \overline{v}\right)$ is given as:

$$D(\tilde{u} , \tilde{v}) = \left[\int_0^1 (\underline{u}(r) - \underline{v}(r))^2 dr + \int_0^1 (\overline{u}(r) - \overline{v}(r))^2 dr\right]^{\frac{1}{2}} \quad (4)$$

Remark(3),[16]: (E^1, D) is a complete metric space.

Definition (3) , Fuzzy Function [16] : The function F: R $\longrightarrow E^1$ is called a fuzzy function.

We call every function defined in set $\widetilde{A} \subseteq E^1$ to $\widetilde{B} \subseteq E^1$ a fuzzy function.

Definition(4),[10]: The fuzzy function F: R $\longrightarrow E^1$ is said to be continuous if:

For an arbitrary $t_1 \in R$ and $\epsilon > 0$ there exists a $\delta > 0$ such that:

$|t - t_1| < \delta \Rightarrow D(F(t), F(t_1)) < \epsilon$, where D is the distance between two fuzzy numbers.

Definition (5),[16]: Let I be a real interval. The r-level set of the fuzzy function y: $I \rightarrow E^1$ can be denoted by:

$$[y(x)]^r = [y_1^r(x), y_2^r(x)],\ x \in I, r \in [0,1] \tag{5}$$

The Seikkala derivative $y'(x)$ of the fuzzy function $y(x)$ is defined by:

$$[y'(x)]^r = [(y_1^r)'(x), (y_2^r)'(x)],\ x \in I, r \in [0,1] \tag{6}$$

Definition (6),[10]: let u and $v \in E^1$. If there exist $w \in E^1$ such that:

$u = v+w$ then w is called the H-difference (Hukuhara-difference) of u and v and it is denoted by $w = u \ominus v$.

In this work, the $\ominus$ sign stands always for H-difference, and let us remark that $u \ominus v \neq u + (-1)\, v$.

Definition (7), Fuzzy Derivative[12]: Let $F : (a,b) \rightarrow E^1$ and $t_0 \in (a,b)$.We say that F is H-differential (Hukuhara-differential) at x_0, if there exists an element $F'(x_0) \in E^1$ such that for all $h > 0$ (sufficiently small), $\exists$ $F(x_0 + h) \ominus F(x_0)$, $F(x_0) \ominus F(x_0 - h)$ and the limits (in the metric D)

$$\lim_{h\to 0} \frac{F(x_0 + h) \ominus F(x_0)}{h} = \lim_{h\to 0} \frac{F(x_0) \ominus F(x_0 - h)}{h} = F'(x_0) \tag{7}$$

Then $F'(x_0)$ is called fuzzy derivative (H-derivative) of F at x_0.

where D is the distance between two fuzzy numbers.

3 Fully fuzzy neural network [6,16]

Artificial neural networks are learning machines that can learn any arbitrary functional mapping between input and output. They are fast machines and can be implemented in parallel, either in software or in hardware. In fact, the computational complexity of ANN is polynomial in the number of neurons used in the network. Parallelism also brings with it the advantages of robustness and fault tolerance. (i.e.) ANN is a simplified mathematical model of the human brain. It can be implemented by both electric elements and computer software. It is a parallel distributed processor with large numbers of connections It is an information processing system that has certain performance characters in common with biological neural networks.

A fuzzy neural network or neuro-fuzzy system is a learning machine that finds the parameters of a fuzzy system (i.e., fuzzy set, fuzzy rules) by exploiting approximation techniques from neural networks. Combining fuzzy systems with neural networks. Both neural networks and fuzzy systems have some things in common. They can be used for solving problems (e. g. fuzzy differential equations, fuzzy integral equations, etc).

If all the adjustable parameters (weights and biases) are fuzzy numbers, then the fuzzy neural network is called fully fuzzy neural network; otherwise it is called partially fuzzy neural network.

4 Solution of FDEs by fully fuzzy neural network

To solve any fuzzy ordinary differential equation, we consider a three-layered fully fuzzy neural network with one unit entry x, one hidden layer consisting of m activation functions and one unit output $N(x)$. The activation function for the hidden units of our fully fuzzy neural network is the hyperbolic tangent function ($s(\propto) = \tanh(\propto)$). Here the dimension of a fully fuzzy neural network is $(1 \times m \times 1)$ (figure1).

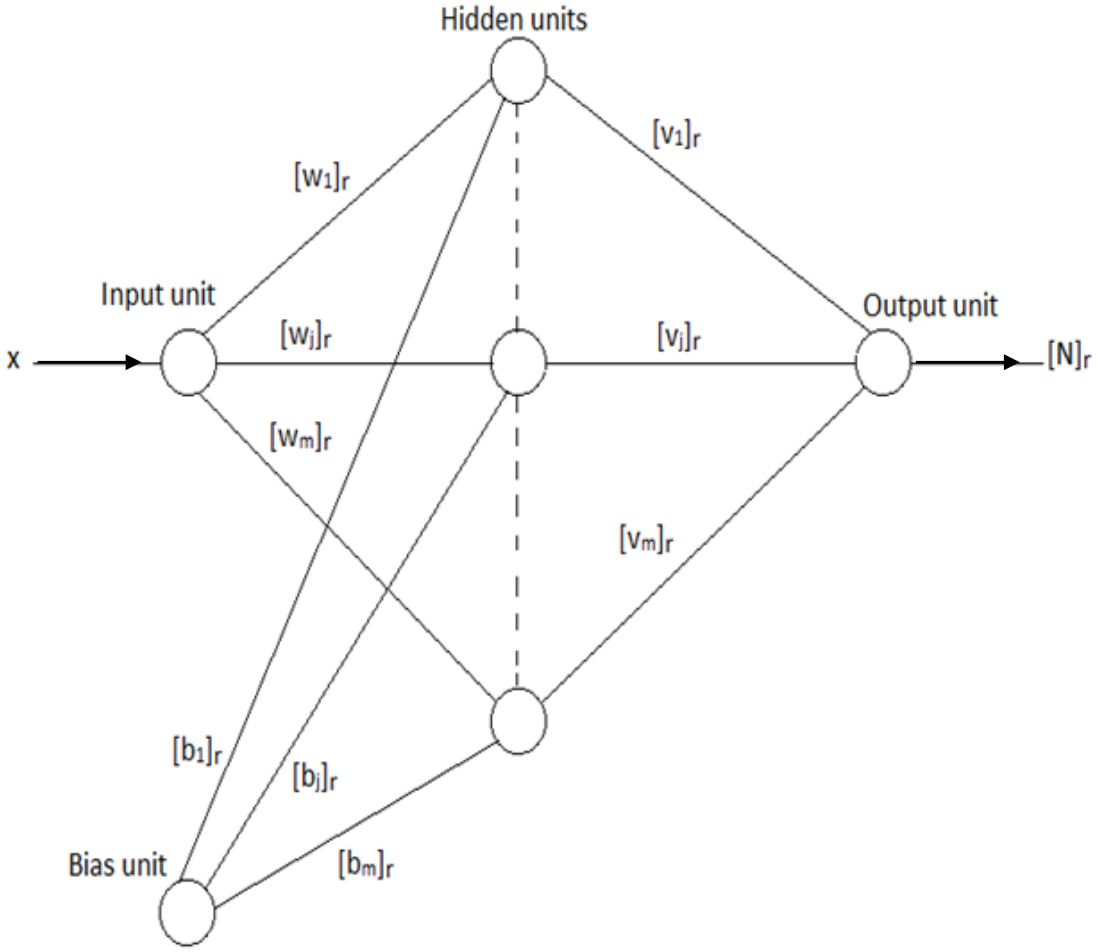

Figure1: (1 × m × 1) Fully fuzzy feed-forward neural network.

For every entry x (where $x \geq 0$) the mathematical operations in the fully fuzzy neural network can be described as:

Input unit: x = x, (8)

Hidden units :

$$[z_j]_r = \left[[z_j]_r^L,\ [z_j]_r^U\right] = \left[s\left([net_j]_r^L\right), s\left([net_j]_r^U\right)\right] \quad (9)$$

where

$$[net_j]_r^L = x\,[w_j]_r^L + [b_j]_r^L \quad (10)$$

$$[net_j]_r^U = x\,[w_j]_r^U + [b_j]_r^U \quad (11)$$

Output unit:

$$[N(x)]_r = [[N(x)]_r^L\ ,\ [N(x)]_r^U] \tag{12}$$

Where

$$[N(x)]_r^L = \sum_{j=1}^{m} \min\{ [v_j]_r^L [z_j]_r^L , [v_j]_r^L [z_j]_r^U , [v_j]_r^U [z_j]_r^L , [v_j]_r^U [z_j]_r^U \} \tag{13}$$

$$[N(x)]_r^U = \sum_{j=1}^{m} \max\{ [v_j]_r^L [z_j]_r^L , [v_j]_r^L [z_j]_r^U , [v_j]_r^U [z_j]_r^L , [v_j]_r^U [z_j]_r^U \} \tag{14}$$

Where

$$[z_j]_r^L = s\left(x\,[w_j]_r^L + [b_j]_r^L\right) \tag{15}$$

$$[z_j]_r^U = s\left(x\,[w_j]_r^U + [b_j]_r^U\right) \tag{16}$$

5 Description of the proposed method

For illustration the proposed method, we will consider the two points fuzzy boundary value problems:

$$y''(x) = f(x, y(x)\ ,\ y'(x))\ ,\quad x \in [a\ , b] \tag{17}$$

with the fuzzy boundary conditions:

$y(a) = A$ and $y(b) = B$, where A and B are fuzzy numbers in E^1 with r-level sets:

$[A]_r = [\underline{A}\ , \overline{A}]$ and $[B]_r = [\underline{B}\ , \overline{B}]$.

The fuzzy trial solution for this problem is:

$$[y_t(x)]_r = \frac{b-x}{b-a}\ [A]_r + \frac{x-a}{b-a}[B]_r\ + (x-a)\,(x-b)\,[N(x)]_r \tag{18}$$

This fuzzy trial solution by intention satisfies the fuzzy boundary conditions in (17).

The error function that must be minimized for problem (17) is in the form:

$$E = \sum_{i=1}^{g}\left(E_{ir}^L + E_{ir}^U\right) \tag{19}$$

where

$$E_{ir}^{L} = \left[\left[\frac{d^2 y_t(x_i)}{dx^2} \right]_r^L - \left[f\left(x_i, y_t(x_i), \frac{d\, y_t(x_i)}{dx} \right) \right]_r^L \right]^2 \quad (20)$$

$$E_{ir}^{U} = \left[\left[\frac{d^2 y_t(x_i)}{dx^2} \right]_r^U - \left[f\left(x_i, y_t(x_i), \frac{d\, y_t(x_i)}{dx} \right) \right]_r^U \right]^2 \quad (21)$$

where $\{x_i\}_{i=1}^{g}$ are discrete points belonging to the interval $[a, b]$ (training set) and in the cost function (19), E_r^L and E_r^U can be viewed as the squared errors for the lower limits and the upper limits of the r – level sets, respectively.

Now, to drive the minimized error function for problem (17):

From (18) we can find:

$$[y_t(x)]_r^L = \frac{b-x}{b-a}\ [A]_r^L + \frac{x-a}{b-a}\ [B]_r^L + (x^2 - (a+b)x + ab)[N(x)]_r^L \quad (22)$$

$$[y_t(x)]_r^U = \frac{b-x}{b-a}\ [A]_r^U + \frac{x-a}{b-a}\ [B]_r^U + (x^2 - (a+b)x + ab)[N(x)]_r^U \quad (23)$$

Then we get:

$$\frac{d[y_t(x)]_r^L}{dx} = \frac{-1}{b-a}\ [A]_r^L + \frac{1}{b-a}\ [B]_r^L + (x^2 - (a+b)x + ab)\ \frac{d[N(x)]_r^L}{dx} + (2x - a - b)[N(x)]_r^L \quad (24)$$

$$\frac{d[y_t(x)]_r^U}{dx} = \frac{-1}{b-a}\ [A]_r^U + \frac{1}{b-a}\ [B]_r^U + (x^2 - (a+b)x + ab)\ \frac{d[N(x)]_r^U}{dx} + (2x - a - b)[N(x)]_r^U \quad (25)$$

Therefore, we have:

$$\left[\frac{d^2 y_t(x)}{dx^2} \right]_r^L = (x^2 - (a+b)x + ab)\ \frac{d^2[N(x)]_r^L}{dx^2} + 2(2x - a - b)\ \frac{d[N(x)]_r^L}{dx} + 2[N(x)]_r^L \quad (26)$$

$$\left[\frac{d^2 y_t(x)}{dx^2} \right]_r^U = (x^2 - (a+b)x + ab)\ \frac{d^2[N(x)]_r^U}{dx^2} + 2(2x - a - b)\ \frac{d[N(x)]_r^U}{dx} + 2[N(x)]_r^U \quad (27)$$

Then (20) and (21) can be rewritten as:

$$E_{ir}^{L} = [(x_i^2 - (a+b)x_i + ab)\ \frac{d^2[N(x_i)]_r^L}{dx^2} + 2(2x_i - a - b)\ \frac{d[N(x_i)]_r^L}{dx} + 2[N(x_i)]_r^L - f(x_i, \frac{b-x_i}{b-a}\ [A]_r^L + \frac{x_i - a}{b-a}\ [B]_r^L + (x_i^2 - (a+b)x_i + ab)[N(x_i)]_r^L, \frac{-1}{b-a}\ [A]_r^L + \frac{1}{b-a}\ [B]_r^L + (x_i^2 - (a+b)x_i + ab)\ \frac{d[N(x_i)]_r^L}{dx} + (2x_i - a - b)[N(x_i)]_r^L)]^2 \quad (28)$$

$$E_{ir}^{U} = [(x_i^2 - (a+b)x_i + ab)\ \frac{d^2[N(x_i)]_r^U}{dx^2} + 2(2x_i - a - b)\ \frac{d[N(x_i)]_r^U}{dx} + 2[N(x_i)]_r^U - f(x_i, \frac{b-x_i}{b-a}\ [A]_r^U + \frac{x_i - a}{b-a}\ [B]_r^U + (x_i^2 - (a+b)x_i + ab)[N(x_i)]_r^U, \frac{-1}{b-a}\ [A]_r^U + \frac{1}{b-a}\ [B]_r^U + (x_i^2 - (a+b)x_i + ab)\ \frac{d[N(x_i)]_r^U}{dx} + (2x_i - a - b)[N(x_i)]_r^U)]^2 \quad (29)$$

Where

$$[N(x_i)]_r^L = \sum_{j=1}^{m} \min\{[v_j]_r^L\, s\left(x_i [w_j]_r^L + [b_j]_r^L\right), [v_j]_r^L\, s\left(x_i [w_j]_r^U + [b_j]_r^U\right), [v_j]_r^U\, s\left(x_i [w_j]_r^L + [b_j]_r^L\right), [v_j]_r^U\, s\left(x_i [w_j]_r^U + [b_j]_r^U\right)\} \quad (30)$$

$$[N(x_i)]_r^L = \sum_{j=1}^{m} \max\{[v_j]_r^L\, s\left(x_i [w_j]_r^L + [b_j]_r^L\right), [v_j]_r^L\, s\left(x_i [w_j]_r^U + [b_j]_r^U\right), [v_j]_r^U\, s\left(x_i [w_j]_r^L + [b_j]_r^L\right), [v_j]_r^U\, s\left(x_i [w_j]_r^U + [b_j]_r^U\right)\} \quad (31)$$

$$\frac{d[N(x_i)]_r^L}{dx} = \sum_{j=1}^{m} \min\{[v_j]_r^L [w_j]_r^L s'\left(x_i [w_j]_r^L + [b_j]_r^L\right), [v_j]_r^L [w_j]_r^U s'\left(x_i [w_j]_r^U + [b_j]_r^U\right), [v_j]_r^U [w_j]_r^L s'\left(x_i [w_j]_r^L + [b_j]_r^L\right), [v_j]_r^U [w_j]_r^U s'\left(x_i [w_j]_r^U + [b_j]_r^U\right)\} \quad (32)$$

$$\frac{d[N(x_i)]_r^U}{dx} = \sum_{j=1}^{m} \max\{[v_j]_r^L [w_j]_r^L s'\left(x_i [w_j]_r^L + [b_j]_r^L\right), [v_j]_r^L [w_j]_r^U s'\left(x_i [w_j]_r^U + [b_j]_r^U\right), [v_j]_r^U [w_j]_r^L s'\left(x_i [w_j]_r^L + [b_j]_r^L\right), [v_j]_r^U [w_j]_r^U s'\left(x_i [w_j]_r^U + [b_j]_r^U\right)\} \quad (33)$$

$$\frac{d^2[N(x_i)]_r^L}{dx^2} = \sum_{j=1}^{m} \min\{[v_j]_r^L ([w_j]_r^L)^2 s''\left(x_i [w_j]_r^L + [b_j]_r^L\right), [v_j]_r^L ([w_j]_r^U)^2\, s''\left(x_i [w_j]_r^U + [b_j]_r^U\right), [v_j]_r^U ([w_j]_r^L)^2 s''\left(x_i [w_j]_r^L + [b_j]_r^L\right), [v_j]_r^U ([w_j]_r^U)^2 s''\left(x_i [w_j]_r^U + [b_j]_r^U\right)\} \quad (34)$$

$$\frac{d^2[N(x_i)]_r^U}{dx^2} = \sum_{j=1}^{m} \max\{ [v_j]_r^L ([w_j]_r^L)^2 s''\left(x_i [w_j]_r^L + [b_j]_r^L\right), [v_j]_r^L ([w_j]_r^U)^2 s''\left(x_i [w_j]_r^U + [b_j]_r^U\right), [v_j]_r^U ([w_j]_r^L)^2 s''\left(x_i [w_j]_r^L + [b_j]_r^L\right), [v_j]_r^U ([w_j]_r^U)^2 s''\left(x_i [w_j]_r^U + [b_j]_r^U\right) \} \quad (35)$$

where s' and s'' are the first and second derivative of the hyperbolic tangent function. Then we substitute (28) and (29) in (19) to find the error function that must be minimized for problem (17).

6. Numerical example

In this section, we will solve two problems about two-point fuzzy boundary value problem. We have used $(1 \times 10 \times 1)$ fully fuzzy feed-forward neural network. The activation function of each hidden unit is the hyperbolic tangent activation function. The analytical solutions $[y_a(x)]_r^L$ and $[y_a(x)]_r^U$ has been known in advance. Therefore, we test the accuracy of the obtained solutions by computing the deviation:

$$\overline{e}(x, r) = |[y_a(x)]_r^U - [y_t(x)]_r^U| \text{ , } \underline{e}(x, r) = |[y_a(x)]_r^L - [y_t(x)]_r^L|$$

To minimize the error function, we have used BFGS quasi-Newton method (For more details, see [16]). The computer programs which we have used in this work are coded in MATLAB 2015.

Example (1): Consider the linear fuzzy boundary value problem:

$$y''(x) - y'(x) = 1 \text{ . with } x \in [0, 0.5]$$

$$y(0) = [\, 2 + r \, , \, 4 - r \,] \, ,$$

$$y(0.5) = [5 + r \, , \, 7 - r] \quad . \quad \text{where} \quad r \in [0, 1].$$

The analytical solutions for this problem are:

$$[y_a(x)]_r^L = \left(2 + r - \frac{3}{e^{0.5} - 1}\right) + \left(\frac{3}{e^{0.5} - 1}\right) e^x$$

$$[y_a(x)]_r^U = \left(4 - r - \frac{3}{e^{0.5} - 1}\right) + \left(\frac{3}{e^{0.5} - 1}\right) e^x$$

The trial solutions for this problem are:

$$[y_t(x)]_r^L = (1 - 2x)(2 + r) + 2x(4 - r) + (x^2 - 0.5\,x)[N(x)]_r^L$$

$$[y_t(x)]_r^U = (1-2x)\,(5+r) \;+\; 2x\,(7-r) + (x^2 - 0.5\,x\,)\,[N(x)]_r^U$$

The fully fuzzy feed forward neural network has been trained by using a grid of ten equidistant points in [0, 0.5].

The error function that must be minimized for this problem will be:

$$E = \sum_{i=1}^{11} \left(E_{ir}^L + E_{ir}^U\right) \tag{36}$$

where

$$E_{ir}^L = [(\,x_i{}^2 - 0.5x_i)\frac{d^2[N(x_i)]_r^L}{dx^2} + (4x_i - 1)\frac{d[N(x_i)]_r^L}{dx} + 2[N(x_i)]_r^L - (\,x_i{}^2 - 0.5x_i)\frac{d[N(x_i)]_r^L}{dx} - (2x_i - 0.5)[N(x_i)]_r^L + 4r - 5\,]^2 \tag{37}$$

$$E_{ir}^U = [(\,x_i{}^2 - 0.5x_i)\frac{d^2[N(x_i)]_r^U}{dx^2} + (4x_i - 1)\frac{d[N(x_i)]_r^U}{dx} + 2[N(x_i)]_r^U - (\,x_i{}^2 - 0.5x_i)\frac{d[N(x_i)]_r^U}{dx} - (2x_i - 0.5)[N(x_i)]_r^U + 4r - 5\,]^2 \tag{38}$$

numerical solutions for this problem can be found in table (1).

Table (1): Numerical result for example (1), x=1.

r	$[y_t(x)]_r^L$	$\underline{e}\,(x\,,r)$	$[y_t(x)]_r^U$	$\overline{e}\,(x\,,r)$
0	9.946164141	3.29137e-7	11.94616425	4.33916e-7
0.1	10.04616401	1.96846e-7	11.84616411	2.93475e-7
0.2	10.14616481	9.95565e-7	11.74616478	9.70548e-7
0.3	10.24616458	7.63284e-7	11.64616385	3.95104e-8
0.4	10.34616447	6.60993e-7	11.54616387	5.67802e-8
0.5	10.44616422	4.09513e-7	11.44616389	7.56011e-8
0.6	10.54616396	1.47232e-7	11.34616391	9.53493e-8
0.7	10.64616391	9.75941e-8	11.24616382	1.15291e-8
0.8	10.74616385	3.39072e-8	11.14616384	2.63433e-8
0.9	10.84616389	7.52383e-8	11.04616386	5.26859e-8
1	10.94616389	7.39070e-8	10.94616386	4.56782e-8

Example (2): Consider the non-linear fuzzy boundary value problem:

$y'' (x) = - \left(y'(x)\right)^2$. with $x \in [0 , 2]$

$y(0) = [r , 2 - r] , y(2) = [1 + r , 3 - r]$ and $r \in [0 , 1]$.

The analytical solutions for this problem are:

$$[y_a(x)]_r^L = \ln \left(x + \frac{2}{e-1}\right) + r - \ln \frac{2}{e-1}$$

$$[y_a(x)]_r^U = \ln \left(x + \frac{2}{e-1}\right) + 2 - r - \ln \frac{2}{e-1}$$

The trial solutions for this problem are:

$$[y_t(x)]_r^L = r \frac{2-x}{2} + (1 + r) \frac{x}{2} + x (x - 2) [N(x)]_r^L$$

$$[y_t(x)]_r^U = (2 - r) \frac{2-x}{2} + (3 - r) \frac{x}{2} + x (x - 2) [N(x)]_r^U$$

The fully fuzzy feed forward neural network has been trained by using a grid of ten equidistant points in [0, 2].

The error function that must be minimized for this problem will be:

$$E= \sum_{i=1}^{11} \left(E_{ir}^L + E_{ir}^U\right) \quad (39)$$

where

$$E_{ir}^L=[(x_i^2 - 2x_i) \frac{d^2[N(x_i)]_r^L}{dx^2} + (4x_i - 4) \frac{d[N(x_i)]_r^L}{dx} + 2[N(x_i)]_r^L + ((x_i^2 - 2x_i) \frac{d[N(x_i)]_r^L}{dx} + (2x_i - 2)[N(x_i)]_r^L + 0.5)^2]^2 \quad (40)$$

$$E_{ir}^L=[(x_i^2 - 2x_i) \frac{d^2[N(x_i)]_r^U}{dx^2} + (4x_i - 4) \frac{d[N(x_i)]_r^U}{dx} + 2[N(x_i)]_r^U + ((x_i^2 - 2x_i) \frac{d[N(x_i)]_r^U}{dx} + (2x_i - 2)[N(x_i)]_r^U + 0.5)^2]^2 \quad (41)$$

Then we use (39) to update the weights and biases.

Numerical solution for this problem can be found in table (2).

Table (2): Numerical result for example (2), x=1.

r	$[y_t(x)]_r^L$	$\underline{e}$ (x , r)	$[y_t(x)]_r^U$	$\overline{e}$ (x , r)
0	0.620114507	3.24734e-10	2.620114507	8.46634e-10
0.1	0.720114507	4.66221-10	2.520114507	9.79602e-10
0.2	0.820114507	2.03208e-10	2.420114507	6.85555e-10
0.3	0.920114507	3.80684e-10	2.320114513	6.62032e-9
0.4	1.020114507	4.09557e-10	2.220114514	7.59010e-9
0.5	1.120114507	3.50405e-10	2.120114508	1.74006e-9
0.6	1.220114507	4.59008e-10	2.020114507	9.00817e-10
0.7	1.320114516	9.46681e-9	1.920114507	9.21604e-10
0.8	1.420114512	5.06564e-9	1.820114507	4.99811e-10
0.9	1.520114507	8.21899e-10	1.720114514	7.15955e-9
1	1.620114514	7.88763e-9	1.620114508	1.02988e-9

For the above two problems we have

$$[N(x_i)]_r^L = \sum_{j=1}^{10} \min\{ [v_j]_r^L s\left(x_i [w_j]_r^L + [b_j]_r^L\right), [v_j]_r^L s\left(x_i [w_j]_r^U + [b_j]_r^U\right), [v_j]_r^U s\left(x_i [w_j]_r^L + [b_j]_r^L\right), [v_j]_r^U s\left(x_i [w_j]_r^U + [b_j]_r^U\right)\}$$

$$[N(x_i)]_r^L = \sum_{j=1}^{10} \max\{ [v_j]_r^L s\left(x_i [w_j]_r^L + [b_j]_r^L\right), [v_j]_r^L s\left(x_i [w_j]_r^U + [b_j]_r^U\right), [v_j]_r^U s\left(x_i [w_j]_r^L + [b_j]_r^L\right), [v_j]_r^U s\left(x_i [w_j]_r^U + [b_j]_r^U\right)\}$$

$$\frac{d[N(x_i)]_r^L}{dx} = \sum_{j=1}^{10} \min\{ [v_j]_r^L [w_j]_r^L s'\left(x_i [w_j]_r^L + [b_j]_r^L\right), [v_j]_r^L [w_j]_r^U s'\left(x_i [w_j]_r^U + [b_j]_r^U\right), [v_j]_r^U [w_j]_r^L s'\left(x_i [w_j]_r^L + [b_j]_r^L\right), [v_j]_r^U [w_j]_r^U s'\left(x_i [w_j]_r^U + [b_j]_r^U\right)\}$$

$$\frac{d[N(x_i)]_r^U}{dx} = \sum_{j=1}^{10} \max\{ [v_j]_r^L [w_j]_r^L s'\left(x_i [w_j]_r^L + [b_j]_r^L\right), [v_j]_r^L [w_j]_r^U s'\left(x_i [w_j]_r^U + [b_j]_r^U\right), [v_j]_r^U [w_j]_r^L s'\left(x_i [w_j]_r^L + [b_j]_r^L\right), [v_j]_r^U [w_j]_r^U s'\left(x_i [w_j]_r^U + [b_j]_r^U\right) \}$$

$$\frac{d^2[N(x_i)]_r^L}{dx^2} = \sum_{j=1}^{10} \min\{ [v_j]_r^L ([w_j]_r^L)^2 s''\left(x_i [w_j]_r^L + [b_j]_r^L\right), [v_j]_r^L ([w_j]_r^U)^2 s''\left(x_i [w_j]_r^U + [b_j]_r^U\right), [v_j]_r^U ([w_j]_r^L)^2 s''\left(x_i [w_j]_r^L + [b_j]_r^L\right), [v_j]_r^U ([w_j]_r^U)^2 s''\left(x_i [w_j]_r^U + [b_j]_r^U\right) \}$$

$$\frac{d^2[N(x_i)]_r^U}{dx^2} = \sum_{j=1}^{10} \max\{ [v_j]_r^L ([w_j]_r^L)^2 s''\left(x_i [w_j]_r^L + [b_j]_r^L\right), [v_j]_r^L ([w_j]_r^U)^2 s''\left(x_i [w_j]_r^U + [b_j]_r^U\right), [v_j]_r^U ([w_j]_r^L)^2 s''\left(x_i [w_j]_r^L + [b_j]_r^L\right), [v_j]_r^U ([w_j]_r^U)^2 s''\left(x_i [w_j]_r^U + [b_j]_r^U\right) \}$$

7 Conclusion

In this work, we have introduced a modified method to find the numerical solution of the two-point fuzzy boundary value problems for the ordinary differential equations. This method based on the fully fuzzy neural network to approximate the solution of the second-order fuzzy differential equations. For future studies, one can extend this method to find a numerical solution of the higher order fuzzy differential equations. Also, one may use this method for solving a fuzzy partial differential equation.

References

[1] H. Lee, I.S. Kang. Neural Algorithms for Solving Differential Equations. Journal of Computational Physics, 91, 110-131. 1990.

[2] A.J. Meade, A.A. Fernandes. The Numerical Solution of Linear Ordinary Differential Equations by Feed-Forward Neural Networks. Mathematical and Computer Modeling, 19(12), 1-25. 1994.

[3] A.J. Meade, A.A. Fernandes. Solution of Nonlinear Ordinary Differential Equations by Feed-Forward Neural Networks. Mathematical and Computer Modeling, 20(9), 19-44. 1994.

[4] I.E. Lagaris, A. Likas. Artificial Neural Networks for Solving Ordinary and Partial Differential Equations. Journal of Computational Physics, 104, 1-26. 1997.

[5] Liu, Jammes. Solving Ordinary Differential Equations by Neural Networks. Warsaw, Poland. 1999.

[6] Tawfiq. On Design and Training of Artificial Neural Network for Solving Differential Equations. Ph.D. Thesis, College of Education, University of Baghdad, Iraq. 2004.

[7] A. Malek, R. Shekari. Numerical Solution for High Order Differential Equations by Using a Hybrid Neural Network Optimization Method. Applied Mathematics and Computation, 183, 260-271. 2006.

[8] S. Pattanaik, R.K. Mishra. Application of ANN for Solution of PDE in RF Engineering. International Journal on Information Sciences and Computing, 2(1), 74-79. 2008.

[9] M. Baymani, A. Kerayechian. Artificial Neural Networks Approach for Solving Stokes Problem, Applied Mathematics, 1, 288-292. 2010.

[10] S. Effati, M. Pakdaman. Artificial Neural Network Approach for Solving Fuzzy Differential Equations. Information Sciences, 180, 1434-1457. 2010.

[11] M. Mosleh, M. Otadi. Fuzzy Fredholm Integro-Differential Equations with Artificial Neural Networks. Communications in Numerical Analysis, Article ID cna-00128, 1-13. 2012.

[12] S Ezadi, N. Parandin. Numerical Solution of Fuzzy Differential Equations Based on Semi-Taylor by Using Neural Network. Journal of Basic and Applied Scientific Research, 3(1s), 477-482. 2013.

[13] M. Mosleh, M. Otadi. Simulation and Evaluation of Fuzzy Differential Equations by Fuzzy Neural Network. Applied Soft Computing, 12, 2817-2827. 2012.

[14] M. Mosleh. Fuzzy Neural Network For Solving a System of Fuzzy Differential Equations. Applied Soft Computing, 13, 3597-3607. 2013.

[15] M. Mosleh, M. Otadi. Solving the Second Order Fuzzy Differential Equations by Fuzzy Neural Network. Journal of Mathematical Extension, 8(1), 11-27. 2014.

[16] Suhhiem. Fuzzy Artificial Neural Network for Solving Fuzzy and Non-Fuzzy Differential Equations. Ph.D. Thesis, College of Sciences, AL-Mustansiriyah University, Iraq. 2016.

Ratio Mathematica Volume 36, 2019, pp. 43-52

The inclusion and exclusion principle in view of number theory

Viliam Ďuriš*

Tomáš Lengyelfalusy †

Abstract

The inclusion and exclusion (connection and disconnection) principle is mainly known from combinatorics in solving the combinatorial problem of calculating all permutations of a finite set or other combinatorial problems. Finite sets and Venn diagrams are the standard methods of teaching this principle. The paper presents an alternative approach to teaching the inclusion and exclusion principle from the number theory point of view, while presenting several selected application tasks and possible principle implementation into the Matlab computing environment.

Keywords: inclusion, exclusion, number theory, combinatorics, Matlab

2010 AMS subject classification: 11B75.‡

* Department of Mathematics, Faculty of Natural Sciences Constantine the Philosopher University in Nitra, Tr. A. Hlinku 1, 949 74 Nitra, Slovakia; vduris@ukf.sk.

† Department of Didactics, Technology and Educational Technologies, DTI University Sládkovičova 533/20, 018 41 Dubnica nad Váhom, Slovakia; lengyelfalusy@dti.sk.

‡Received on May 2nd, 2019. Accepted on June 3rd, 2019. Published on June 30th, 2019. doi: 10.23755/rm.v36i1.465. ISSN: 1592-7415. eISSN: 2282-8214.

1 Introduction

In traditional secondary school mathematics (in combinatorics, number theory or even in probability theory), the notion of factorial and combinatorial numbers is introduced [1]. If *n* and *k* are two natural numbers with $n \geq k$, then we call *a combinatorial number* the following notation

$$\binom{n}{k} = \frac{n!}{(n-k)!\,k!} = \frac{n(n-1)\ldots(n-k+1)}{1 \cdot 2 \cdot \ldots \cdot k}$$

while (*factorial of the number n*) $n! = 1 \cdot 2 \cdot \cdots \cdot n$, where $n > 1$, $0! = 1$, $1! = 1$.

For combinatorial numbers, the basic properties apply:

$$\binom{n}{1} = n \quad \binom{n}{0} = 1 \quad \binom{0}{0} = 1 \quad \binom{n}{k} = \binom{n}{n-k} \quad \binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1}$$

The relation $\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1}$ is the basis for placing combinatorial numbers in the plane in the shape of a triangle (a so-called *Pascal's triangle*) [2], in which combinatorial numbers can be gradually calculated using the fact that $\binom{n}{0} = \binom{n}{n} = 1$ for each *n*.

$$\begin{array}{ccccccc} & & & \binom{0}{0} & & & \\ & & \binom{1}{0} & & \binom{1}{1} & & \\ & \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & \\ \binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} \\ & & & \ldots & & & \end{array}$$

If *n* is a natural number, and if *a*, *b* are arbitrary complex numbers, then the *binomial theorem* can be applied by using the form:

$$(a+b)^n = \binom{n}{0} a^n + \binom{n}{1} a^{n-1}b + \cdots + \binom{n}{n-1} ab^{n-1} + \binom{n}{n} b^n$$

The special cases of the binomial theorem are as follows:

a) if $a = 1, b = -1$:

$$1 - \binom{n}{1} + \cdots + (-1)^{n-1}\binom{n}{n-1} + (-1)^n = 0$$

b) if $a = 1, b = 1$:

$$(1+1)^n = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n-1} + \binom{n}{n} = 2^n$$

Let us consider now *N* given objects and *K* properties $a_1, \dots, a_K$. Let us denote $N(0)$ as the number of objects that do not have either of these properties, $N(a_i)$ as the number of those that have the property a_i, $N(a_i a_j)$ as the number of those that have the property a_i as well as a_j etc. Then

$$N(0) = N - \sum N(a_i) + \sum N(a_i a_j) - \sum N(a_i a_j a_s) + \cdots + (-1)^K N(a_1 a_2 \dots a_K),$$

where, in the first addition, we sum up using numbers $i = 1, 2, \dots, K$, in the second addition, using all pairs of these numbers, in the third addition, using all threesomes of these numbers, etc. We call this relationship *the inclusion and exclusion principle* [3].

The validity of the inclusion and exclusion principle can be shown from the number theory point of view the way that if an object has no property from the properties $a_i, i = 1, \cdots, K$, so it contributes by the unit value to the left equality, though contributing at the same time to the right side, that is, to the number *N* (in the following additions it does not reappear). Let an object now have *t* properties ($t \geq 1$). Then, it does not contribute to the left side as there is a number of objects on the left side that do not have any of the properties. Let us calculate the contribution of this object to the right side. In the first addition, it appears *t*-times. In the second addition, it appears $\binom{t}{2}$–times because from *t* properties it is possible to choose pairs of the properties in $\binom{t}{2}$ ways. In the third addition, it appears $\binom{t}{3}$–times, etc., so the total contribution to the right side is as follows:

$$1 - t + \binom{t}{2} - \binom{t}{3} + \dots + (-1)^{t-1}\binom{t}{t-1} + (-1)^t = 0,$$

which is a special case of the binomial theorem. Thus, the total contribution of such an object to both sides is zero and the right side is actually equal to the number of objects that do not have any of the given properties.

2 Selected examples of the inclusion and exclusion principle

The first example requires some mathematical concepts to be recalled. By the *Cartesian product* of sets A, B we mean set $A \times B = \{[x, y]: x \in A \wedge y \in B\}$, with the symbol $|A|$ we denote the number of elements (so-called *cardinality*) of the finite set A. If $|A| = a$, $|B| = b$, the Cartesian product then contains $a \cdot b$ of ordered pairs. Since the Cartesian product contains ordered pairs, $A \times B$ is not the same set as $B \times A$. [4]
The relation f of set A to set B is called a function of set A to set B if $\forall x \in A \exists y \in B: [x, y] \in f$ and simultaneously if $[x, y] \in f \wedge [x, z] \in f$, so $y = z$. The symbol B^A denotes a set of all functions $A \to B$.
If f is a function of set A into set B and $\forall x_1, x_2 \in A: x_1 \neq x_2 \Rightarrow f(x_1) \neq f(x_2)$, the function f is called an injective function of set A into set B (or simply an *injection*; we also say that the function f is ordinary).
Let us now consider two finite sets A, B, where $|A| = n$ and $|B| = m$. Then the number of all injective functions from A into B is $m \cdot (m-1) \cdot \cdots \cdot (m-n+1) = \prod_{i=0}^{n-1}(m-i)$. Injections from set $A = \{1,2,\cdots,n\}$ into set B, where $|B| = m$, are called *variations without repetition (or simply variations) of the n-th class from m elements* (of the set B). For these functions, the term $V_n(m)$ is used in practice. It is easier to write the expression $m \cdot (m-1) \cdot \cdots \cdot (m-n+1)$ with the following factorial notation $V_n(m) = \frac{m!}{(m-n)!}$.
Variations of the n-th class from n elements of the set B are bijective functions $A \to B$ and their number is $n \cdot (n-1) \cdot \cdots \cdot 2 \cdot 1 = n!$. They are called *permutations* (of set B) and denote $P(n) = n!$.
Let us now consider basic set A with the cardinality $|A| = n$. *Combinations (without repetition) of the k-th class (or k-combinations) from n elements* are k-element subsets of set A. We denote them as $C_k(n)$. If A is a finite set, with $|A| = n$, then, the number of k-combinations of elements of set A is $C_k(n) = \binom{n}{k} = \frac{n!}{(n-k)!k!} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1}$. [5]

Example 2.1. A group of N men is to take part in a chess tournament. Before entering the room, they place their coats in the locker room. However, when they are about to leave, they are unable to recognize their coats. What is the probability that none of them will take their own coat?

Solution. Let us denote the coats $1,2,\cdots,N$. Then the distribution of the coats on the chess players can be made $N!$, since these are the permutations of the set $\{1,2,\cdots,N\}$. First, we determine the number $N(0)$ of permutations, for which there is no coat on the right player. The number of permutations that do not leave

in place the k-element set of coats is $(N-k)!$ The number of k-sets can be chosen in $\binom{N}{k}$ ways.

Then, based on the inclusion and exclusion principle, there applies

$$N(0) = N - \binom{N}{1}(N-1)! + \binom{N}{2}(N-2)! - \cdots + (-1)^N \binom{N}{N}(N-N)!$$

$$N(0) = \sum_{k=0}^{N} (-1)^k \binom{N}{k}(N-k)!$$

Next, we get

$$N(0) = \sum_{k=0}^{N} (-1)^k \frac{N!}{k!\,(N-k)!}(N-k)! = N! \sum_{k=0}^{N} \frac{(-1)^k}{k!}$$

All permutations of N elements is $N!$, hence the likelihood that no chess player is wearing his coat when leaving the tournament is

$$\frac{N! \sum_{k=0}^{N} \frac{(-1)^k}{k!}}{N!} = \sum_{k=0}^{N} \frac{(-1)^k}{k!}$$

Example 2.2. A tennis centre has a certain number of players and 4 groups A, B, C, D. Each player trains in at least one group, while some players train in multiple groups at once according to the table.

A.....26	AC.....18	ABC.....5
B.....17	AD.....3	ABD.....0
C.....58	BC.....9	ACD.....2
D.....19	BD.....0	BCD.....0
AB.....7	CD.....5	ABCD.....0

We will show how many players have a tennis centre.

Solution. Let us denote M_1 as the set of all players in group A, M_2 as the set of all players in group B, M_3 as the set of all players in group C and M_4 as the set of all players in group D. Then, set $N = M_1 \cup M_2 \cup M_3 \cup M_4$ is a set of all players in the centre.

Based on the inclusion and exclusion principle, there applies:

$$0 = |M_1 \cup M_2 \cup M_3 \cup M_4| - (26 + 17 + 59 + 19) + (7 + 18 + 3 + 9 + 5) - (5 + 2) + 0$$

From which $|M_1 \cup M_2 \cup M_3 \cup M_4| = 26 + 17 + 59 + 19 - 7 - 18 - 3 - 9 - 5 + 5 + 2 = 85$. As a result, the tennis centre has 85 players.

Example 2.3. Let $n > 1$ be a natural number. In number theory, the symbol $\varphi(n)$ denotes the number of natural numbers smaller than *n* and relatively prime s *n*, where $\varphi(n)$ is called Euler's function [3]. Let $n = p_1^{\alpha_1} \dots p_k^{\alpha_k}$ be a canonical decomposition of the number *n*. We will show that the following relation applies:

$$\varphi(n) = n\left(1 - \frac{1}{p_1}\right)\left(1 - \frac{1}{p_2}\right) \dots \left(1 - \frac{1}{p_k}\right)$$

Solution. Once more, we will use the inclusion and exclusion principle. Let $n = p_1^{\alpha_1} p_2^{\alpha_2} \dots p_k^{\alpha_k}$ is a canonical decomposition of the number *n*. The natural numbers that are relatively prime with the number *n* are those that are not divisible by either of the prime numbers $p_1, p_2, \dots, p_k$. So, let a_i mean the property that "the number *m* is divisible by the prime number $p_i, i = 1, \dots, k$". The number of numbers that are smaller or equal to the number *n* and are divisible by the number p_i is $N(a_i) = \frac{n}{p_i}$. It is an integer since $p_i \mid n$. Next, we get $N(a_i a_j) = \frac{n}{p_i p_j}$ and other members of the notation.

Then:

$$\varphi(n) = n - \sum \frac{n}{p_i} + \sum \frac{n}{p_i p_j} - \sum \frac{n}{p_i p_j p_s} + \dots + (-1)^k \frac{n}{p_1 p_2 \dots p_k}$$

This expression can be simplified to the form:

$$\varphi(n) = n\left(1 - \frac{1}{p_1}\right)\left(1 - \frac{1}{p_2}\right) \dots \left(1 - \frac{1}{p_k}\right)$$

Several other interesting tasks and applications of the inclusion and exclusion principle can be found e.g. in the resources [6], [7].

3 Implementation of the inclusion and exclusion principle in the Matlab computing environment

When solving various practical tasks with pupils, it is possible and appropriate to use some computing environment, e.g. Matlab. We will now solve a simple task of divisibility.

Example 3.1. We will show how many numbers there are up to 1000 that are not divisible by three, five, or seven.

Solution. Before proceeding to the solution of the task, we will use divisibility relations to determine the number of all natural numbers smaller than 1000, each of which can be divided simultaneously by three, five, and seven.

First, we will generally show that if $3|a$, $5|a$, then $3 \cdot 5 = 15|a$, being valid if $3|a$, so $a = 3b$, if $5|a$, so $a = 5c$. The left sides are equal, so the right sides must be equal, too. Then

$$3b = 5c$$

Since $(3,5) = 1 \Rightarrow 3|\mathrm{c} \Rightarrow c = 3d$. Then $a = 5c = 15d \Rightarrow 15|a$.

Now, we will show that if $15|a$, $7|a$, then $15 \cdot 7 = 105|a$ is valid if $15|a$, so $a = 15e$, if $7|a$, so $a = 7f$. Since $a = a$, it holds true that

$$15e = 7f$$

From the relation $(15,7) = 1 \Rightarrow 15|\mathrm{f} \Rightarrow f = 15g$. Then $a = 7f = 105g \Rightarrow 105|a$.

We will do the division $\frac{1000}{105} = 9 + \frac{55}{105}$ and we see that there exist 9 numbers with the required property.

Let us get back to our basic task. There, we have $N = 1000$. Let a_1 be the property that "the number *n* is divisible by three", property a_2 stand for "the number *n* is divisible by five", property a_3 stand for "the number *n* is divisible by seven". At the same time, $N(0)$ is the number of searched numbers not divisible by any of the numbers 3, 5, 7.
Every third natural number is divisible by three since $1000 = 3 \cdot 333 + 1$. We have the number $N(a_1) = 333$, that is 333 numbers up to 1000 are divisible by three. By similar consideration, we determine $N(a_2) = 200,\ N(a_3) = 142$.

Based on the previous considerations, we determine the number $N(a_1a_2)$. It holds true that if a number is divisible by three and five, it is also divisible by its product, i.e. by the number 15 (inasmuch as the numbers 3 and 5 are relatively prime). Hence, $N(a_1a_2)$ equals the number of numbers up to 1000 divisible by 15 and $N(a_1a_2) = 66$. Similarly, we determine $N(a_2a_3) = 28$ and $N(a_1a_3) = 47$. For the number $N(a_1a_2a_3)$ it is valid that it will be equal to the number of numbers up to 1000 that are divisible by the product $3 \cdot 5 \cdot 7 = 105$, hence $N(a_1a_2a_3) = 9$.

Then, based on the inclusion and exclusion principle, we have in total

$$N(0) = 1000 - (333 + 200 + 142) + (66 + 28 + 47) - 9 = 457$$

Now we implement the given task into the Matlab computing environment to verify the result. First we create the function "`count_the_divisors`", which is the application of the inclusion and exclusion principle:

```
function cnt = count_the_divisors(N, a, b, c)
    cnt_3  =  floor(N  /  a);  %counts  of  numbers
divisible by a
    cnt_5  =  floor(N  /  b);  %counts  of  numbers
divisible by b
    cnt_7  =  floor(N  /  c);  %counts  of  numbers
divisible by c

    cnt_3_5 = floor(N / (a * b)); %counts of numbers
divisible by a and b
    cnt_5_7 = floor(N / (b * c)); %counts of numbers
divisible by b and c
    cnt_3_7 = floor(N / (a * c)); %counts of numbers
divisible by a and c

    cnt_3_5_7 = floor(N / (a * b * c)); %counts of
numbers divisible by a, b and c

    %and now inclusion-exclusion principle applied
    cnt = N - (cnt_3 + cnt_5 + cnt_7) + (cnt_3_5 +
cnt_5_7 + cnt_3_7) - cnt_3_5_7;
```

We will call the function from the command line:

```
>> N = 1000;
>> count_the_divisors(N, 3, 5, 7)
```

```
ans =

   457
```

When creating functions or scripts solving various problems based on the inclusion and exclusion principle, it is possible to use various set operations (functions) built directly in Matlab without the need to create one's own structures. [8]

4 Conclusion

The principle of inclusion and exclusion is a "set problem" that falls within the field of discrete mathematics with different applications in combinatorics. However, this principle also plays a significant role in number theory when defining the so-called Euler's function or Fermat's theorem, or in clarifying and exploring the fundamental problems of number theory, such as expressing the distribution of prime numbers among natural numbers on the numerical axis and many other questions still open today.

The paper offered something different than just a set view of the inclusion and exclusion principle and its definition using number theory knowledge and the properties of combinatorial numbers. Our work is a guideline for solving selected practical tasks in which the involvement of the principle might not be expected at first sight. We also showed the possible application of ICT and the Matlab computing environment in solving computational problems in the field of number theory, which can be concurrently involved in mathematics teaching.

In conclusion, the inclusion and exclusion principle has much more application than we allege in our short contribution and can be used to solve more difficult tasks, e.g. in algebra to solve specific systems of equations or to solve various problems in combination with the Dirichlet principle. Some research shows that the ability to solve problems also depends on the substitution thinking, which makes possible to use mathematical knowledge effectively in various areas of number theory [9].

References

[1] J. Sedláček. *Faktoriály a kombinační čísla*. Praha, Mladá fronta, 1964.

[2] A. Vrba. *Kombinatorika*. Praha: Mladá fronta, 1980.

[3] Š. Znám. *Teória čísel*. Bratislava, SPN, 1975.

[4] M.T. Keller, W.T. Trotter. *Applied Combinatorics*. American Institute of Mathematics, 2017.

[5] M. Škoviera. *Úvod do diskrétnej matematiky*. Bratislava, Katedra informatiky FMFI UK, 2007.

[6] K. H. Rosen. *Discrete mathematics and its applications*. 4th ed., WCB/McGraw Hill, Boston, 1999.

[7] M.J. Erickson. *Introduction to Combinatorics*. John Wiley & Sons, New York, ISBN: 0-471-15408-3, 1996.

[8] Mathwork. *Online documentation*. 2019. Available at: https://www.mathworks.com/help/matlab/set-operations.html, Accessed 15th of April 2019.

[9] D. Gonda: *The Elements of Substitution Thinking and Its Impact On the Level of Mathematical Thinking*. In: IEJME — MATHEMATICS EDUCATION, vol. 11, no. 7, p. 2402-2417, Look Academic Publishers, 2016.

Ratio Mathematica Volume 36, 2019, pp. 53-68

Mathematics and radiotherapy of tumors

Luciano Corso[*]

Abstract

The present work takes inspiration from the scientific degree plan of the Italian Ministry of Education and has a didactic and cultural character. It pursues three objectives: the first is to make young people understand the importance of mathematics in medicine; the second is to stimulate students to use mathematical tools to give rational answers in the therapeutic field, in particular in the treatment of some types of nodular tumors; the third is to inform people on the effectiveness of mathematical methods and their indispensability in the rigorous treatment of some human pathologies.
Using the experimental data about the development of a tumor, we move on to the analysis of the mathematical models able to allow a rational control of its behavior. The method we used in the development of this therapeutic process is essentially deterministic, even if some passages implicitly have a probabilistic nature.

Keywords: population, cells, tumor, carrying capacity, differential equation

2010 AMS subject classification: 92B08, 92D08, 34K02, 39A06, 62P07, 97D06. [†]

[*]Founder and director of the journal MatematicaMente - President of the Verona section of Mathesis – Verona, Italy; e-mail: lcorso@iol.it.
[†]Received on May 1st, 2019. Accepted on June 20rd, 2019. Published on June 30th, 2019. doi:10.23755/rm.v36i1.471. ISSN: 1592-7415. eISSN: 2282-8214.

1. Premise

Usually, when we talk about the therapeutic treatment of serious pathologies it is difficult to consider the contribution of mathematics and statistics to the success of the interventions. Most often it is thought that positive results correspond to the abilities and knowledge of the luminaries of surgery and medicine. This article aims to provide additional information: to demonstrate that applied mathematics (in particular statistics) offers indispensable tools for a rational approach to these therapies. The method we used in the development of this therapeutic process is essentially deterministic, although some passages implicitly provide a probabilistic reference; in particular, when the least squares principle is applied for the research of the theoretical model of interpolation. The basic hypothesis is that the deviations of the experimental values from the theoretical values of the model have a Normal distribution.

2. Mathematics as a measure of the world

The field in which Mathematics moves has become vast. Usually, it is divided into two major sectors: the pure and that applied mathematics. The first sector has a purely speculative nature and is concerned with a rigorous arrangement of the basic principles of the discipline; the second, instead, relates to the applications of mathematical methods to Natural Sciences, Medicine, Engineering and Economics. It is in this second sector that interesting applications can be found that can help man solve several technical-scientific problems. It is necessary, however, to warn this is only an exemplifying division. Actually, mathematics is a unitary whole and it is difficult to know where its theoretical part ends and its experimental soul begins and vice versa. Often, problems arise in an application environment that requires in-depth theoretical analysis. So, it is necessary to refer to an experience, to a useful operational path.

A wider approach, not only descriptive, to natural phenomena requires a considerable knowledge of the mathematics that allows:

- Their measurement (Analysis, Probability Calculus, Statistics);
- The study of their possible forms (Analysis, Geometry, Statistics);
- The coherent arrangement of the rules followed (Logic, Algebra).

All scientific methodologies require compliance with these three points.

3. Problem analysis

Biology is one of the sciences that is proving to be very ductile to use mathematical techniques for a rational response to problems. It enables, with

genetics good practices and good procedures to improve the lives of human beings. The mathematical fields that can be applied to Biology range from Combinatorial Calculus to Probability Calculus, to Geometry, to Statistics and they offer a vast set of procedures.

The problem I am presenting is, certainly, of undoubted effect. It is an efficient and effective treatment to counteract, and eventually block, the progress of a particular type of tumor: the glioblastoma. It is a nodular tumor that lurks in the brain tissues and soon leads to the death of the host (the patient). We start from an experimental model of the tumor nodule, which, growing in the laboratory, gives us a lot of biological and kinetic measures of its growth (Figure 1). In particular, we can determine the growth time, the number of the cells for each instant of time and the critical limit of their growth beyond which there is nothing left to do (for example, for the compression of the tissues or for metastasis). In the dynamics of the tumor, we also consider the necrosis of many of its cells for the lack of food and of oxygen. It is also necessary to know the clinical picture of the patient and his immune response.

After that, we analyze the mathematical models able to guarantee a rigorous control of the behavior of this type of tumor.

4. The choice of mathematical models

On the basis of what we previously analyzed, the process requires the selection of mathematical models, as the first approach, in order to quantitatively describe the natural growth of the tumor mass over time and to find a mathematical model that allows to give to the patient a therapy that increases his life expectancy compared to the natural one, starting from the observation of the neoplasm.

The mathematical models able to control the growth of biological populations are studied by that part of mathematics that is known as population dynamics [8]. When dealing with a problem of growth of biological populations, we take on known and tested standard models. Usually, any changes to be made to the models are arranged during the work, keeping the standard model used as fixed as possible. One of the most well-known growth models is that of Verhulst [8]. In our case, however, the Verhulst equation does not adapt well to describe the growth dynamics of the glioblastoma tumor cells. It has been observed, from previous studies, that the most suitable model to describe this growth is given by the differential equation of B. Gompertz.

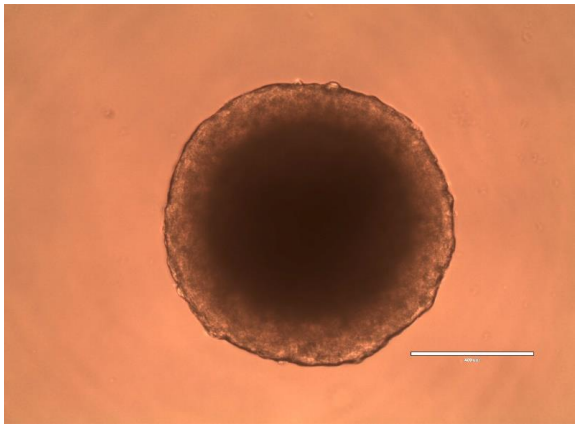

Figure 1. Photomicrograph of an experimental tumor nodule (tumor spheroid). The reference bar is 400 μm long. The central area of the nodule, darker and denser, is mainly formed by dead cells because of the poor availability of oxygen and the accumulation of toxic substances produced by the cells themselves with their metabolism, due to problems related to the diffusion of these molecules in the tissue. This area is generally referred to as the necrotic heart. Photo courtesy of Dr Roberto Chignola, Department of Biotechnology, University of Verona.

5. Gompertz model and tumor growth

This model can be expressed as a system of differential equations

$$\begin{cases} \dfrac{dX(t)}{dt} = kp(t) \cdot X(t) \\ \dfrac{dkp(t)}{dt} = -\beta \cdot kp(t) \end{cases} \tag{1}$$

or as a differential equation that includes both equations (2).

$$\frac{dX(t)}{dt} = \beta \cdot X(t) \cdot Log\left[\left(\frac{X(t)}{K}\right)^{-1}\right] \tag{2}$$

Model (2) derives from (1), as can be demonstrated.

We now present the parameters and variables of models (1) and (2). $X(t)$ is the number of tumor cells at time t; K is the carrying capacity of the environment in which the tumor cells live and is equal to $K = \mathrm{Max}(X(t))$: it represents the critical limit beyond which a tumor mass cannot go (otherwise would kill the host); $X(t) / K$ is the occupancy rate of the environment; $kp(t)$ is the time-dependent growth rate of the tumor cell population; β is a parameter that dampens the genetic growth of the population of individuals considered.

The differential equation (2) admits an integral curve in a closed form. It is given by:

$$X(t) = K \cdot e^{-C \cdot e^{-\beta \cdot t}} . \tag{3}$$

As shown, (3) depends on the parameters K, β, C.

The tumor has a mass whose volume is estimated on an experimental basis as follows:

$$Vol(t) = \frac{4}{3} \cdot \pi \cdot r_0^3 \qquad , \qquad r_0 = \frac{1}{2} \cdot \sqrt{d_{min} \cdot d_{max}} \ , \tag{4}$$

where $Vol(t)$ is the volume of the tumor mass at time t, r_0 is the geometric mean of the two rays $d_{min}/2$ and $d_{max}/2$, where d_{min} and d_{max} are the minimum and the maximum of the diameters of the spheroid. Once the volume is known, taking into account that a tumor cell has a known size (usually estimated in $10^{-9} cm^3$), one can determine the number of cells in the nodule in the following way:

$$X(t) = Vol(t)/Vol_{cellula}. \tag{4bis}$$

$X(t)$ of (4bis) is a very large value and therefore not very useful for calculations. Since the volume of a cell is known and is constant, the size of the population of tumor cells is conveniently replaced by the volume of the tumor mass $Vol(t)$. Starting from this substitution, $X(t)$ becomes $Vol(t)$ and, considering the multiplicative constant ($1 / Vol_{cellula}$), is also the population numerousness.

It is now necessary to estimate the parameters of the model (3).

6. Discretization and parameter estimation

The inevitable step to estimate the parameters of the model (2) or (3) with the least squares method is the discretization of the model. In practice, it consists to replacing the derivative with the incremental ratio and with the application of the finite difference operator first. Let $\Delta X_t = X_{t+1} - X_t$, from (2) we obtain:

$$\frac{\Delta X_t}{\Delta t} = \beta \cdot X_t \cdot Log\left[\left(\frac{X_t}{K}\right)^{-1}\right], \tag{5}$$

where $\Delta t = 1$. With easy algebraic steps, we get to:

$$\frac{X_{t+1}}{X_t} = 1 + \beta \cdot Log\left(\frac{K}{X_t}\right). \tag{6}$$

Equation (6) can be set in the following way:

$$\hat{Y}_t = A + B \cdot Log(X_t) , \tag{7}$$

where $\hat{Y}_t = X_{t+1}/X_t$, $A = 1 + \beta \cdot Log(K)$, $B = -\beta$.

Equation (7) is a linear model in the parameters. Thus we can apply the least squares method to estimate parameters *A* and *B* based on the experimental data in our possession. We obtain:

$$S(A,B) = \sum_{j=1}^{n} \left(Y_j - \hat{Y}_j\right)^2 = \sum_{j=1}^{n} \left(Y_j - A - B \cdot Log(X_t)\right)^2. \qquad (8)$$

Passing to the partial derivatives with respect to *A* and to *B*, setting them equal to zero and solving the system, we have:

$$\begin{pmatrix} n & \sum_{j=1}^{n} Log(X_j) \\ \sum_{j=1}^{n} Log(X_j) & \sum_{j=1}^{n} [Log(X_j)]^2 \end{pmatrix} \cdot \begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{n} Y_j \\ \sum_{j=1}^{n} Y_j \cdot Log(X_j) \end{pmatrix}. \qquad (9)$$

In this case, it is not necessary to proceed to the calculation of the second derivatives since the Hessian is a positive semidefinite matrix and therefore the solutions of the system (9) give precisely the minimum of *S*(*A*, *B*) [9].

Once we have found the values for *A* and *B*, β and *K* are easily obtained. It is then calculated X_t. For the calculation of the constant *C* in (3), the initial condition is taken into account: at time $t = 0$ we have $X(0) = K \cdot e^{-C}$, and hence we get $C = LogK - LogX(0)$.

7. Processing

To verify the validity of the method presented above, one uses the experimental measurements daily obtained with glioblastoma tumor nodules grown in laboratory (spheroids). The measures are relative to the variations of nodular size, taken for 77 days. We start, therefore, from the set *W* of the experimental data, where the first term of each pair represents the discrete time expressed in days of each observation and the second the volume of the tumor mass expressed in mm^3:

W= {{0, 3.57}, {1, 7.37}, {2, 10.9025},{3, 14.435},{4, 21.5},{5, 28.6},
{6, 37.14}, {7, 41.98}, {8, 52.89}, {9, 57.805},{10, 62.72},{11, 72.55},
{12, 88}, {13, 105.6}, {14, 96.5}, {15, 105.6},{16, 116.05},{17, 126.5},
{18, 147.4}, {19, 147.4}, {20, 185.2},{21, 172},{22, 199},{23, 199},
{24, 199}, {25, 199}, {26, 213.6}, {27, 199},{28, 199},{29, 199},{30, 199},
{31, 199}, {32, 199}, {33, 213.6}, {34, 199},{35, 213.6}, {36, 206.5},
{37, 199.4}, {38, 193}, {39, 185.2}, {40, 199}, {41, 199}, {42, 213.6},

{43, 213.6}, {44, 213.6}, {45, 213.6}, {46, 213.6}, {47, 185.2}, {48, 213.6}, {49, 199}, {50, 213.6}, {51, 209.95}, {52, 206.3}, {53, 199}, {54, 199}, {55, 213.6}, {56, 199}, {57, 199}, {58, 199}, {59, 199}, {60, 199}, {61, 213.6}, {62, 185.2}, {63, 185.2}, {64, 185.2}, {65, 185.2}, {66, 185.2}, {67, 213.6}, {68, 213.6}, {69, 199}, {70, 213.6}, {71, 203.2}, {72, 192.8}, {73, 172}, {74, 199}, {75, 185.2},{76, 199}, {77, 199}}.

From (9) we get: A = 1.93967, $\beta \cong 0.18076$, $K \cong 180.991$ mm^3, $X_0 \cong 3.57$ mm^3, $C \cong 3.92588$.

It, therefore, turns out to be

$$\hat{X}(t) = 180.991 \cdot e^{-3.92588 \cdot e^{-0.18076 \cdot t}} . \qquad (10)$$

It is not linear and therefore the goodness of fit is measured by the following fit index (which is a particular coefficient of variation):

$$I_2 = \frac{1}{M(\hat{X})} \cdot \sqrt{\frac{\sum_{j=1}^{n} (X_j - \hat{X}_j)^2}{n}} , \qquad (11)$$

where X_j are the second terms of the data pairs W, $\hat{X}_j$ are the theoretical results of the application of (10), M is the average of the theoretical values $\hat{X}_j$ and n is the sample size.

In our case the value is $I_2 \cong 0.147582$.

The value of I_2 seems acceptable; moreover, given the difficulty of data collection, we can be satisfied with this approach even if, according to the international standard, a value lower than 0.1 should be recommended [10].

We now present the graph of the theoretical model and the distribution of experimental data around it (Figure 2).

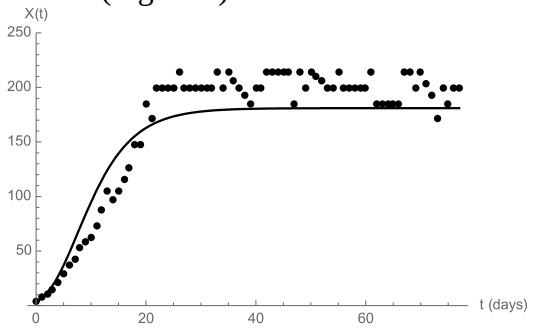

Figure 2. On the t-axis there is time in days, on the ordinates there is the volume of tumor.

Calculating the second derivative of (10) and placing it equal to zero, we obtain the inflection point [7]. It is equal to (7.56578 days, 66.5829 mm^3). We have thus finished studying the Gompertz model applied to our experimental

data. Let us now turn to the study of the optimal therapy to be applied to the nodule to control its growth.

8. The radiobiological treatment of tumor

The goal of radiological treatment of the cancer is to reduce its mass by killing its cells, without simultaneously damaging healthy cells. Radiotherapies aim to achieve this goal. This treatment, however, is rather dangerous since, in the irradiation of the tumor mass, healthy tissue cells are unfortunately also affected. In short, the following problem must be addressed: how much minimum radiant dose should be given to the patient to maximize the number of cancer cells killed with minimal damage to healthy cells? To answer this question, we need to address some preliminary aspects on the subject.

We have shown that the Gompertz model is valid in the interpretation of the dynamics of the tumor mass of an experimental nodule of glioblastoma. At this point we apply the model also to evaluate the dynamic behavior of the same tumor in a patient.

Before tackling the preliminaries, we consider that $X_0 = K \cdot e^{-c}$ and we put it in (3), obtaining the following formula (algebraic steps are simple and are omitted):

$$X(t) = X_0 \cdot e^{\frac{\alpha_0}{\beta}\cdot\left(1-e^{-\beta\cdot t}\right)}\ , \tag{12}$$

where $\alpha_0/\beta = C$, the parameter α_0 assumes the meaning of instantaneous spheroid growth rate at time t = 0 and β is a generic factor that deaden the tumour growth. From (12) it is confirmed that

$$Max[X(t)] = \lim_{t\to+\infty} X_0 \cdot e^{\frac{\alpha_0}{\beta}\cdot\left(1-e^{-\beta\cdot t}\right)} = X_0 \cdot e^{\frac{\alpha_0}{\beta}} = K. \tag{13}$$

Equation (13) represents a constraint on the growth of the spheroid. On the basis of a consolidated case series, it is believed that the maximum volume of the tumour borne by a patient can reach 25 cm^3, after which the effects are devastating and lead to the death of the guest in a short time. Then from (13) we have:

$$Log(K) = Log(X_0) + \frac{\alpha_0}{\beta}\ ,$$

$$\frac{\alpha_0}{\beta} = Log\left(\frac{K}{X_0}\right) \cong Log\left(\frac{25\ cm^3}{10^{-9} cm^3}\right) \cong 23.94, \tag{14}$$

where X_0 in this case corresponds to the volume in cm^3 of a tumor cell at the beginning of the process; that is $X_0 = Vol_{cellula}$.

9. Some notions of radiobiology

Often only possible therapy in the treatment of tumors is the radiotherapy, especially when the tumor involves important tissues of the human body or is located in places of difficult surgical access. From a clinical point of view, radiotherapy is an indispensable treatment even when it is considered necessary to intervene with more invasive therapies such as surgery and chemotherapy. Currently, biomedical research is further progressing with promising studies on the interaction between tumor cells and subatomic particles obtained with appropriate accelerators. At the moment encouraging results have been achieved, but the journey is still long. The treatment of tumor masses with radiation has the purpose of inducing massive molecular damage to the diseased cells so as to lead them to death. The decisive problem is to avoid as far as possible damage to healthy cells when one intervenes on sick cells. The damage induced by radiotherapy treatment depends on the intensity of the radiant dose. There are international indications that establish the effects of any radiation therapy. The radiant dose is expressed in Gray (Gy), which corresponds to the energy of 1 joule absorbed by 1 kg of biological tissue. Moreover, this basic unit must be multiplied by a suitable parameter that allows to take into account the effect on biological tissues of different nature of this radiant dose (RBE = Relative Biological Effectiveness). Finally, the product between Gy and RBE gives the equivalent biological dose to be administered, which is measured in Sievert (Sv). It should be considered that for radiations of clinical interest, radiation γ [4], we consider RBE = 1 and Gy = Sv. Table 1 highlights from a descriptive point of view the effects on human beings of exposure to radiant doses of different degrees of intensity [5].

Dose (Sv)	Effects
(0.05 - 0.2]	No symptoms, but risk of DNA mutations
(0.2 - 0.5]	Temporary drop in red blood cells
(0.5 - 1]	Drop in immune system cells and risk of infection
(1 - 2]	Immunodepression, nausea and vomiting. Mortality of 10% at 30 days from exposure
(2 - 3]	Severe immunodepression, nausea and vomiting 1-6 hours after exposure. Latency phase of 7-14 days after which symptoms appear such as hair loss. Mortality of 35% at 30 days from exposure
(3 – 4]	Bleeding of the mouth and urinary tract. Mortality of 50% at 30 days from exposure
(4 – 6]	Mortality of 60% at 30 days from exposure. Female infertility. The convalescence lasts from a few months to a year
(6 – 10]	Complete injury of the bone marrow (the organ that produces red blood cells and all cells of the immune system). Symptoms appear between 15 and 30 minutes after exposure and mortality is 100% at 14 days after exposure

(10 – 50]	Immediate nausea, bleeding from the gastrointestinal tract and diarrhea, coma and death within 7 days. No medical intervention is possible
(50 – 80]	Immediate coma. Death occurs in a few hours due to the collapse of the nervous system
> 80	Exposure to these doses occurred in two circumstances. Both subjects died within 49 hours of the accident

Table 1: Effects of radiation on human beings

10. The modeling of therapy

At this point, we must find a therapeutic process that allows us to stop the growth of the tumor or, even better, to reduce its mass to extinction. The model should take into account the disposition of the cells within the tumor mass, their microenvironment and the toxic effects induced on the healthy tissues of the surrounding cells and any other factor that may inform about the dynamics of the tumor. Studies conducted so far in various research institutes around the world have led to confirm, as an acceptable model to be considered in the treatment of tumors with radiant dose, the following one:

$$\widehat{SF}(D) = e^{-a\cdot D - b\cdot D^2}, \tag{15}$$

where $\widehat{SF}$ is the survival rate, a and b are two arbitrary parameters and D is the radiant dose. We must estimate the parameters a and b of the model as a function of the experimental data. Even in this case we linearize the model and apply the least squares method.

Dose (Gy)	SF		Dose (Gy)	SF
0.0000	1.0000		5.5036	0.19609
0.53957	0.87780		6.0072	0.18372
1.0072	0.84048		6.5108	0.14785
1.5468	0.73778		7.0144	0.11642
2.0144	0.78746		7.5180	0.097850
2.5180	0.62009		7.9856	0.073780
3.0216	0.55627		8.4892	0.058100
3.5252	0.46753		8.9928	0.043800
3.9928	0.36816		9.4964	0.036020
4.5324	0.33752		10.000	0.033750
5.0360	0.26007		10.504	0.026010

Table n. 2: Numerical data relating to the graph in figure 3, further on

11. Assumptions for the radiotherapy

When we face the problem of finding the relationship between a dynamic model of natural growth of a tumor and its radio-therapeutic treatment, collateral effects inevitably arise that create states other than those we would have liked to encounter. The complete modeling of a radiotherapy treatment requires the consideration of numerous variables that influence the interaction between tumor cells and radiant doses. For this reason, as a first approximation, we put some valid hypotheses to simplify the method. The choice of the hypotheses useful for the simplification of an effective model for the treatment of a tumor is in any case indispensable every time the control of the final results is desired. If we consider the analysis of the problem from a mathematical point of view, it is necessary to think about the implication of having to replace differential equations, defined in the continuous, with equivalent equations defined in the discrete. At this point we present the list of the necessary hypotheses to get on with the analysis of the process.

Assumption 1: The Gompertz model is a good representation of the growth dynamic of a tumor mass, starting from a first degenerated cell up to asymptotically reaching a volume of 25 cm^3. Thus, it is possible to simulate tumor growth using the equations (1), (2) and (3).

Assumption 2: A solid tumor, in general, consists of proliferating cells P, quiescent cells Q and dead cells U. The number of total cells N at time t is therefore given by

$$N(t) = P(t) + Q(t) + U(t). \tag{16}$$

Table 3 and Figure 5 refer only to proliferating cells since ionizing radiations are much less effective if directed against quiescent cells.

Assumption 3: In a solid tumor, on an experimental basis, it is possible to state that the number of quiescent and dead cells becomes significant with respect to the total of cells at the inflection point of the Gompertz curve (3) and (12).

Assumption 4: Radiation therapy has instantaneous effects, causing the immediate death of the cancer cells. These effects should at least be faster than the growth of tumor cells. This avoids a detailed kinetic analysis of the toxicity of radiation.

Assumption 5: After undergoing radiotherapy treatment, the tumor grows with the same dynamic modalities that preceded the treatment. It is a common convention in scientific treatises; however, there are also different points of view on this matter [6].

Assumption 6: The maximum dose in a single treatment is 3 Gy. You can also perform multiple treatments if and only if they are repeated at 24-hour intervals. It is not possible, however, to exceed 65 Gy. This assumption is

indicated by the radiotherapeutic protocols followed in the therapy of some tumors. The 3 Gy dose allows healthy tissues affected by radiation to recover from damage.

Assumption 7: We assume the existence of two critical thresholds in the treatment phase: 1) if after treatment a tumor falls below 1 mm^3, then we consider a therapy to be successful; 2) if, on the other hand, the volume increases beyond the dimension corresponding to the inflection point of the Gompertz curve, the therapy must be considered as failed. In practice, nothing justifies this assumption from a clinical or biological point of view and yet we accept it as work hypothesis.

Based on these hypotheses, we can proceed with the estimation of the model parameters (15) and with the application of the programmed therapy.

12. Procedure for a rational therapy

The method we will use for the treatment of glioblastoma, meets the following two objectives:

1) Check if there is a relationship between the effectiveness of the radiotherapy treatment used and the rate of tumor growth.

2) In case of an affirmative answer to the first objective, find a specific treatment protocol that allows to optimize the relationship between the benefits of the therapy and the costs due to the induction of toxic effects; in concrete terms, it is necessary to find the minimum amount of radiation to be used with the maximum destructive effect of cancer cells.

We start with the estimation of the parameters of the model (15) using the well-known method of least squares and, also in this case, evaluating the goodness of fit with index I_2 (11). The model (15) must be linearized:

$$Log\left[\widehat{SF}(D)\right] = -a \cdot D - b \cdot D^2 \tag{17}$$

and applying the least squares method we have:

$$S(a,b) = \sum_{j=1}^{n} \left(Log\left(SF(D_j)\right) - Log\left(\widehat{SF}(D_j)\right)\right)^2 = \\ = \sum_{j=1}^{n} \left(Log\left(SF(D_j)\right) + a \cdot D + b \cdot D^2 \right)^2.$$

Calculating the partial derivatives of $S(a, b)$ with refer to a and to b, we obtain the system

$$\begin{cases}\left(\sum_{j=1}^{n} D_j^2\right)\cdot a+\left(\sum_{j=1}^{n} D_j^3\right)\cdot b=-\sum_{j=1}^{n} D_j\cdot Log\left[SF(D_j)\right]\\ \left(\sum_{j=1}^{n} D_j^3\right)\cdot a+\left(\sum_{j=1}^{n} D_j^4\right)\cdot b=-\sum_{j=1}^{n} D_j^2\cdot Log\left[SF(D_j)\right]\end{cases} \tag{18}$$

Considering the data in Table 2 and solving (18) with refer to a and b we obtain:

$a \cong 0.124275;\ b \cong 0.0264028.$

The model adapted to the data in table 2 is, therefore:

$$\widehat{SF}(D) \cong e^{-0.124275\,D-0.0264028\,D^2} \tag{19}$$

Using the coefficient of variation:

$$I_{2,SF}=\frac{1}{M\left(\widehat{SF}(D)\right)}\cdot\sqrt{\frac{\sum_{j=1}^{n}\left(SF(D_j)-\widehat{SF}(D_j)\right)^2}{n}} \tag{20}$$

and taking into account both the data in table 2, and the theoretical values calculated with (19), we get the goodness of fit: $I_{2,SF} \cong 0.0998765$. This value shows that our approach is good. Figure 3 presents both the trend of experimental data and the interpolated model.

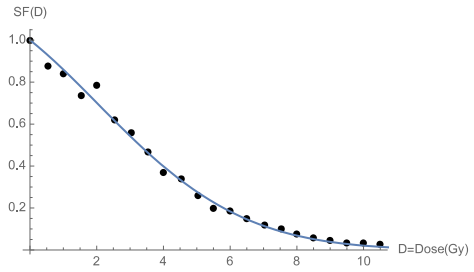

Figure 3: the graph shows the link between the radiant dose and the fraction of surviving individuals (19). The points represent, in Cartesian coordinates, the data of table 2. We put on the *D*-axis the radiant dose, on the ordinate axis the survival rate *SF*(*D*).

13. Research of the inflection point

At this point, it is necessary to start the therapy taking into account what has resulted from these preliminary procedures. We consider again the model (10) and figure 3. Furthermore, on the basis of the assumption 3, the most effective radiotherapy treatment is the one which begins at the inflection point of the Gompertz curve.

We consider the model (12):

$$\hat{X}(t) = X_0 \cdot e^{\frac{\alpha_0}{\beta}\cdot\left(1-e^{-\beta\cdot t}\right)} \quad , \tag{21}$$

and taking into account that a cancer cell has a volume of 10^{-9} cm^3 and that $\beta = 0.016$ we have:

$$\hat{X}(t) = 10^{-9} \cdot e^{23.9421\left(1-e^{-0.016\,t}\right)} \quad . \tag{22}$$

Calculating the second derivative of $X(t)$ and setting it equal to zero we get the inflection point (198.477 days, 9.19654 cm^3) [7]. We note that this result is different from that obtained using the model (10). Here, in fact, the parameters of the Gompertz model are changed, which are now imposed not by the experimental data of the single experimental nodule (which in our case led to the model (10)), but by a different operating standard that requires both a start from a single tumor cell, whose volume is fixed at 10^{-9} cm^3, and from a critical maximum limit of tumor expansion equal to $K \cong 25$ cm^3. Figure 4 presents the function with the flex point.

At this point the radiant doses should be applied at intervals that allow the patient's average life to be maximized. The first simulation (Fig. 4) considers a single-dose therapy to hit the tumor mass with a single dose of radiation (from 1 to 3 Gy with intervals of 0.4). Starting at the time of the cancer diagnosis observation, when the tumor mass can vary from a minimum of 0.0050 cm^3 to a maximum marked by the flex point, we have to measure the effect of the therapy on the cancer using the delay time of its growth. This time corresponds to the one that the tumor mass needs, after having been treated with radiotherapy, to return to the mass it had before the treatment was carried out. Methods and procedures are reported in [2]. In the last two graphs we report two other simulations in which, with respect to the protocol for the search for an optimal result, two different outcomes are observed. In figure 5, the result is not satisfactory; instead, in figure 6 the protocol gives a favorable outcome and the mass of the glioblastoma is reduced below the desired minimum threshold.

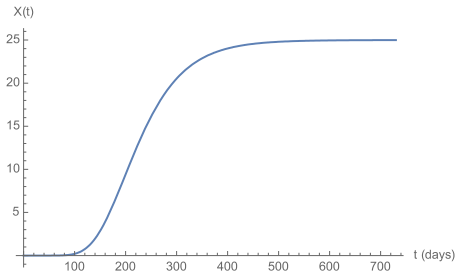

Figure 4. Gompertz curve (22) related to the investigated tumor. The inflection point is (198.477 days, 9.19654 cm^3). On the *t*-axis there is the time in *days* and on the ordinate axis the tumor volume in cm^3.

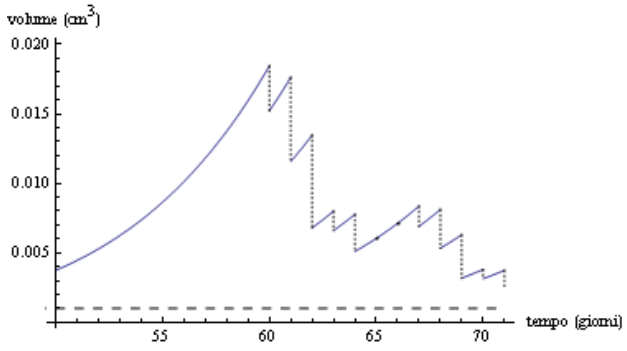

Figure 5. The graph describes the effect of 12 radiotherapy treatments on a glioblastoma. The doses, in Gray, is (1, 2, 3, 1, 2, 0, 0, 1, 2, 3, 1, 2). Note that the treatment did not give the desired result. The mass of the tumor has not been reduced below the critical threshold set by the protocol.

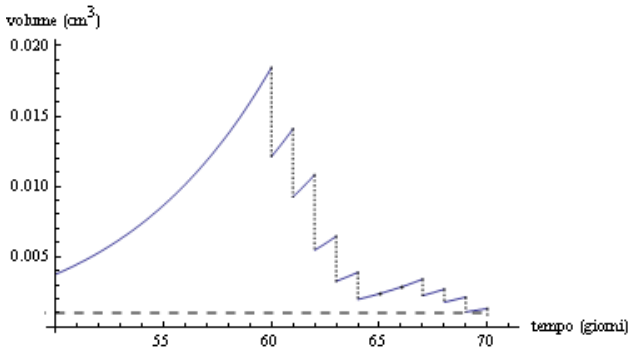

Figure 6. The graph describes the effect of 12 radiotherapy treatments on a glioblastoma. The doses, in Gray, is (2, 2, 3, 3, 3, 0, 0, 2, 2, 3, 3, 3). In this case, it should be noted that the treatment has reached the desired result. The tumor mass was reduced under the critical threshold established by the protocol.

Each cusp corresponds to the flex point of the various curves that sequentially describe the progression of tumor growth after each treatment.

References

[1] Chignola R, Castelli F., Corso L., Pezzo G., Zuccher S., La biomatematica in un problema di oncologia sperimentale, I Quaderni del Marconi, ITIS G. Marconi di Verona, anno 2006, ISBN 88-902125-9-4. Piano lauree scientifiche UNIVR

[2] Burato A., Chignola R., Castelli F., Corso L., Pezzo G., Zuccher S., Matematica e radioterapia dei tumori, Sviluppo e applicazioni di un modello predittivo semplificato, I Quaderni del Marconi, ITIS G. Marconi di Verona, anno 2007, ISBN 978-88-95539-01-0. Piano lauree scientifiche UNIVR.

[4] http://en.wikipedia.org/wiki/Gamma_ray

[5] http://en.wikipedia.org7wiki/Radiation_poisoning

[6] Guirardo D. et al., Dose dependence of growth rate of multicellular tumour spheroids after irradiation, The British Journal of Radiology, vol. 76, 2003, pp. 109-116.

[7] Mathematica 5. 1, Wolfram research – www.wolfram.com.

[8] Smith J. M., *L'ecologia e i suoi modelli*, A. Mondadori editore, Milano, 1975, ISBN 0303-2752.

[9] Apostol Tom M., *Calcolo*, Vol. 1 e 3, Boringhieri ed, Torino, 1985, ISBN 88-339-5033-6.

[10] Gambotto Manzone A. M., Susara Longo C., Probabilità e statistica 2, ed. Tramontana, Milano, 2010. ISBN 978-88-23322-98-1.

Ratio Mathematica Volume 36, 2019, pp. 69-78

En Route for the Calculus of Variations

Jan Coufal[*]
Jiří Tobíšek[†]

Abstract

Optimal control deals with the problem of finding a control law for a given system such that a certain optimality criterion is achieved. An optimal control is an extension of the calculus of variations. It is a mathematical optimization method for deriving control policies. The calculus of variations is concerned with the extrema of functionals. The different approaches tried out in its solution may be considered, in a more or less direct way, as the starting point for new theories. While the true "mathematical" demonstration involves what we now call the calculus of variations, a theory for which Euler and then Lagrange established the foundations, the solution which Johann Bernoulli originally produced, obtained with the help analogy with the law of refraction on optics, was empirical. A similar analogy between optics and mechanics reappears when Hamilton applied the principle of least action in mechanics which Maupertuis justified in the first instance, on the basis of the laws of optics.

Keywords: Calculus of variations, Johann Bernoulli, Jacob Bernoulli, Euler, Lagrange, Maupertuis, principle of least action.

2010 AMS subject classification: 01A50, 49K50[‡]

[*] University of Economics and Management, Nárožní 2600/9a, 158 00 Praha 5, Czech Republic, e-mail: jan.n.coufal@seznam.cz.
[†] University of Economics and Management, Nárožní 2600/9a, 158 00 Praha 5, Czech Republic, e-mail: jiri.tobisek@vsem.cz.
[‡]Received on May 4th, 2019. Accepted on May 27th, 2019. Published on June 30th, 2019. doi: 10.23755/rm.v36i1.467. ISSN: 1592-7415. eISSN: 2282-8214.

1 Introduction

Our intention here is to write the history of the brachistrone and its remarkable consequences. In the contemporary socio-cultural context, the question would essentially be formulated in the following text: what shape should we make slides in children's playgrounds so that the time of descent should be minimized? The considerable importance of this question is well understood when we consider how children behave, and they want to obtain the best performance, but the question is also important in a more general way, and a great number of scholars have attempted to solve this problem.

Unfortunately the problem appears to be particularly tricky, and it depends upon a number of parameters, including the variable value of the friction between the clothes of the child and the surface of the slide. We shall not attempt to solve that particular problem here, but content ourselves with theory of the idealized problem, simplifying the situation sufficiently in order to be able to find a solution. In fact we shall replace the child by a perfectly smooth marble, and we assume that it rolls down a smooth surface, thus assuming that friction forces are negligible with respect to gravity.

Now, we are simply confronted with the *problem of brachistrone* as Johann Bernoulli expressed it in the *Acta Eruditorum* published in Leipzig in June 1696 ([1], vol. 1, p. 161): *Datis in plano vertacali duobus punctis A & B, assignare Mobili M viam AMB, per quam gravitate sua descenden, & moveri incipiens a puncto A, brevissimo tempore perveniat ad alterum punctum B.*

The expression *brevissimo tempore* is the latin translation of the greek term *brachistochrone* (brachys is brief, brachisto is quickest, chronos is time and brachistochrone is the shortes time). In a modern style: Given two points *A* and *B* in a vertical plane, what is the curve traced out by a point subject only to the force gravity, starting from rest at *A,* such that it arrives at *B* in the shortest time?

Common sense suggests that this curve is necessarily situated in the vertical plane containing the points *A* and *B.* Common sense also leads us to think that the quickest route is the shortest, and is given by the line segment joining the points *A* and *B.* But this is not the case. We know, for example that a longer journey on a motorway be faster than going a shorter distance on an ordinary road. Here, in order to try to solve the problem of brachistochrome, it is necessary to consider all the curves joining points *A* and *B* and compare all the corresponding times of travel. Taking everything into account, even under these restrictions, the problem turns out to be a subtle one. The brachistochrone problem, a priori a simple game for mathematicians, turns out in the end to be a considerable problem.

2 Falling bodies, reflection and refraction

In 1638, well before the problem had been explicitly stated, Galileo gave his solution to the brachistochrone problem in the course of the Third Day of his [5]. It is here that he studied uniform acceleration – Galileo called it "natural acceleration" – comparing it with uniform motion, and showed that a body falling in space traverses a distance proportional to the square of the time of descent (Theorem II in [4]). With regard to bodies moving on inclined planes he deduced ([5]):

Theorem V. The times of descent along planes of different length, slope and height bear to one another a ratio which is equal to the product of the ratio of the lengths by square root of inverse ratio of their heights.

We interpret the proportionality to be: a body travels a distance L and descends a height H in time t such that:

$$t = \frac{k \cdot L}{\sqrt{H}}.$$

Galileo then proves the following neat result ([5]):

Theorem VI. If from the highest or lowest point in a vertical circle there be drawn any inclined planes meeting the circumference, the times of descent along these chords are each equal to the other.

At the end of the Third Day, Galileo shows that it is also possible to improve on this descent ([5]):

Theorem XXII. If from the lowest point of a vertical circle, a chord is drawn subtending an arc not greater than a quadrant, and if from the two ends of this chord two other chords be drawn to any point on the arc, the time of descent along the two later chords will be shorter than along the first, and shorter also, by the same amount, than along the lower of these two latter chords.

This result is false, since arguing the case from two to three segments is based on a faulty intuition from arguing from one to two segments. The brachistrochrone problem is considerably more subtle than the one of the research into optimum inclination of planes, which is a simple problem of the extremum for a function of single variable.

The demonstration by Johann Bernoulli [1] also derives from an intuitive approach. This approach, an analogy with the law of refraction, leads to the curve solution which one cannot find *a priori,* without an arsenal of sufficiently sophisticated techniques. Let us begin by recalling the first laws of Optics, which are in fact consequences of the principles of optimization.

Experience tells us that light travels in straight lines. This phenomenon is stated as a principle: light chooses the shortest path. This formulation led to a real theoretical advance since it allowed Hero of Alexandria in the first century AD to explain the law of reflection, namely, the equality of the angles of incidence and reflection. In the case of reflection, the speed remains constant. It

is not so for refraction, where the speed of light $\frac{c}{n}$ varies as a function of the index n of the medium traversed. However, the principle stated above could have been stated in the following form as the Fermat's Principle: light chooses the fastest route, which in a homogenous medium where its speed is constant, is equivalent to the previous principle.

So, to go from A to B, passing from a medium of index n_1 to medium of index n_2, the trajectory of the light will not be the line segment AB, but broken line AIB such that the trajectory AIB will have the shortest time of all trajectories from A to B. Using the initial conditions we calculate that the angle of incidence i and the angle of refraction r are related to the respective speeds by the formula:

$$\frac{\sin i}{v_i} = \frac{\sin r}{v_r}, \tag{1}$$

or using the indices n_i and n_r we have the sine formula

$$n_i \cdot \sin i = n_r \cdot \sin r.$$

This formula, discovered by the Dutch scientist Snell in 1621, received its correct interpretation with Fermat. In a letter of the 1st of January 1662 to M De la Chambre, Fermat explains ([4], vol. II, pp. 457-463): *As i said in my previous letter, M. Descartes has never demonstrated his principle; because not only do the comparisons hardly serve as a foundation for the demonstrations, but he uses them in the opposite sense and supposes that the passage of light is more easy in dense bodies than in rare bodies, which is clearly false. I will not say anything to you about the shortcomings of the demonstration itself ...*

Fermat puts his principle to work, and proves the sine formula using his method 'de maximis et minimis' ([4]). Another example of a non-homogeneous medium where the shortest trajectory is not the quickest occurs in mechanics, where the effect of gravity is in the vertical direction. And this is the context for Johann Bernoulli brachistochrone problem. Johann Bernoulli in the *Acta Eruditorum* of May 1697 ([1], vol. 1, pp. 187-193). His method typically corresponds to what we now call a discretisation of the problem. He imagines space carved into small lamina, sufficiently fine so that within each one it is possible to imagine that the speed is constant. Within each strip the trajectory becomes the shortest route, and necessarily a segment. The complete trajectory appears as a sequence of segments. But how we move from one strip to another? We must always optimize the time of travel. As in refraction of light, this is done by using Fermat's principle. Thus, if v_i is the speed in a given band and v_r in the band immediately below, the angle i is the angle made with the vertical by segment of the trajectory in the first band, an the angle r in the neighboring band, then they are connected by the rule of sines (1). If we now imagine that the horizontal strips become progressively thinner, and their number increases indefinitely, the line of segments tends towards a curve. The tangents at each

point of this curve approach the sequence of segments. The angle u which the tangent makes with the vertical is then connected to the speed v by the relation:

$$\frac{\sin u}{v} = const.$$

Here, the speed v of a particle is known; it is result of the action of gravity and, as we know from Galileo, it is a function of the distance fallen y, according to the formula

$$v = \sqrt{2gy}.$$

And so the rule of sines leads to the equation:

$$\frac{\sin u}{\sqrt{y}} = const.$$

In particular, for $y = 0$, the tangent is vertical.

That is a characteristic equation of a well-known curve of the time, the cycloid.

We have just seen that the solution to the curve is a cycloid. But how can we construct such a curve, starting from a point *A,* an arriving exactly at a point *B*? Newton gave a simple solution in a letter to Montague on the 30th of January 1697 (see [10], p. 223). In addition to Newton's contribution to the solution of the problem of the brachistochrone, we must also mention Leibniz, and in a lesser role, the Marquis de l'Hospital, and most of all, Jacob Bernoulli, the older brother of Johann ([1], vol. 1, p. 194-204): … *my elder brother made up the fourth of these, that the three great nations, Germany, England, France, have given us each one of their own to unite with myself in such a beautiful search, all finding the same truth.*

The method used by Jacob Bernoulli is laborious, but quite general. Also, Jacob, in wanting to show the singular character of Johann's method, extended the problem by posing new questions. Indeed, Johann's method, founded on an analogy, does not work except in a particular case, and cannot be used for more general problems of this type. In particular, Jacob Bernoulli put the following question to his brother" given a vertical line which of all the cycloids having the same starting point and the same horizontal base, is the one which will allow a heavy body passing along it to arrive at the vertical line the soonest? Such statement reminds us of Calileo's first version, which was about finding the inclined plane through a given point which gave the shortest time to reach a given vertical. Johann Bernoulli ([1], vol. 1, p. 206-213) replied and showed that the cycloid in question is the one which meets the given line horizontally. More generally, the cycloid which allows us to achieve the swiftest possible descent to a given oblique line is the one which meets the line at right angles. This cycloid which, as we have just said, is a brachistochrone curve, was also known to Huygens fro 1659 as the tautochrone curve: bodies which fall in an inverted cycloid arrive at the bottom at the same time, no matter from what height they are released. This property was perhaps closer to that observed by Galileo: the

equality of the times for the distance on the chords of the same circle. Among the other problems posed by Jacob Bernoulli to Johann are those which are called isoperimetric problems, which together with brachistochrone problem are prototypes of optimization problems. These scientific exchanges between the two brothers were carried out in the form of letters. Here is a sample of Johann's response to same criticisms by Jacob ([1], vol. 1, p. 194-204): *So there it is, his imagination, stronger and more vivid than those claiming to be sorcerers who believe they have found themselves bodily present at a Sabbath, has seduced him; he is carried along by a torrent of vain conjectures; in a word, he is longer ready to give reign to reason ...* The resolution of these problems is then the object – reason or excuse? – for a long dispute between the two brothers; a dispute which developed into a major row, but which gave birth to new area in mathematics, the Calculus of Variations.

3 The Calculus of Variations

When we look for boundary values of a function *f* of a variable *x,* i.e. when we look for values of the variable *x* for which the value $f(x)$ is a maximum or minimum, we look for the points where the graph of f has a horizontal tangent, or we say we look for the values where $f'(x) = 0$. In the case of a function *f* of two variables *x* and *y,* we have to consider the points where the tangent plane is horizontal to the surface which has the equation $z = f(x, y)$. Alternatively we could say we seek the number pairs [*x*, *y*] for which

$$\frac{\partial f}{\partial x}(x, y) = \frac{\partial f}{\partial y}(x, y) = 0.$$

Or we can say we are looking for the points where the function f has a stationary value. In the case of a finite number of variables, the difficulties seem surmountable, and the approach to the problem may be effected with the aid of the differential calculus of Newton and Leibniz. Here the object which changes is not a number or a point, but a curve, a function, and the corresponding quantity to maximize or minimize is a number depending on this curve or on this function. It is necessary to conceive an extension of the differential calculus. The new theory which was created is called the calculus of variations, the variations being those of the function. But, in 1696, this theory had not been formulated and our problem becomes *a priori* somewhat subtle. A problem in the calculus of variations can be presented generally in the following fashion: we try to find a curve, being the graphical representation of a function *y* of *x*, which minimizes or maximizes a certain quantity among all the curves

constrained by certain conditions[§]. The quantity whose extreme value has to be found[**] is expressed generally in the form of an integral:

$$I(y) = \int_a^b F(x, y, y')\, dx$$

where y represented the unknown function, y' its derivative, x variable and F a particular function.

Among the typical problems of the calculus of variations, besides the isoperimetric problems above are investigations of the geodesic lines on surface, i.e. the curves of minimum length joining two points of a surface. Also, the investigation of the shapes of the surfaces of revolution which offer the least resistance to movement, a problem which Newton tackled in 1687 in the *Principia.* The statement of the brachistochrone problem in 1696 could be considered as the definitive origin of the calculus of variations, for it is the problem which generated general methods of investigation which were gradually developed in a competitive context.

Johann Bernoulli himself posed the problem of geodetics to Euler. Euler reworked the ideas of Jacob Bernoulli, simplified them, and finally was the first to formulate the general methods which allowed them to be applied to the principal problems of the calculus variation. He developed these ideas systematically in 1744 in [3]. In a way like Jacob Bernoulli, Euler tackles the problem as a problem of limits in an investigation of the ordinary extremum. Euler derived the differential equation:

$$\frac{\partial F}{\partial y}(x, y, y') - \frac{d}{dx}\left(\frac{\partial F}{\partial y'}(x, y, y')\right) = 0 \qquad (2)$$

which satisfies each solution y. It is only a necessary condition and the method does not establish the existence of a solution. The equation (2), today called the Euler-Lagrange equation, is a second order differential equation in y:

$$\frac{\partial F}{\partial y}(x, y, y') - \frac{\partial^2 F}{\partial y' \partial x}(x, y, y') - \frac{\partial^2 F}{\partial y' \partial y}(x, y, y') - \frac{\partial^2 F}{\partial y'^2}(x, y, y') = 0.$$

In 1760, Lagrange greatly simplified matters by introducing the differential symbol δ, specifically for the calculus of variations, corresponding to a variation of the complete function. He makes the point of it in the introduction to [6]: *For as little as we know the principles of the differential calculus, we know the method for determining the largest and smallest ordinates of curves; but there are questions of maxima and minima at a higher level which, although depending on the same method, are not able to be applied so easily. They are*

[§] For brachistochrone problem – the curve joining two points A and B.
[**] Here – the time of the journey.

those where it is needed to find the curves themselves, in which a given integral expression becomes a maximum or minimum with respect to all the other curves. ... Now here is a method which only requires a straightforward use of the principles of the differential and integral calculus; but above all I must give warning that while this method requires that the same quantities vary in two different ways, in order not to mix up these variations, I have introduced into my calculations a new symbol δ. In this way, δZ expressed a difference of Z which is not the same as dZ, but which, however, will be formed by the same rules; such that where we have for any equation dZ=m dx, we can equally have δZ=m δx, and likewise for other cases.

A century later, Mach was able to write in [7]: *In this way, by analogy, Johann Bernoulli accidentally found a solution to the problem. Jacob Bernoulli developed a geometric method for the solution of analogous problems In one stroke, Euler generalized the problem and the geometrical method, Lagrange finally freed it completely from the consideration of diagrams, and provided an analytical method.*

4 The Principle of Least Action

We shall make a digression, the purpose of which will soon become clear Maupertuis stated his Principle of Least Action in 1744 in [8]. He explains and justifies his principle from the law of refraction: *In thinking deeply upon this matter, I reflected that light, as it passes from one medium to another, yet not taking the shortest path, which is a straight line, might just as well not take the shortest time. Actually, why should there be a preference here for time over space? Light cannot go at the same time by the shortest path and by the quickest route, so why does it go by one route rather than another? In fact, it does not take either of these; it takes a route that has the greater real advantage: the path taken is the one where the quantity of action is the least.*

Now I must explain what I mean by the quantity of action. When a body is moved from one place to another, a certain action is needed: this action depends neither on the speed of the body and the distance travelled; but it depends on the speed nor the distance taken separately. The quantity of action is moreover greater when the speed of the body is greater and when the path travelled is greater; it is proportional to the sum of the distance multiplied respectively by the speed travelled over each space. ... It is quantity of action which is the true expenditure of Nature, and which she uses as sparingly as possible in the motion of light. Let there be two different media, separated by a surface represented by the line CD, such that the speed of light in the medium above is m. and the speed in the medium below is n.

Let a ray of light, starting from point A, reach a point B: to find the point R where the ray changes course, we look for the point where if the ray bends the

quantity of action is the least: and I have $m \cdot AR = n \cdot RB$ *which must be a minimum. ...*

That is to say, the sine of the angle of incidence to the sine of the angle of refraction is in inverse proportion to the speed with the light traverses each medium.

All the phenomena of refraction now agree with the central principle that Nature, in the production of its effects, always tends towards the most simple means. So this principle follows, that when light passes from one medium to another the sine of the angle of refraction to the sine of the angle of incidence is in inverse ratio to the speed with which the light traverses each medium.

And so for Maupertuis, light is propagated so as to minimize $AR \cdot v_1 = RB \cdot v_2$ and not the quantity $\frac{AR}{v_1} = \frac{RB}{v_2}$. For these conclusions to agree with the experimental results of the time, and so that his principle would lead to the sine law. It is true that at that time no one knew how to measure the speed of light and no one could find a way of deciding between the different theories. The experimental proof that light travels faster in air than in water was not established until 1850 Foucault.

In 1746, Maupertuis extended his principle from optics to mechanics ([9]): *When a body is carried from one place to another, the action is greater when the mass is heavier, when the speed is faster, when the distance over which it is carried is longer. ... Whenever a change in Nature takes place, the quantity of action necessary for this change is the smallest possible.*

With this general principle, Maupertuis established a kind of union between philosophy, physics and mathematics: Nature works in such a way as to minimize its action; the idea of causality is abandoned in favor of the idea achieving an aim, characterized by a harmony between the physical world and rational thought.

5 Conclusion

It would be right to conclude by revisiting our initial problem of the slides in the playground. We are circumspect, and content ourselves with noticing that in the course of this wander through diverse disciplines, the theme of minimization or maximization briefly the problem of optimalization is ever present, and should not be underestimated during these unhappy times.

Acknowledgements

This contribution is a follow-up to the project of the Centre of Economic Studies of University of Economics and Management

References

[1] Bernoulli, J. *Opera Omnia.* Lausanne and Geneva, 1742.

[2] Chabert, J.-L. The Brachistrone Problem. History of Mathematics, Histories of Problems, Elipse, Paris, 1997, pp. 183-202.

[3] Euler, L. Methodus inveniendi lineas curvas maximi minimive proprietate gaudentes: sive solution problematic isoperimetrici lattissimo sensu accepti, Lausanne and Geneva, 1774 (Œvres, vol. 24, Berne, Orel Füssli, 1952).

[4] Fermat, P. Œvres, ed. Tannery, P. and Henry, C., Gauthier-Villars, Paris, 1894.

[5] Galileo, G. Discorsi e dimostrazioni matematiche intorno a duo nuove scienze, Leyden, 1638.

[6] Lagrange, J.-L. Essai d'une nouvelle méthode pour déterminer les maxima et les minima des formules integrals defines. Miscellanea Taurinensia, vol II, 1760-1761 (Œvres, vol. I, Paris, pp. 333-468).

[7] Mach, E. Die Mechanik in ihrer Entwickelung : historisch-kritisch dargestellt, F. A. Brockhaus, Leipzig, 1883.

[8] Maupertius, P. L. M. de. Accord de différentes lois de la nature qui avaient jusqu'ici paru incompatibles. Mémoires de l'Acadenie des Sciences de Paris, 1744, pp. 417-426.

[9] Maupertius, P. L. M. de. Les lois du movement et du repos déduites d'un principe métaphysique. Mémoires de l'Acadenie des Sciences de Berlin, 1746, pp. 267-294.

[10] Newton, I. Correspondence, ed. by Scott, J. F., vol. IV (1694-1709), Cambridge University Press, Cambridge, 1967.

Ratio Mathematica Volume 36, 2019, pp. 79-87

The distinguishing number and the distinguishing index of co-normal product of two graphs

Saeid Alikhani*
Samaneh Soltani †

Abstract

The distinguishing number (index) $D(G)$ ($D'(G)$) of a graph G is the least integer d such that G has an vertex labeling (edge labeling) with d labels that is preserved only by a trivial automorphism. The co-normal product $G \star H$ of two graphs G and H is the graph with vertex set $V(G) \times V(H)$ and edge set $\{\{(x_1, x_2), (y_1, y_2)\} | x_1y_1 \in E(G) \text{ or } x_2y_2 \in E(H)\}$. In this paper we study the distinguishing number and the distinguishing index of the co-normal product of two graphs. We prove that for every $k \geq 3$, the k-th co-normal power of a connected graph G with no false twin vertex and no dominating vertex, has the distinguishing number and the distinguishing index equal two.

Keywords: distinguishing number; distinguishing index; co-normal product.
2010 AMS subject classifications: 05C15, 05C60. [1]

*Department of Mathematics, Yazd University, Yazd, Iran; alikhani@yazd.ac.ir
†Department of Mathematics, Yazd University, Yazd, Iran; s.soltani1979@gmail.com

[1]Received on February 12th, 2019. Accepted on May 3rd, 2019. Published on June 30th, 2019. doi: 10.23755/rm.v36i1.452. ISSN: 1592-7415. eISSN: 2282-8214.

1 Introduction and definitions

Let $G = (V, E)$ be a simple graph of order $n \geq 2$. We use the the following notations: The set of vertices adjacent in G to a vertex of a vertex subset $W \subseteq V$ is the *open neighborhood* $N(W)$ of W. Also $N(W) \cup W$ is called a *closed neighborhood* of W and denoted by $N[W]$. A *subgraph* of a graph G is a graph H such that $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$. If $V(H) = V(G)$, we call H a *spanning subgraph* of G. Any spanning subgraph of G can be obtained by deleting some of the edges from G. Two distinct vertices u and v are called *true twins* if $N[v] = N[u]$ and *false twins* if $N(v) = N(u)$. Two vertices are called *twins* if they are true or false twins. The number $|N(v)|$ is called the *degree* of v in G, denoted as $\deg_G(v)$ or $\deg(v)$. A vertex having degree $|V(G)| - 1$ is called a *dominating vertex* of G. Also, $\mathrm{Aut}(G)$ denotes the automorphism group of G, and graphs with $|\mathrm{Aut}(G)| = 1$ are called *rigid* graphs.

A labeling of G, $\phi : V \to \{1, 2, \ldots, r\}$, is said to be *$r$-distinguishing*, if no non-trivial automorphism of G preserves all of the vertex labels. The point of the labels on the vertices is to destroy the symmetries of the graph, that is, to make the automorphism group of the labeled graph trivial. Formally, ϕ is r-distinguishing if for every non-trivial $\sigma \in \mathrm{Aut}(G)$, there exists x in V such that $\phi(x) \neq \phi(\sigma(x))$. The *distinguishing number* of a graph G is defined by

$$D(G) = \min\{r|\ G \text{ has a labeling that is } r\text{-distinguishing}\}.$$

This number has defined in [1]. Similar to this definition, the *distinguishing index* $D'(G)$ of G has defined in [8] which is the least integer d such that G has an edge colouring with d colours that is preserved only by a trivial automorphism. If a graph has no nontrivial automorphisms, its distinguishing number is 1. In other words, $D(G) = 1$ for the asymmetric graphs. The other extreme, $D(G) = |V(G)|$, occurs if and only if G is a complete graph. The distinguishing index of some examples of graphs was exhibited in [8]. For instance, $D(P_n) = D'(P_n) = 2$ for every $n \geq 3$, and $D(C_n) = D'(C_n) = 3$ for $n = 3, 4, 5$, $D(C_n) = D'(C_n) = 2$ for $n \geq 6$, where P_n denotes a path graph on n vertices and C_n denotes a cycle graph on n vertices. A graph and its complement, always have the same automorphism group while their graph structure usually differs, hence $D(G) = D(\overline{G})$ for every simple graph G.

Product graph of two graphs G and H is a new graph having the vertex set $V(G) \times V(H)$ and the adjacency of vertices is defined under some rule using the adjacency and the nonadjacency relations of G and H. The distinguishing number and the distinguishing index of some graph products has been studied in literature (see [2, 6, 7]). The *Cartesian product* of graphs G and H is a graph, denoted by $G\Box H$, whose vertex set is $V(G) \times V(H)$. Two vertices (g, h) and (g', h') are adjacent if either $g = g'$ and $hh' \in E(H)$, or $gg' \in E(G)$ and $h = h'$.

In 1962, Ore [10] introduced a product graph, with the name Cartesian sum of graphs. Hammack et al. [4], named it co-normal product graph. The *co-normal product* of G and H is the graph denoted by $G \star H$, and is defined as follows:

$$V(G \star H) = \{(g, h) | g \in V(G) \text{ and } h \in V(H)\},$$
$$E(G \star H) = \{\{(x_1, x_2), (y_1, y_2)\} | x_1 y_1 \in E(G) \text{ or } x_2 y_2 \in E(H)\}.$$

We need knowledge of the structure of the automorphism group of the Cartesian product, which was determined by Imrich [5], and independently by Miller [9].

Theorem 1.1. [5, 9] *Suppose ψ is an automorphism of a connected graph G with prime factor decomposition $G = G_1 \Box G_2 \Box \ldots \Box G_r$. Then there is a permutation π of the set $\{1, 2, \ldots, r\}$ and there are isomorphisms $\psi_i : G_{\pi(i)} \to G_i$, $i = 1, \ldots, r$, such that*

$$\psi(x_1, x_2, \ldots, x_r) = (\psi_1(x_{\pi(1)}), \psi_2(x_{\pi(2)}), \ldots, \psi_r(x_{\pi(r)})).$$

Imrich and Klavžar in [7], and Gorzkowska et.al. in [3] showed that the distinguishing number and the distinguishing index of the square and higher powers of a connected graph $G \neq K_2, K_3$ with respect to the Cartesian product is 2.

The relationship between the automorphism group of co-normal product of two non isomorphic, non rigid connected graphs with no false twin and no dominating vertex is the same as that in the case of the Cartesian product.

Theorem 1.2. [12] *For any two non isomorphic, non rigid graphs G and H, $\mathrm{Aut}(G \star H) = \mathrm{Aut}(G) \times \mathrm{Aut}(H)$ if and only if both G and H have no false twins and dominating vertices.*

Theorem 1.3. [12] *For any two rigid isomorphic graphs G and H, $\mathrm{Aut}(G \star H) \cong S_2$.*

Theorem 1.4. [12]*The graph $G \star H$ is rigid if and only if $G \ncong H$ and both G and H are rigid graphs.*

In the next section, we study the distinguishing number of the co-normal product of two graphs. In section 3, we show that the distinguishing index of the co-normal product of two simple connected non isomorphic, non rigid graphs with no false twin and no dominating vertex cannot be more than the distinguishing index of their Cartesian product. As a consequence, we prove that all powers of a connected graph G with no false twin and no dominating vertex distinguished by exactly two edge labels with respect to the co-normal product.

2 Distinguishing number of co-normal product of two graphs

We begin this section with a general upper bound for the co-normal product of two simple connected graphs. We need the following theorem.

Theorem 2.1. [12] *Let G and H be two graphs and $\lambda : V(G \star H) \to V(G \star H)$ be a mapping.*

(i) If $\lambda = (\alpha, \beta)$ defined as $\lambda(g, h) = (\alpha(g), \beta(h))$, where $\alpha \in \mathrm{Aut}(G)$ and $\beta \in \mathrm{Aut}(H)$, then λ is an automorphism on $G \star H$.

(ii) If G is isomorphic to H and $\lambda = (\alpha, \beta)$ defined as $\lambda(g, h) = (\beta(h), \alpha(g))$, where α is an isomorphism on G to H and β is an isomorphism on H to G, then λ is an automorphism on $G \star H$.

Theorem 2.2. *If G and H are two simple connected graphs, then*

$$\max\big\{D(G\square H), D(G), D(H)\big\} \leq D(G\star H) \leq \min\big\{D(G)|V(H)|, |V(G)|D(H)\big\}.$$

Proof. We first show that $\max\{D(G), D(H)\} \leq D(G\star H)$. By contradiction, we assume that $D(G \star H) < \max\{D(G), D(H)\}$. Without loss of generality we suppose that $\max\{D(G), D(H)\} = D(G)$. Let C be a $(D(G\star H))$-distinguishing labeling of $G \star H$. Then the set of vertices $\{(g, h^*) \ : \ g \in V(G)\}$, where $h^* \in V(H)$ have been labeled with less than $D(G)$ labels. Hence we can define the labeling C' with $C'(g) := C(g, h^*)$ for all $g \in V(G)$. Since $D(G \star H) < D(G)$, so C' is not a distinguishing labeling of G, and so there exists a nonidentity automorphism α of G preserving the labeling C'. Thus there exists a nonidentity automorphism λ of $G\star H$ with $\lambda(g, h) := (\alpha(g), h)$ for $g \in V(G)$ and $h \in V(H)$, such that λ preserves the distinguishing labeling C, which is a contradiction. Now we show that $D(G\square H) \leq D(G \star H)$, and so we prove the left inequality. By Theorems 1.1 and 2.1, we can obtain that $\mathrm{Aut}(G\square H) \subseteq \mathrm{Aut}(G \star H)$, and since $V(G\square H) = V(G \star H)$, we have $D(G\square H) \leq D(G \star H)$.

Now we show that $D(G \star H) \leq \min\{D(G)|V(H)|, |V(G)|D(H)\}$. For this purpose, we define two distinguishing labelings of $G \star H$ with $D(G)|V(H)|$ and $|V(G)|D(H)$ labels, respectively. Let C be a $D(G)$-distinguishing labeling of G and C' be a $D(H)$-distinguishing labeling of H. We suppose that $V(G) = \{g_1, \ldots, g_n\}$ and $V(H) = \{h_1, \ldots, h_m\}$, and define the two following distinguishing labelings L_1 and L_2 of $G \star H$ with $D(G)|V(H)|$ and $|V(G)|D(H)$ labels.

$$L_1(g_j, h_i) := (i-1)D(G) + C(g_j),$$
$$L_2(g_j, h_i) := (j-1)D(H) + C'(h_i).$$

We only prove that the labeling L_1 is a distinguishing labeling, and by a similar argument, it can be concluded that L_2 is a distinguishing labeling of $G \star H$. If f is an automorphism of $G \star H$ preserving the labeling L_1, then f maps the set $H_i := \{(g_j, h_i) \ : \ g_j \in V(G)\}$ to itself, setwise, for all $i = 1, \ldots, m$. Since the restriction of f to H_i can be considered as an automorphism of G preserving the distinguishing labeling C, so for every $1 \leq i \leq m$, the restriction of f to H_i is the identity automorphism. Hence f is the identity automorphism of $G \star H$. $\square$

The bounds of Theorem 2.2 are sharp. For the right inequality it is sufficient to consider the complete graphs as the graphs G and H. In fact, if $G = K_n$ and $H = K_m$, then $G \star H = K_{nm}$. For the left inequality we consider the non isomorphic rigid graphs as the graphs G and H. Then by Theorem 1.4, we conclude that $G \star H$ and $G \square H$ are a rigid graph and hence $\max\{D(G \square H), D(G), D(H)\} = D(G \star H)$.

With respect to Theorems 1.1 and 1.2, we have that the automorphism group of a co-normal product of connected non isomorphic, non rigid graphs with no false twin and no dominating vertex, is the same as automorphism group of the Cartesian product of them, so the following theorem follows immediately:

Theorem 2.3. *If G and H are two simple connected, non isomorphic, non rigid graphs with no false twin and no dominating vertex, then $D(G \star H) = D(G \square H)$.*

Since the path graph P_n ($n \geq 4$), and the cycle graph C_m ($m \geq 5$) are connected, graphs with no false twin and no dominating vertex, then by Theorem 2.3 we have $D(P_n \star P_q) = D(P_n \star C_m) = D(C_m \star C_p) = 2$ for any $q, n \geq 3$, where $q \neq n$ and $m, p \geq 5$, where $m \neq p$. (see [7] for the distinguishing number of Cartesian product of these graphs).

To prove the next result, we need the following lemmas.

Lemma 2.1. [13] *For any two distinct vertices (v_i, u_j) and (v_r, u_s) in $G \star H$, $N((v_i, u_j)) = N((v_r, u_s))$ if and only if*

(i) $v_i = v_r$ *in G and* $N(u_j) = N(u_s)$ *in H, or*

(ii) $u_j = u_s$ *in H and* $N(v_i) = N(v_r)$ *in G, or*

(iii) $N(v_i) = N(v_r)$ *in G and* $N(u_j) = N(u_s)$.

Lemma 2.2. [13] *A vertex (v_i, u_j) is a dominating vertex in $G \star H$ if and only if v_i and u_j are dominating vertices in G and H, respectively.*

Theorem 2.4. [12] *For a rigid graph G and a non rigid graph H, $|\mathrm{Aut}(G \star H)| = |\mathrm{Aut}(H)|$ if and only if G has no dominating vertex and H has no false twin.*

Now we are ready to state and prove the main result of this section.

Theorem 2.5. *Let G be a connected graph with no false twin and no dominating vertex, and $\star G^k$ the k-th power of G with respect to the co-normal product. Then $D(\star G^k) = 2$ for $k \geq 3$. In particular, if G is a rigid graph, then for $k \geq 2$, $D(\star G^k) = 2$.*

Proof. By Lemmas 2.1 and 2.2, we can conclude that $G \star G$ has no false twin and no dominating vertex. We consider the two following cases:

Case 1) Let G be a non rigid graph. If $H := G \star G$, then $D(\star G^3) = 2$ by Theorem 2.3. Now by induction on k, we have the result.

Case 2) Let G be a rigid graph. In this case, $|\mathrm{Aut}(G \star G)| = 2$, by Theorem 1.3, and so $D(G \star G) = 2$. If $H := G \star G$, then $|\mathrm{Aut}(G \star H)| = |\mathrm{Aut}(H)|$, by Theorem 2.4. Hence $|\mathrm{Aut}(\star G^3)| = 2$. By induction on k and using Theorem 2.4, we obtain $D(\star G^k) = 2$ for $k \geq 2$, where G is a rigid graph. □

3 Distinguishing index of co-normal product of two graphs

In this section we investigate the distinguishing index of co-normal product of graphs. Pilśniak in [11] showed that the distinguishing index of traceable graphs, graphs with a Hamiltonian path, of order equal or greater than seven is at most two.

Theorem 3.1. [11] *If G is a traceable graph of order $n \geq 7$, then $D'(G) \leq 2$.*

We say that a graph G is almost spanned by a subgraph H if $G - v$, the graph obtained from G by removal of a vertex v and all edges incident to v, is spanned by H for some $v \in V(G)$. The following two observations will play a crucial role in this section.

Lemma 3.1. [11] *If a graph G is spanned or almost spanned by a subgraph H, then $D'(G) \leq D'(H) + 1$.*

Lemma 3.2. *Let G be a graph and H be a spanning subgraph of G. If $\mathrm{Aut}(G)$ is a subgroup of $\mathrm{Aut}(H)$, then $D'(G) \leq D'(H)$.*

Proof. Let to call the edges of G which are the edges of H, H-edges, and the others non-H-edges, then since $\mathrm{Aut}(G) \subseteq \mathrm{Aut}(H)$, we can conclude that each automorphism of G maps H-edges to H-edges and non-H-edges to non-H-edges. So assigning each distinguishing edge labeling of H to G and assigning non-H-edges a repeated label we make a distinguishing edge labeling of G. □

Since for two distinct simple non isomorphic, non rigid connected graphs, with no false twin and no dominating vertex we have $\mathrm{Aut}(G\star H) = \mathrm{Aut}(G\Box H)$, so a direct consequence of Lemmas 3.1 and 3.2 is as follows:

Theorem 3.2. *(i) If G and H are two simple connected graphs, then $D'(G \star H) \leq D'(G\Box H) + 1$.*

(ii) If G and H are two simple connected non isomorphic, non rigid graphs with no false twin and no dominating vertex, then $D'(G \star H) \leq D'(G\Box H)$.

Theorem 3.3. *Let G be a connected graph with no false twin and no dominating vertex, and $\star G^k$ the k-th power of G with respect to the co-normal product. Then for $k \geq 3$, $D'(\star G^k) = 2$. In particular, if G is a rigid graph, then for $k \geq 2$, $D'(\star G^k) = 2$.*

Proof. By Lemmas 2.1 and 2.2, we can conclude that $G \star G$ has no false twin and no dominating vertex. We consider the two following cases:

Case 1) Let G be a non rigid graph. If $H = G \star G$, then $D(\star G^3) = 2$ by Theorem 3.2(ii). Now by an induction on k, we have the result.

Case 2) Let G be a rigid graph. In this case, $|\mathrm{Aut}(G \star G)| = 2$, by Theorem 1.3, and so $D(G \star G) = 2$. If $H := G \star G$, then $|\mathrm{Aut}(G \star H)| = |\mathrm{Aut}(H)|$, by Theorem 2.4. Hence $|\mathrm{Aut}(\star G^3)| = 2$. By an induction on k and using Theorem 2.4, we obtain $D(\star G^k) = 2$ for $k \geq 2$, where G is a rigid graph. □

Theorem 3.4. *Let G be a connected graph of order $n \geq 2$. Then $D'(G\star K_m) = 2$ for every $m \geq 2$, except $D'(K_2 \star K_2) = 3$.*

Proof. Since $|\mathrm{Aut}(G \star K_m)| \geq 2$, so $D'(G \geq K_m) = 2$. With respect to the degree of vertices $G \star K_m$ we conclude that $G \star K_m$ is a traceable graph. We consider the two following cases:

Case 1) Suppose that $n \geq 2$. If $m \geq 3$, or $m = 2$, and $n \geq 4$, then the order of $G\star K_m$ is at least 7, and so the result follows from Theorem 3.1. If $m = 2$, $n = 3$, then $G = P_3$ or K_3. In each case, it is easy to see that $D'(G \star K_m) = 2$.

Case 2) Suppose that $n = 2$. Then $G = K_2$, and so $G \star K_m = K_{2m}$. Thus $D'(G \star K_m) = 2$ for $m \geq 3$, and $D'(K_2 \star K_2) = D'(K_4) = 3$. □

By the value of the distinguishing index of Cartesian product of paths and cycles graphs in [3] and Theorem 3.2, we can obtain this value for the co-normal product of them as the two following corollaries.

Corolary 3.1. (i) *The co-normal product $P_m\star P_n$ of two paths of orders $m \geq 2$ and $n \geq 2$ has the distinguishing index equal to two, except $D'(P_2\star P_2) = 3$.*

(ii) *The co-normal product $C_m \star C_n$ of two cycles of orders $m \geq 3$ and $n \geq 3$ has the distinguishing index equal to two.*

(iii) *The co-normal product $P_m \star C_n$ of orders $m \geq 2$ and $n \geq 3$ has the distinguishing index equal to two.*

Proof.

(i) If $n, m \geq 4$, then the result follows from Theorem 3.2 (ii). If $n = 2$ or $m = 2$, then we have the result by Theorem 3.4. For the remaining cases, with respect to the degree of vertices in $P_m \star P_n$, we obtain easily the distinguishing index.

(ii) If $n, m \geq 5$, then the result follows from Theorem 3.2 (ii). If $n = 3$ or $m = 3$, then we have the result by Theorem 3.4. For the remaining cases we use of Hamiltonicity of $C_m \star C_n$ and Theorem 3.1.

(iii) If $n \geq 5$ and $m \geq 4$, then the result follows from Theorem 3.2 (ii). If $n = 3$ or $m = 2$, then we have the result by Theorem 3.4. The remaining cases are $C_n \star P_3$ and $C_4 \star P_m$. In the first case and with respect to the degree of vertices in $C_n \star P_3$, we obtain easily the distinguishing index. In the latter case, we use of Hamiltonicity of $C_4 \star P_m$ and Theorem 3.1. □

4 Acknowledgements

The authors would like to express their gratitude to the referee for her/his careful reading and helpful comments.

References

[1] M.O. Albertson and K.L. Collins, *Symmetry breaking in graphs*, Electron. J. Combin. 3 (1996), #R18.

[2] S. Alikhani and S. Soltani, *The distinguishing number and distinguishing index of the lexicographic product of two graphs*, Discuss. Math. Graph Theory 38 (2018) 853-865.

[3] A. Gorzkowska, R. Kalinowski, and M. Pilśniak, *The distinguishing index of the Cartesian product of finite graphs*, Ars Math. Contem. 12 (2017), 77-87.

[4] R. Hammack, W. Imrich and S. Klavžar, *Handbook of product graphs (second edition)*, Taylor & Francis group 2011.

[5] W. Imrich, *Automorphismen und das kartesische Produkt von Graphen*, Oesterreich. Akad. Wiss. Math.-Natur. Kl. S.-B. II 177 (1969), 203-214.

[6] W. Imrich, J. Jerebic and S. Klavžar, *The distinguishing number of Cartesian products of complete graphs*, European J. Combin. 29 (4) (2008), 922-929.

[7] W. Imrich and S. Klavžar, *Distinguishing Cartesian powers of graphs*, J. Graph Theory, 53.3 (2006), 250-260.

[8] R. Kalinowski and M. Pilśniak, *Distinguishing graphs by edge colourings*, European J. Combin. 45 (2015), 124-131.

[9] D.J. Miller, *The automorphism group of a product of graphs*, Proc. Amer. Math. Soc. 25 (1970), 24-28.

[10] O. Ore, *Theory of Graphs*, Amer. Math. Society 1962.

[11] M. Pilśniak, *Improving upper bounds for the distinguishing index*, Ars Math. Contemp. 13 (2017), 259-274.

[12] S. Rehman and I. Javaid, *Fixing number of co-noraml product of graphs*, arXiv:1703.00709 (2017).

[13] S. Rehman and I. Javaid, *Resolving, dominating and locating dominating sets in co-normal product of graphs*, submitted.

Ratio Mathematica Volume 36, 2019, pp.89-95

The theorem of the complex exponentials

Alberto Daunisi[*]

Abstract

This paper describes a new theorem that relates the lengths of the legs of a right triangle with the ratio of three complex exponentials. The big novelty of the theorem consists in transforming two real measures of legs derived from Euclidean geometry into a combination of imaginary elements obtained from the complex analysis.

Keywords: complex analysis, complex geometry

2010 subject classification: 37K20.[†]

[*] Researcher in Mathematics, Bologna (Italy). alberto.daunisi@gmail.com

[†] Received on December 10th, 2018. Accepted on April 24th, 2019. Published on June 30th, 2019. doi: 10.23755/rm.v36i1.472. ISSN: 1592-7415. eISSN: 2282-8214.

1. Introduction

It seems difficult, apparently, to imagine that the difference between the lengths of the legs of a right triangle may have some connection with the ratio of complex numbers, or vice-versa, that a ratio of complex numbers may be obtained from the difference of the legs of a right triangle, but this theorem, that we will call, precisely, *Theorem of the complex exponentials,* shows that it is possible.

Let's start with some historical and mathematical considerations, from which our research is inspired. In ancient Greece, the right triangles were basically solved by the *first and second theorem of Euclide* (IV-III century B.C.) and by the *theorem of Pythagoras* (about 575-495 B.C.). This happened because the Greek trigonometry, that was only applied to the study of astronomy, was based on the measurement of the ropes of a circle (subtended by a certain angle), rather than on that of sines and cosines. The functions sine and cosine, developed by the Indians in the IV-V century A.C., have been imported in the Arab world around the VIII century A.C., and then, to the West world, a few centuries later. From this moment, the triangles started to be solved by the relations that bind the lengths of the sides of the triangle with the values of the trigonometric functions of its angles. In particular, two fundamental trigonometric theorems were introduced, through which it has been possible to solve any problem related to the elements of a triangle: *the theorem of sines and the theorem of Carnot*. The first one states that *in any triangle the ratio between one side and the sine of the opposite angle is always constant and equal to the diameter of the circle circumscribed to the given triangle*; the second one states that *in any triangle the square of one side is equal to the sum of the squares of the other two, plus their product to the cosine of the angle included*. From the theorem of sines, applied to the right triangles, it descends the theorem according to which *in a right triangle a leg is equal to the product of the hypotenuse for the sine of the angle opposite to the leg*. Finally, we arrive at the XVIII century A.C., where, in another branch of mathematics completely different from the above one, is developed, in all its entirety, the theory of complex numbers of the form $x+iy$, with x and y real numbers and $i = \sqrt{-1}$ the imaginary unit. In particular, the studies of Abraham de Moivre (1667-1754) and Leonhard Euler (1707-1783) provided to the complex numbers a definitive and systematic structure from which descended the complex trigonometric functions and the complex exponential functions. De Moivre left us the famous formula (1739) that calculates the power of a complex number expressed in the form trigonometric $(\cos \alpha + \sin \alpha)^m = \cos(m\alpha) + i \sin(m\alpha)$, while Euler left us the equally famous formula (1748) that binds the trigonometric functions *sine* and *cosine* to the complex exponential function $e^{iw} = \cos w + i \sin w$.

Our theorem is inspired by a very specific motivation: considering that the sides of a right triangle may be expressed by a trigonometric function, and this one by a complex variable, we wanted to discover if the same sides may have a relationship with the elements of the complex analysis and, in case of positive response, in which way and form. With this purpose, we discovered two important results: the first one is that the ratio between the difference of the legs of a right triangle and the difference of their projections on the hypotenuse, multiplied by the cosine of half-difference of two angles opposite to the legs, is always constant; the second one is that this constant is given by the ratio between complex exponentials, or their powers, where the most important constants of the all mathematics are appearing: the constant of Napier **e** (or of Euler), introduced by John Napier on 1618 and used systematically by Euler (1736) for its exponentials; the imaginary unit **i**, officially introduced by Friedrich Gauss (1777-1855) in an essay of 1832; the constant of Archimedes (287-212 B.C.) **π**, calculated with approximation by the greek mathematician in the III century B.C. and definitively calculated, with 35 decimal digits, by Ludolf van Ceulen on 1610. Both the above important results are set forth and proved in the following theorem.

2. The theorem of the complex exponentials

Statement: In a right triangle CAB (rectangle in A), where **a** is the hypotenuse, **c** and **b** the legs, **m** and **n** the projections of the respective legs **c** and **b** on the hypotenuse, **γ** and **β** the angles respectively opposed to **c** and **b**, it results:

$$\frac{c-b}{m-n}\cos\frac{\gamma-\beta}{2}=\frac{e^{i\pi/4}}{e^{2i\pi}+e^{i\pi/2}}$$

where **e**=2,71….is the Napier's constant, **π**=3,14…..is the pi and **i**=$\sqrt{-1}$ is the imaginary unit.

Proof. Let us consider the right triangle CAB of Figure 1, rectangle in A (α=90°), having hypotenuse a, height h, minor leg b and major leg c, n and m the respective projections of b and c on the hypotenuse a, γ the angle opposed to c and β the angle opposed to b.

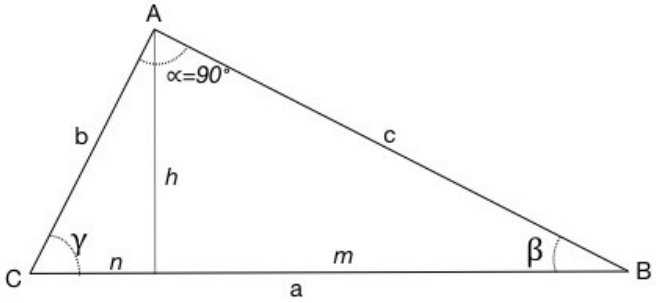

Figure 1

With reference to the right triangle of Figure 1, we know that the first Euclid's theorem asserts:

$$\begin{aligned} c^2 &= a \cdot m \\ b^2 &= a \cdot n \end{aligned} \tag{1}$$

from which, subtracting member to member, it derives:

$$c^2 - b^2 = a\,(m - n) \tag{2}$$

namely:

$$(c + b) \cdot (c - b) = a \cdot (m - n) \tag{3}$$

Taking into account that it is: c=a sinγ and b=a sinβ, from (3) it derives:

$$\frac{c - b}{m - n} = \frac{a}{a(\sin\gamma + \sin\beta)} \tag{4}$$

Simplifying and applying the formulas Prosthaphaeresis to the denominator of second member (4), from (4) it's obtained:

$$\frac{c-b}{m-n} = \frac{1}{2\sin\frac{\gamma+\beta}{2}\cos\frac{\gamma-\beta}{2}} \tag{5}$$

But we know that $\gamma + \beta = 90°$, so in the denominator of (5) it's $2 \sin\frac{\gamma+\beta}{2} = \sqrt{2}$, therefore from (5) it derives:

$$\frac{c-b}{m-n} = \frac{1}{\sqrt{2}\cos\frac{\gamma-\beta}{2}} \tag{6}$$

namely:

$$\frac{c-b}{m-n}\cos\frac{\gamma-\beta}{2} = \frac{1}{\sqrt{2}} \tag{7}$$

We know, from complex number's theory, that the trigonometric form of the complex number 1+i is:

$$1+i = \sqrt{2}\,(\cos\frac{\pi}{4} + i\sin\frac{\pi}{4}) \tag{8}$$

For the formula of Euler it's:

$$\cos\frac{\pi}{4} + i\sin\frac{\pi}{4} = e^{i\pi/4} \tag{9}$$

Replacing (9) in (8), we obtain:

$$1+i = \sqrt{2}\cdot e^{i\pi/4} \tag{10}$$

namely:

$$\frac{1}{\sqrt{2}} = \frac{e^{i\pi/4}}{1+i} \tag{11}$$

Let us remind now that it is:

$$1=e^{2i\pi} \quad \text{and} \quad i=e^{i\pi/2} \tag{12}$$

Replacing (12) in (11), we obtain:

$$\frac{1}{\sqrt{2}} = \frac{e^{i\pi/4}}{e^{2i\pi} + e^{i\pi/2}} \tag{13}$$

Finally, replacing the second member of (7) with the second member of (13), we obtain:

$$\frac{c-b}{m-n}\cos\frac{\gamma-\beta}{2} = \frac{e^{i\pi/4}}{e^{2i\pi} + e^{i\pi/2}}$$

And the theorem is thus proven.

Conclusions

We have shown a theorem born from the motivation to investigate and solve a problem: to link a geometric result of III century B.C., although it reworked by the trigonometric functions of XVI century, to the last theories of complex numbers of XVIII century, apparently irreconcilables with the Euclidean geometry. We think to have got two relevant teachings: on the one hand we have bound the elements of a right triangle (legs and angles) to a constant of complex analysis, given by the combination of three most important constants of mathematics; on the other hand we have notably pointed out a precise methodological procedure of the proof, based strictly on the deductive method, where, starting from a general axiom alleging geometric structure of the right triangles, we reached, through a series of rigorous logical concatenations, a particular result alleging new structure of complex analysis.

We finally think that from this article we also can draw another useful teaching: to discover this theorem allowed us to investigate on three completely different (among their) branches of mathematics (Euclidean geometry, trigonometric functions, complex analysis), born and developed in different

ages, transmitted by several men separated by time and by different languages, cultures and religions, who, although not knowing themselves with each other, have always improved the ideas of their predecessors and transmitted it to the future generations. They have been united only by their love for mathematics, in addition to the desire to contribute to its development. We think that we all must pick up an example from this act of faith, that only mathematics, between all sciences, is able to provide.

References

[1] Alberto Daunisi (2014). L'Ultimo Teorema di Fermat, Storia Matematica, BookSprint Edizioni, Salerno-Italy.

[2] Howard Levi (1960). Foundations of geometry and trigonometry, Prentice-Hall, Inc. Englewood Cliffs, New Jersey.

[3] Joseph Bak and Donald Newman (1997). Complex Analysis, Springer-Verlag New York Inc.

[4] Morris Kline (1998). Calculus, Dover Publications, Inc. Mineola, New York.

[5] Barry Mazur (2004). Imaging numbers, Picador (Farrar, Straus and Giroux), New York.

Ratio Mathematica Volume 36, 2019, pp. 97-108

Creation of the concept of zero-point method in teaching mathematics

Tomáš Lengyelfalusy[*]
Dalibor Gonda[†]

Abstract

Pupils learn different calculating algorithms. The effective use of learned algorithms requires creativity in their application to solving diverse tasks. To achieve this goal, it is necessary to create a concept of the calculating algorithm for pupils. The present paper describes a method of creating a zero-point method. The teaching of this method is divided into two stages. In the first stage, the student masters the basic algorithm and becomes familiar with the main ideas of this method, while in the second stage a student learns how to apply this method with some modifications in other types of tasks. In our article, we present the application of a zero-point method in solving quadratic inequalities.

Keywords: concept creation, method of the zero-point, algorithm, pupils' understanding.

2010 AMS subject classification: 97D40. [‡]

[*] Department of Didactics, Technology and Educational Technologies, DTI University, Sládkovičova 533/20, 018 41, Dubnica nad Váhom. lengyelfalusy@dti.sk.
[†] Faculty of Humanities, University of Žilina, Univerzitná 8215/1, 010 26 Žilina. daliborgonda@gmail.com.
[‡] Received on May 12nd, 2018. Accepted on June 24th, 2019. Published on June 30th, 2019. doi: 10.23755/rm.v36i1.466. ISSN: 1592-7415. eISSN: 2282-8214.

1 Introduction

Recently the education at primary and secondary schools has undergone several reforms. One of the essential features of these reforms has been a reduction of the curriculum of individual subjects and reducing the number of lessons, especially science lessons. The main aim of reducing the curriculum and thus reducing the demands was monitoring the improvements in educational achievements of our students [1]. But PISA 2015 test results say otherwise. Slovak students achieved in 2015, on average, significantly worse results than the OECD average. It is worthy of reflection that our students achieve the best results-the results almost on an average of the best students in the OECD. Another feature of this educational reform is teaching a "playful" way. Pupils should acquire new knowledge and skills not by memorizing and practising, but above all by the playful way. PISA testing in 2015 showed that, in terms of pupils' attitudes to learning, our students declare significantly lower endurance to solve complex problems, lower openness to solve tasks and less belief in their own abilities. It can also negatively be reflected on their results in mathematics. Compared to 2003, many of Slovak students' attitudes to learning significantly deteriorated. 2015 PISA test results are in substantial agreement with the results of the external part of the school leaving examination (maturita). All Slovak students have to pass maturita from Slovak language and literature and a foreign language. Only those students have to pass maturita from mathematics, who choose math as a maturita subject. Nevertheless, over the past three years, the average percentage of school maturita exam in mathematics is always worse than the average percentage of school maturita exam in compulsory subjects. We think that the ideas of school reforms are correct, but it turns out that it is not right to use the same methods to achieve the goals for all subjects.

Mathematics affects almost every area of human life. In the education of our youth, who should be, according to the reference of John Paul II., our hope for the future. Math is challenging in its own way but at the same time can also be beautiful. We think it is necessary to seek such forms and methods of teaching mathematics [2], that we make the beauty of math available to students [3]. In the following lines, we will outline one possible way of teaching mathematics.

2 Two stages of mathematical education

Mathematical education can be divided into two stages. The first, basic stage is the acquisition of basic calculating algorithms. These calculating algorithms are acquired by students, who practice them on the appropriate number of tasks. We can talk about math "drill", without which it is

impossible to be a successful solver of mathematical problems. The information-receptive didactic method with a combination of the reproductive method is mainly used in this stage. It is very important that the student acquires the necessary skill of how to use them by repeated use of basic calculating algorithms. The teacher, by the right choice of tasks, ensures that pupils acquire these calculating algorithms at least at the level of understanding, not only at the level of memorization. The second, application stage is the application of the acquired algorithms in different areas of mathematics and other disciplines or in practical everyday life. At this stage, the mathematical "drill" is replaced by mathematical thinking. Based on the assignment a student considers what math knowledge and skills he can use to solve the task. Unlike the first stage, he must learn that the first step of task solution is not to count but to think. Based on a detailed consideration and possible task mathematization the student chooses a suitable calculating algorithm. At this stage, the teacher becomes a moderator of solution and uses a heuristic didactic method. At this stage, in terms of the taxonomy of educational objectives, the level of acquirement of calculating algorithms will be increased for the minimum to the application level. If the teaching is correct, we can say, that at this stage, the students do not learn new calculating algorithms. At this stage, students gain new, mainly theoretical knowledge of mathematics, and also learn how to apply already gained calculating algorithms in a new context. The above-described stages are illustrated on the example of the method of zero points.

3 Method of zero points

Solving of the most mathematical problems includes solving of various equations and inequalities, or their systems. The tasks, where it is necessary to solve equations, inequalities and their systems belong to the declaratory mathematical tasks [4].

Declaratory mathematical tasks are historically the oldest mathematical tasks. When solving these tasks the mathematical concepts and methods. Those are the tasks that require finding, calculating, constructing etc. of all mathematical objects of a particular type, having the desired properties. In each declaratory task, we can define as the frame of considerations some non-empty set $\boldsymbol{M}$ of mathematical objects, which is a carrier of a particular structure. Using the terms belonging to this structure, it is then possible to express the desired properties of those objects of the set $\boldsymbol{M}$ that we are looking for. To characterize the elements of the set $\boldsymbol{M}$ we use propositional form $\boldsymbol{V}$ (x) which verity domains then create subsets of the set $\boldsymbol{M}$. In each determinative task there is a subset $\boldsymbol{K}$ of set $\boldsymbol{M}$, which elements have the characteristics required by task assignment. The task and the objective of the investigator are to determine the set $\boldsymbol{P}$ by naming of its elements or to operate

with already known subsets of the set $\boldsymbol{M}$. We can solve the mathematical declaratory task with the direct and indirect methods.

The direct method of solving means a process by which we determine the set of solutions $\boldsymbol{K}$ so that we work exclusively with sets that belong to the chain of sets inclusions

$$\varnothing \subset \ldots \subset \boldsymbol{K} \subset \ldots \subset \boldsymbol{M},$$

where $\boldsymbol{M}$ is a non-empty set of mathematical objects, among which elements we are looking for the solving of the task. Indirect methods consist in the fact that instead of solving the task that is defined we solve the other task or other tasks (using some direct method) and the results are used to obtain the results of the original task. One of the indirect methods is to switch to subtasks on the same set. We divide the set $\boldsymbol{M}$ to individual subsets and we investigate the specific location of each original task. We will obtain partial solutions to the original task on each of these subsets. The overall result for the task will be obtained by the unification of partial results. Method of zero points can be included precisely into that category of indirect methods (in some literature this method is also called the method of intervals).

The essential feature of the method of zero points is the attempt to divide tasks into several "sub-tasks", solving them on the corresponding subsets - intervals. To deal with this method it is necessary to learn the algorithms of expression modifying, polynomial factorization to the product of the root factors and solving various types of equations [5].

4 Teaching the method of zero points

The teaching of this method is recommended to be realized in three levels.

Level 1: Acquisition of the method

The students meet the method of zero points for the first time when they solve inequalities with an unknown in the denominator. Its basic steps are learned through leading example.

Example 1: On the set **R** solve the inequality $\frac{2x+3}{x-1} < 1$

Solution: Most students have the following knowledge on solving the inequalities: Inequalities are solved using the same equivalent adjustment as the equations. If the inequality is multiplied or divided by a negative number, the sign of inequality is changed to the opposite. On the basis of this knowledge the first step of solving is an attempt to remove a fraction of the assigned inequality, that is, they multiply the inequality by the expression (x-1). Already in the introduction of the model example, students learn another difference between solving the equations and inequalities. Inequalities, unlike equations, cannot be multiplied by the expression of which I cannot clearly decide whether it is positive or negative. If we want students to use the

proposed adjustment, it is first necessary to determine for which values of the variable the expression $(x-1)$ is positive and for which negative. Consequently, it is necessary to divide the solving of the inequalities to, in this case, two parts - when the expression $(x-1)$ is positive and the sign of inequality does not change after multiplication, and when the expression $(x-1)$ is negative and the sign of inequality changes to opposite one after multiplication. Basically, the assigned inequality should be tackled twice. We recommend concentrating on the issue of "multiplying inequalities" and pay sufficient attention, because it is needed to change students fixed “definition” of solving the inequalities. The method of zero points does not require multiple solving of the same inequalities and therefore it, is considered to be mora effective method. It can be divided into the following steps:

1. Annulling the right side of the equation:

$$\frac{2x+3}{x-1}-1<0$$

2. Simplifying the expression on the left side of the inequality:

$$\frac{x+4}{x-1}<0 \qquad (1)$$

After these adjustments, we draw the students’ attention to the intermediate target of our solutions. We compare the fraction to zero. Therefore, we only need to determine the sign of the expression $\frac{x+4}{x-1}$. Our partial objective is to determine for what value of x it is positive and for what value negative.

3. Determining the zero points:

Zero points are the values of variable x for which numerator and denominator separately on the left side of the inequality takes the zero value. Zero points can be determined based on solving the equation $x+4=0;\ \ x-1=0$. Zero points are NB: -4; 1.

4. Adjusted numerical axis:

We come to the core of the method. First, we explain the function of zero points. Zero points divided real numbers, in this case, into the three sets - intervals. For each interval is true: The expression $\frac{x+4}{x-1}$ is positive or negative in the whole interval, in other words, it does not change the resulting sign. The adjacent intervals the expression $\frac{x+4}{x-1}$ has different resulting signs. Based on the above it is sufficient, if we want to determine the final sign, to substitute any number belonging to this interval to the expression. If we know the final sign in one of the intervals, we automatically recognize the resulting sign in all intervals as signs alternate. Using that knowledge, we can create a customized numerical axis (Fig. 1):

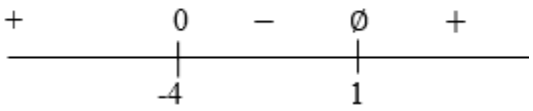

Figure 1.

There are numbers under the axis to be substituted for variable x in the expression; above the axis are values of the expression after substitution. That is, if we substitute any number from the interval (-4, 1) the resulting sign of the expression $\frac{x+4}{x-1}$ is negative, after the substituting $x = -4$, the resulting value of the expression is zero. Symbol ∅ means that for the value $x = 1$ the expression is not defined.

The adjusted numerical axis can be created as follows. First, on the numerical axis (from the bottom), we mark zero points. (Students often automatically show zero even if it is not the zero point on the numerical axis. There should be **only** zero points on the numerical axis).

We substitute any number different from zero points to the expression on the left side of the inequality. If the zero point is not zero, we substitute number **zero** to the variable. After substituting the number zero to the variable x, the expression $\frac{x+4}{x-1}$ has the value of - 4. Then we write a minus sign above the numerical axis in the part corresponding to the interval, from which we substituted the number zero. The signs in the other intervals will be completed without calculations, whilst complying with the principle of alternating signs. We complete 0 above the zero point "of the numerator" and the sign ∅ above the zero point "of the denominator".

5 Determination of results

Those values of variable x for which the expression $\frac{x+4}{x-1}$ acquire negative values will be the solution to the inequality (1). Based on the adjusted numbering axis, the search solution to the assigned inequality is the interval from -4 to 1. Finally, we determine the "brackets" of the final interval. Zero point, above which is symbol ∅, cannot be the solution, therefore it will be at zero point "of the denominator" always round bracket. If there is the symbol 0 above the zero point, it means, that after its substituting, the resulting value of the expression is zero. However, we are looking for negative values of the expression and therefore the number -4 has a round bracket. The ultimate solution is $x\in$ (-4.1).

After solving the model example we recommend to discuss with students how the solution would change if we solve the inequality

$$\frac{2x+3}{x-1} > 1 \text{ and the inequality } \frac{2x+3}{x-1} \leq 1.$$

Students should be aware, that in both cases, the first four steps will be identical with the model example. In the fifth step, based on the same considerations, the solution of the inequality would be $\frac{2x+3}{x-1} > 1 \;\; x \in$

$(-\infty;-4) \cup (1;\infty)$. The solution of inethe quality $\frac{2x+3}{x-1} \leq 1$ is $x \in \langle -4;1)$

Level 2: Understanding of the method

The main idea behind the method of zero points can be considered a comparison of the fraction with zero. A student knows that if a numerator and a denominator have the same final sign, so the fraction is positive if they have different sign fraction is negative. The correct application of this idea leads to an understanding of the method of zero points and also to a more efficient using of this method. The correct application of the main idea it is essential to understand the "functioning" of zero points. The zero point for this expression, in principle, divides the set of real numbers (NA) into three subsets. On one of the subsets, it acquires only positive values, on another one just negative. The third subset is only composed of zero point and the expression of the set acquires a value of 0. For example, the expression $x - 5$ has a zero point 5. Then, the expression acquires negative values on the set $M_1 = (-\infty; 5)$, on the set $M_2 = (5; \infty)$ it acquires positive values and on the set $M = \{5\}$ it takes the value 0. Thus, we can simplistically say, that there is a different sign of the expression from the various sides of the zero point. If the expression is in productive form, the zero points of individual members of the product create the zero points of all expression.

Example 2: On the set R solve the inequality $\frac{(x-9)(x+1)^2}{(x-4)(x+5)} > 0$

Solution: Zero points -5; -1; 4; 9.

At first, we draw attention to the expression $(x+1)^2$. This expression acquires for all $x \in$ R non-negative values. Therefore, it has no influence to the final sign of the expression $\frac{(x-9)(x+1)^2}{(x-4)(x+5)}$. The zero point of the expression $(x+1)^2$ can be described as "unnecessary" zero point and it will not be showed on the adjusted numbering axis. (If we showed it there, the theory of alternation marks would not apply.)

To obtain the solution of the inequality we only need to know the final sign of the expression $\frac{(x-9)(x+1)^2}{(x-4)(x+5)}$. Therefore. after substituting, for example $x = 0$, it is not necessary to know the numerical value. At the same time, we know that it is not necessary to substitute to the expression $(x+1)^2$. By applying the above mentioned ideas after substituting $x = 0$ we obtain "a signed" value of the expression: $\frac{-}{-.+}$.

We set the adjusted numbering axis (Fig. 2):

Figure 2.

The solution of the assigned inequality based on the adjusted numbering line and the sign of inequality is $x \in (-5; 4) \cup (9; \infty)$. To obtain the final solution, we must once again pay attention to "needless" zero point. We know that for $= -1$, the expression acquires the resulting value zero on the left side. Therefore, the number -1 does not belong to the solution of our set of ithe nequality. The ultimate solution of the inequality $x \in (-5; -1) \cup (-1; 4) \cup (9; \infty)$.

Level 3: Application of the method

After mastering the basic algorithm and understanding the method of zero points we recommend to focus on the teaching of its application in other types of examples, such as those in which students can penetrate into its mysteries. The closest type of tasks is inequalities in the productive form. The student already knows that there are the same rules for comparison zero to the product as for the comparison of the quotient to zero. Therefore, in solving qualities in productive form, the method of zero points can be used identically as in solving the inequalities in productive form. Quadratic inequality can be seen as inequality in the productive form. In example 3 we show a sample solution.

Example 3: On the set **R** solve the inequality $x^2 + 3x - 4 \geq 0$.

Solution: Quadratic trinomial on the left side of the inequality must be adjusted to the product of the root factors, and therefore we obtain the inequality in the form of productive form

$$(x - 1)(x + 4) \geq 0$$

Zero points are -4; 1. The quadratic trinomial, after substitution $x = 0$, acquires negative value. In fact, zero is not necessary to be substituted, because for $x = 0$ is the final "a signed" value of quadratic trinomial, identical to the sign in front of the absolute member. We set the adjusted numbering axis (Fig. 3):

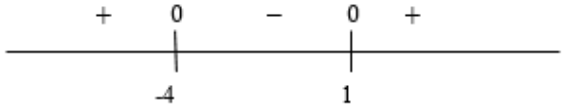

Figure 3.

Based on the sign of inequality, in assigned inequality, we search for which values of unknown x the expression $x^2 + 3x - 4$ acquires positive or zero values. Therefore, the solution is inequality is

$$x \in (-\infty; -4\rangle \cup \langle 1; \infty).$$

If the quadratic equation corresponding to the assigned inequality has less than two real roots, the method of zero points is modified. At this

modification, we primarily rely on understanding the "functioning" of zero points.

Example 4: On the set $\boldsymbol{R}$ solve the inequality $x^2 - 4x + 4 \geq 0$.
Solution: The inequality should be adjusted to productive form

$$(x - 2)(x - 2) \geq 0.$$

The left side of inequality will not be left in this form, because the students would incorrectly use the principle of alternation marks around the zero points. The expression on the left side of the inequality will be written in simplified form, and we receive the inequality

$$(x - 2)^2 \geq 0$$

Zero point is 2. Since the expression $(x - 2)^2$ is for all $x \in \boldsymbol{R}$ non-negative, number 2 is "unnecessary" zero point. Number 2 is the only zero point and so it is not needed to set the adjusted numbering axis. The solution of the inequality is $x \in \boldsymbol{R}$ and it was discovered when we were considering the zero point.

After solving example 4 we suggest a discussion on solving inequalities:

$$x^2 - 4x + 4 > 0, \qquad x^2 - 4x + 4 \leq 0\,, \qquad x^2 - 4x + 4 < 0.$$

Note: A common mistake at solving the inequality $(x - 2)^2 \geq 0$ is the extract of the root of both sides of the inequality, after which students have the wrong inequality $x - 2 \geq 0$. The following consideration can bring them to the fact, that the inequality is incorrect. Both sides of the inequality were non-negative before extracting the root and the left side can also takes negative values. If we want, even after extracting, both sides being non-negative, we must put the left side of inequality to an absolute value($\sqrt{a^2} = |a|$). After correct extracting, we get the inequality with absolute value which can also be solved by the method of zero points.

Example 5: On the set $\boldsymbol{R}$ solve the inequality $x^2 + 2x + 6 < 0$.
Solution: On the set $\boldsymbol{R}$ it is not possible to modify the quadratic trinomial to the product, as the appropriate quadratic equation

$$x^2 + 2x + 6 = 0$$

have no real roots. Based on the understanding of the function of zero points we know, that expression $x^2 + 2x + 6$ has for all x ∈ R a signed value. It is identical with the sign in front of the absolute term. So the expression on the left side of the inequality is for all real numbers positive. The solution of the inequality is $x = \{\}$. Even after solving this inequality we recommend the discussion about solutions for different variants of the sign of inequality.

If we want to see if the students understand the method, they must be able to apply the basic ideas of the method to solving the task. In other words,

we understand the method of solving if it developed our mathematical thinking. The following example can be solved by applying the basic ideas of the method of zero points.

Example 6: For which parameter values $a \in R$ is each $x \in$ R the solution of inequality

Solution: The expression $x^2 - 8x + 20$ has no zero points and according to the sign in front of the absolute member we know, that it acquires positive values for all $x \in R$. If all real numbers should be the solution of the assigned inequality, the expression in the denominator of the inequality fraction must be negative for all $x \in R$. Using the basic ideas of the method of zero points, we consider the following. We need the expression $ax^2 + 2(a+1)x + 9a + 4$ "still" negative, and that does not change the final sign, and therefore we cannot have the zero points. That is, the quadratic equation

$$ax^2 + 2(a+1)x + 9a + 4 = 0$$

has no solution. Thus, discriminant has to be negative. This way we get the inequality

$$\frac{x^2 - 8x + 20}{ax^2 + 2(a+1)x + 9a + 4} < 0$$

The solution to this inequality that we solve using the method of zero points is $a \in \left(-\infty; -\frac{1}{2}\right) \cup (1; \infty)$. Now, we secure the final sign will be negative. We know from the method of zero points, that by substituting zero to quadratic trinomial, the final sign is identical with a sign in front of the absolute term. The denominator in the assigned inequality is a quadratic trinomial with parameter. For $a_1 \in \left(-\infty; -\frac{1}{2}\right) \cup (1; \infty)$ has the constant sign for all x ∈ R. If the absolute member is negative, the resulting sign of trinomial will be negative. Therefore we solve the inequality

$$9a + 4 < 0.$$

Its solution is $a_2 \in \left(-\infty; -\frac{4}{9}\right)$. Based on the previous considerations, the parameter a must meet both conditions. The ultimate solution is $a \in a_1 \cap a_2 = \left(-\infty; -\frac{1}{2}\right)$.

Conclusion

The basis for the success of a student in solving mathematical tasks is acquiring the calculating algorithms [6], [7]. To achieve this goal it is necessary to solve, especially alone, the sufficient number of tasks, more or less, of the same type. We believe that the mastery of basic calculating algorithms is necessary but not sufficient condition for student success in

dealing with the tasks. It is not enough just to learn the calculating algorithm, it is necessary, after its acquisition, also think about its individual elements. This is the way when the basic ideas, used in the algorithm, occur. The discovering these main ideas of calculating algorithm lead to understanding, as well as acquiring the algorithm at a higher level. The understanding causes the method to be is a powerful tool in students dealing with tasks. It affects his mathematical thinking. The method of zero points is a method that should be understood and not only learned. If a student enters its secrets, it becomes flexible and he will be able to use it in different types of tasks and, as appropriate, be adapted. By understanding the method will become effective tool in the hands of the investigator. The students know that the method of zero points is mainly used to solve inequalities. If the students know the method, it heads their initial ideas, when solving inequality, to adjust the inequality to a productive or quotient form. This fact can be used in teaching solutions to quadratic inequalities. Using the method of zero points the student does not learn new calculating algorithm, but he learns how to apply already acquired knowledge and skills. We think that one of the possible ways to increase the efficiency in mathematical learning is the emphasis on understanding the calculating algorithms and their subsequent application in various areas of mathematics. While we make sure that we choose those tasks, where the main ideas can be applied. This way helps us to create the thought linking of mathematics as a whole and mathematics with other disciplines, e.g. those involving computers into the pedagogical process [8], in the mind of the students. Basically, there is no need to reduce the amount of subject matter, just to organize the mathematical knowledge better in the mind of the students.

References

[1] Bušek, I. *Řešené maturitní úlohy z matematiky*. Praha: SPN. 1985.

[2] Šedivý, O., Ďuriš, V., Fulier, J. *Konštruktivizmus vo vyučovaní matematiky*. In. Veda-vzdelávanie-prax - 2 diel: zborník z medzinárodnej vedeckej konferencie. Nitra, UKF 2007, p. 319-323, ISBN 978-80-8094-203-8.

[3] Fulier, J., Šedivý, O. *Motivácia a tvorivosť vo vyučovaní matematiky*. Nitra: Fakulta prírodných vied UKF v Nitre. 2001.

[4] Hejný, M. *Teória vyučovania matematiky 2*. Bratislava: SPN. 1991.

[5] Odvárko, O. *Metody řešení matematických úloh.* Praha: SPN. 1990.

[6] Turek, I. *Didaktika.* Bratislava: Iura Edition. 2008.

[7] Hošková-Mayerová, Š., Rosická, Z. *Programmed learning*. In: Procedia - Social and Behavioral Sciences, Vol. 31, p. 782-787, 2012 DOI: 10.1016/j.sbspro.2011.12.141.

[8] Ďuriš, V.: Prečo začleňovať počítačové programy do vyučovania. In. Technológia vzdelávania: vedecko-pedagogický časopis, ISSN 1335-003X, Vol. 6, 2005

Publisher:
Accademia Piceno – Aprutina dei Velati in Teramo (A.P.A.V.)

Periodicity:
every six months

Printed in 2019 in Pescara (Italy)

Authorisation n. 9/90 of 10/07/1990 released by Tribunale di Pescara
ISSN: 1592-7415 (printed version)

Authorisation n. 16 of 17/12/2013 released by Tribunale di Pescara
ISSN: 2282-8214 (online version)

Accademia Piceno - Aprutina dei Velati in Teramo
ACCADEMIA DI SCIENZE, LETTERE, ARTI E TECNOLOGIA

www.eiris.it – www.apav.it

Ratio Mathematica, 36, 2019

Contents

Published by Accademia Piceno - Aprutina dei Velati in Teramo (A.P.A.V.)