

Studies in Systems, Decision and Control 217

Martine Ceberio  
Vladik Kreinovich *Editors*

# Decision Making Under Uncertainty and Constraints

A Why-Book

 Springer

# **Studies in Systems, Decision and Control**

Volume 217

## **Series Editor**

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences,  
Warsaw, Poland

The series “Studies in Systems, Decision and Control” (SSDC) covers both new developments and advances, as well as the state of the art, in the various areas of broadly perceived systems, decision making and control—quickly, up to date and with a high quality. The intent is to cover the theory, applications, and perspectives on the state of the art and future developments relevant to systems, decision making, control, complex processes and related areas, as embedded in the fields of engineering, computer science, physics, economics, social and life sciences, as well as the paradigms and methodologies behind them. The series contains monographs, textbooks, lecture notes and edited volumes in systems, decision making and control spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the worldwide distribution and exposure which enable both a wide and rapid dissemination of research output.

Indexed by SCOPUS, DBLP, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

Martine Ceberio · Vladik Kreinovich  
Editors

# Decision Making Under Uncertainty and Constraints

A Why-Book

 Springer

*Editors*

Martine Ceberio  
Department of Computer Science  
University of Texas at El Paso  
El Paso, TX, USA

Vladik Kreinovich  
Department of Computer Science  
University of Texas at El Paso  
El Paso, TX, USA

ISSN 2198-4182

ISSN 2198-4190 (electronic)

Studies in Systems, Decision and Control

ISBN 978-3-031-16414-9

ISBN 978-3-031-16415-6 (eBook)

<https://doi.org/10.1007/978-3-031-16415-6>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

*I keep six honest serving-men  
(They taught me all I knew);  
Their names are What and Why and When  
And How and Where and Who.*

*Rudyard Kipling*

In the first approximation, decision making is nothing else but an optimization problem: we want to select the best alternative. This description, however, is not fully accurate: it implicitly assumes that we know the exact consequences of each decision, and that, once we have selected a decision, no constraints prevent us from implementing it. In reality, we usually know the consequences with some uncertainty, and there are also numerous constraints that need to be taken into account. The presence of uncertainty and constraints makes decision making challenging.

To resolve these challenges, we need to go beyond simple optimization, we also need to get a good understanding of how the corresponding systems and objects operate, a good understanding of why we observe what we observe—this will help us better predict what will be the consequences of different decisions. All these problems—in relation to different application areas—are the main focus of this book.

Because of this focus, we encouraged authors to include the Why word into the titles of their papers. Several authors agreed, so this book can truly be called a why-book, a true tribute to Rudyard Kipling.

Most papers from this book are extended and selected versions of papers presented at the 14th International Workshop on Constraint Programming and Decision Making CoProD'2021 (Szeged, Hungary, September 12, 2021); 26th UTEP/NMSU Workshop on Mathematics, Computer Science, and Computational Science (El Paso, Texas, November 6, 2021); and several other conferences.

We are greatly thankful to all the authors and referees, and to all the participants of the CoProd'2021 and other workshops. Our special thanks to Prof. Janusz Kacprzyk, the editor of this book series, for his support and help. Thanks to all of you!

El Paso, USA  
August 2022

Martine Ceberio  
[mceberio@utep.edu](mailto:mceberio@utep.edu)

Vladik Kreinovich  
[vladik@utep.edu](mailto:vladik@utep.edu)

# Contents

## Applications to Arts

<b>Baudelaire’s Ideas of Vagueness and Uniqueness in Art: Algorithm-Based Explanations</b> .....	3
Luc Longpré, Olga Kosheleva, and Vladik Kreinovich	

## Applications to Biosciences

<b>Selfish Gene Theory Explains Oedipus Complex</b> .....	13
Olga Kosheleva and Vladik Kreinovich	

## Applications to Education

<b>How to Teach Advanced Highly Motivated Students: Teaching Strategy of Iosif Yakovlevich Verebeichik</b> .....	19
Olga Kosheleva and Vladik Kreinovich	
<b>Why 70/100 Is Satisfactory? Why Five Letter Grades? Why Other Academic Conventions?</b> .....	25
Christian Servin, Olga Kosheleva, and Vladik Kreinovich	
<b>Shall We Ignore All Intermediate Grades?</b> .....	33
Christian Servin, Olga Kosheleva, and Vladik Kreinovich	
<b>Why <math>\infty</math> is a Reasonable Symbol for Infinity</b> .....	39
Olga Kosheleva and Vladik Kreinovich	
<b>What Is 1/0 from the Practical Viewpoint: A Pedagogical Note</b> .....	43
Olga Kosheleva and Vladik Kreinovich	
<b>Historical Diversity Through Base-10 Representation of Mayan Math</b> .....	49
Julian Viera and Olga Kosheleva	



<b>Why Base-20, Base-40, and Base-60 Number Systems?</b> .....	63
Sean R. Aguilar, Olga Kosheleva, and Vladik Kreinovich	
<b>Why Chomsky Normal Form: A Pedagogical Note</b> .....	69
Olga Kosheleva and Vladik Kreinovich	
<b>How to Best Write Research Papers: Basic English? Sophisticated English?</b> .....	75
Martine Ceberio, Christian Servin, Olga Kosheleva, and Vladik Kreinovich	
<b>Applications to Engineering</b>	
<b>How to Select Typical Objects</b> .....	83
Mariana Benitez, Jeffrey Weidner, and Vladik Kreinovich	
<b>Why Homogeneous Membranes Lead to Optimal Water Desalination: A Possible Explanation</b> .....	89
Julio Urenda, Martine Ceberio, Olga Kosheleva, and Vladik Kreinovich	
<b>Fault Detection in a Smart Electric Grid: Geometric Analysis</b> .....	93
Hector Reyes, Dillon Trinh, and Vladik Kreinovich	
<b>Applications to Geosciences</b>	
<b>Why Geological Regions?</b> .....	103
Daniela Flores, Olga Kosheleva, and Vladik Kreinovich	
<b>Applications to Machine Learning</b>	
<b>Why, in Deep Learning, Non-smooth Activation Function Works Better Than Smooth Ones</b> .....	111
Daniel Cruz, Ricardo Godoy, and Vladik Kreinovich	
<b>Why Residual Neural Networks</b> .....	117
Sofia Holguin and Vladik Kreinovich	
<b>Why Semi-supervised Learning Makes Sense: A Pedagogical Note</b> .....	121
Olga Kosheleva and Vladik Kreinovich	
<b>How to Gauge the Quality of a Multi-class Classification When Ground Truth Is Known with Uncertainty</b> .....	125
Ricardo Mendez, Osagumwenro Osaretin, and Vladik Kreinovich	
<b>An AlphaZero-Inspired Approach to Solving Search Problems</b> .....	129
Evgeny Dantsin, Vladik Kreinovich, and Alexander Wolpert	

## Applications to Physics

<b>Fuzzy Techniques, Laplace Indeterminacy Principle, and Maximum Entropy Approach Explain Lindy Effect and Help Avoid Meaningless Infinities in Physics</b> .....	141
Julio Urenda, Sean Aguilar, Olga Kosheleva, and Vladik Kreinovich	

<b>Dimension Compactification Naturally Follows from First Principles</b> .....	153
Julio C. Urenda, Olga Kosheleva, and Vladik Kreinovich	

<b>Is Our World Becoming Less Quantum?</b> .....	159
Lidice Castro and Vladik Kreinovich	

## Applications to Psychology and Decision Making

<b>As Complexity Rises, Meaningful Statements Lose Precision—But Why?</b> .....	167
Miroslav Svítek, Olga Kosheleva, and Vladik Kreinovich	

<b>Why People Overestimate Small Probabilities?</b> .....	173
David Amparan and Vladik Kreinovich	

<b>Why Ovals in Eliciting Intervals?</b> .....	177
Joshua Zamora and Vladik Kreinovich	

<b>Why Moments (and Generalized Moments) Are Used in Statistics and Why Expected Utility Is Used in Decision Making: A Possible Explanation</b> .....	181
R. Noah Padilla and Vladik Kreinovich	

<b>Decision Making Under Uncertainty: Cases When We Only Know an Upper Bound or a Lower Bound</b> .....	189
Toshiki Kamio, Gavin Baechle, and Vladik Kreinovich	

<b>Why Do People Become Addicted: Towards a Theoretical Explanation for Eyal's Experiment-Based Hook Model</b> .....	193
Christopher Reyes and Vladik Kreinovich	

<b>Why Decimal System? Why Communities with More Than 150 Folks Tend to Split? New Consequences of the Seven Plus Minus Two Law</b> .....	201
Leonardo Orea Amador and Vladik Kreinovich	

<b>Lev Landau's Marital Advice Explained</b> .....	207
Olga Kosheleva and Vladik Kreinovich	

<b>Why Too Much Interaction Between Different Parts of the Brain Leads To Unhappiness</b> .....	211
Ricardo Alvarez, Yamel Hernandez, and Vladik Kreinovich	

**Applications to Religion**

**Gödel’s Proof of Existence of God Revisited** ..... 217  
Olga Kosheleva and Vladik Kreinovich

**Blessings, God, Sacrifices: Possible Rational Explanations of Biblical Ideas** ..... 223  
Olga Kosheleva and Vladik Kreinovich

**General Computational Aspects**

**Why Model Order Reduction** ..... 233  
Salvador Robles, Martine Ceberio, and Vladik Kreinovich

**Bounding the Range of a Sum of Multivariate Rational Functions** ..... 239  
Mohammad Adm, Jürgen Garloff, Jihad Titi, and Ali Elgayar

**Fourier Transform and Other Quadratic Problems Under Interval Uncertainty** ..... 251  
Oscar Galindo, Christopher Ibarra, Vladik Kreinovich, and Michael Beer

**B-Matrices and Doubly B-Matrices in the Interval Setting** ..... 257  
Matyáš Lorenc

**Commonsense “And”-Operations** ..... 285  
Javier Tellez, Wenbo Xie, and Vladik Kreinovich

**Extension to Multidimensional Problems of a Fuzzy-Based Explainable and Noise-Resilient Algorithm** ..... 289  
Javier Viaña, Stephan Ralescu, Kelly Cohen, Anca Ralescu, and Vladik Kreinovich

**Additional Spatial Dimensions Can Help Speed Up Computations** ..... 297  
Luc Longpré, Olga Kosheleva, and Vladik Kreinovich

# **Applications to Arts**

# Baudelaire's Ideas of Vagueness and Uniqueness in Art: Algorithm-Based Explanations



Luc Longpré, Olga Kosheleva, and Vladik Kreinovich

**Abstract** According to the analysis by the French philosopher Jean-Paul Sartre, the famous French poet and essayist Charles Baudelaire described (and followed) two main—somewhat unusual—ideas about art: that art should be vague, and that to create an object of art, one needs to aim for uniqueness. In this paper, we provide an algorithm-based explanation for these seemingly counter-intuitive ideas, explanation related to Kolmogorov complexity-based formalization of Garrett Birkhoff's theory of beauty.

## 1 Formulation of the Problem

**Baudelaire's ideas about art.** In this book [32] about the famous 19 century French poet and essayist Charles Baudelaire, Jean-Paul Sartre emphasizes the following two somewhat unusual aspects of Baudelaire's attitude to art.

The first aspect is explicit in Baudelaire's essays: vagueness. In a well-studied passage of his book *Fusées*, Baudelaire defines beautiful as "Something a little vague, which leaves room for conjecture". This may sound almost trivial now, after the Impressionists changed our understanding of Beauty, but in Baudelaire's time, when

---

L. Longpré · V. Kreinovich (✉)

Department of Computer Science, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA

e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

L. Longpré

e-mail: [longpre@utep.edu](mailto:longpre@utep.edu)

O. Kosheleva

Department of Teacher Education, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA

e-mail: [olgak@utep.edu](mailto:olgak@utep.edu)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

M. Ceberio and V. Kreinovich (eds.), *Decision Making Under Uncertainty*

and *Constraints*, Studies in Systems, Decision and Control 217,

[https://doi.org/10.1007/978-3-031-16415-6\\_1](https://doi.org/10.1007/978-3-031-16415-6_1)

beauty was still mostly measured by the Renaissance giants such as Leonardo da Vinci or Rafael, with their highly realistic details, this was definitely an almost heretical thought.

The second aspect is not so explicit, but can also be traced to many of his essays and letters: uniqueness, that in order to create someone beautiful, one needs to create something truly unique, repetition is an antithesis of beauty. This also sounds somewhat heretical: there seems to be often a lot of similarity between several beautiful paintings.

**A natural question.** How can we explain these ideas?

**What we do in this paper.** In this paper, we show that actually, both seemingly counterintuitive ideas can be explained within a proper algorithm-based formalization of what is beautiful and how can we design a beautiful object.

## 2 What Is Beauty—Birkhoff’s Approach and Its Algorithm-Related Formalization

**Birkhoff’s approach.** According to the theory developed by the 20th century mathematician Garrett D. Birkhoff—one of the founding fathers of lattice theory—beauty  $B$  can be described as the ratio

$$B = \frac{O}{C} \quad (1)$$

between properly defined order  $O$  and complexity  $C$ ; see, e.g., [3–9]. In the simplest cases, he formalized these notions—and showed that his formula is indeed working.

In his examples:

- Birkhoff defined complexity  $C$  as the number of construction steps needed to construct the given object, and
- he defined order as a simplicity of the description: if we can describe an object by using a shorter text, then its order is higher.

**Birkhoff’s approach reformulated in general algorithmic terms.** Birkhoff’s theory appeared before the general development of algorithm theory. Now that we are accustomed to the notion of algorithms, it is natural to reformulate his theory in precise algorithmic terms. In these terms, the number of construction steps simply becomes the number of computational steps—i.e., the computation time  $t(p)$  of the algorithm  $p$  that generates the given object.

The notion of order is a little more difficult to formalize. In his examples, by a description of the objects, Birkhoff meant a complete description, i.e., a description which is detailed enough so that, given this description, we can uniquely reconstruct the object. In other words, the description can serve as a program for a computational device which, given this description, reconstructs the object. In these terms, the length of the description is equal to the length  $\ell(p)$  of this program  $p$ .

In these terms, the beauty  $B$  of an object should be a function of the time  $t(p)$  and the length  $\ell(p)$  of a program  $p$  that generates this object:  $B = B(t(p), \ell(p))$ . It is well known in computer science that there is a trade-off between the program time and the program length. A short program usually uses only a few ideas of speeding up computations, and thus, takes a reasonable amount of time to run. If we want to speed up the computations, we must add some complicated ideas and modify the algorithm. As a result, to make the program faster, we must usually make it longer. Vice versa, we can often shorten the program by eliminating some of the time-saving parts and thus, by making its running time longer.

In general, if we cut a bit from the program that generates the object  $x$ , we get a new program  $p'$  which is exactly one bit shorter ( $\ell(p') = \ell(p) - 1$ ). To generate the desired object  $x$ , since we do not know whether the deleted bit was 0 or 1, we can try both possible values of this bit (i.e., run two programs  $p'0$  and  $p'1$ ) and find out which of the two objects is better. Thus, if we delete a bit, then instead of running the original program  $p$  once, we run two programs  $p'0$  and  $p'1$ . Hence, crudely speaking, when we decrease the length of the program by 1, we thus get a double increase in the running time:  $t(p') = 2t(p)$ .

The new situation is, in effect, the same, the resulting object is the same, the only difference is that we now have  $\ell(p') = \ell(p) - 1$  and  $t(p') = 2t(p)$ . It is therefore reasonable to require that the beauty value  $B(t, \ell)$  does not change under this transformation, i.e., that for all possible values of  $t$  and  $\ell$ , we have

$$B(t, \ell) = B(2t, \ell - 1). \quad (2)$$

It can be shown (see, e.g., [24]) that every function satisfying this property can be described as a function of the following ratio:

$$r(p) \stackrel{\text{def}}{=} \frac{2^{-\ell(p)}}{t(p)}. \quad (3)$$

Thus, the beauty of the object can be described as the largest possible value of the ratio (2) over all the programs  $p$  that generate this object.

**Is this an adequate formalization?** The ratio (3) is in perfect accordance with Birkhoff's formula (1):

- the time  $t(p)$  is exactly what Birkhoff meant by complexity and
- the numerator  $2^{-\ell(p)}$  is a decreasing function of the program's (thus description's) length—in perfect accordance with Birkhoff's idea of order.

**How this formalization is related to other algorithmic notions.** Maximizing the ratio (3) is equivalent to minimizing its inverse  $t(p) \cdot 2^{\ell(p)}$  and to minimizing the binary logarithm  $\ell(p) + \log_2(t(p))$ . From this viewpoint, the beauty of an object is related:

- to the notion of *Kolmogorov complexity*—which is defined as the length of the shortest possible program that generates the given object [28], and

- to resource-bounded versions of Kolmogorov complexity [28] that minimize a combination of the program's length and time.

In this sense, Birkhoff's beauty can be viewed as a particular variant of the resource-bounded Kolmogorov complexity.

### 3 How This Explains the Need for Vagueness

**What is vagueness.** Birkhoff's definition is usually applied to abstract objects. However, many objects of art describe real-life objects and/or events: e.g., a painting can reflect a person or a landscape, a poem can describe some events and/or feelings, etc.

Real-life objects can be reproduced with different number of details. For example, we can have a photograph that captures all the details of an object, or we can have a blurred image or even a silhouette, where only some features are reproduced and many details are missing. This is exactly what is meant by vagueness—that some details are missing.

**Why is vagueness important for beauty.** For each object of art  $a$ , we can define its beauty  $B(a)$  as the largest possible value of the ratio (3) over all programs that generate this object.

For the same original real-life object  $x$ , for reproductions  $x_v$  corresponding to different levels of vagueness  $v$ , we have, in general different value of beauty  $B(x_v)$ . If our goal is to make the most beautiful object of art, we should select the level  $v$  for which the corresponding beauty  $B(x_v)$  is the largest possible.

There are many possible levels; let us denote this number by  $L \gg 1$ . A priori, we have no reason to assume that one of these levels is more susceptible to beauty: we can enjoy Leonardo's madonnas with lots of detail, and we can enjoy impressionistic painting where most details are missing. Since we do not have any reason to believe that one of these levels is more probable as the most beautiful one, it is reasonable to conclude that each of these levels is equally probable to be the most beautiful one; this reasoning goes back to the 18–19 centuries' mathematician Pierre-Simon Laplace—one of the founders of probability theory—and is therefore known as Laplace's Indeterminacy Principle; see, e.g., [14]. So, each of the  $L$  levels has the same probability  $1/L$  to be the most beautiful.

In particular, this means that only with probability  $1/L \ll 1$ , the most beautiful level is the level of all the details. In all other cases, the most beautiful level corresponds to some vagueness—which explains Baudelaire's observation that in the overwhelming majority of cases, vagueness is an important attribute of beauty.



## 4 Why Uniqueness: An Algorithmic Explanation

**We want the most beautiful representation of a real-life object.** As we have mentioned earlier, there are many possible representations of an object. Our goal is to select the most beautiful representation.

In abstract terms, our goal is to select a representation  $a$  that maximizes the corresponding beauty  $B(a)$ .

**Let us analyze this problem from the algorithmic viewpoint.** In contrast to science—that studies objects that already exist—art is about creating new objects. So, it makes sense to think of algorithms that can help in this creation.

Art can reflect everything, so the corresponding optimization problems are very generic. In general, the problem of finding the object that maximizes a given computable function is not algorithmically solvable (see, e.g., [1, 10–13, 23, 26, 31]), but there is an important case when, under some reasonable condition, the corresponding algorithm is possible: the case when there is exactly *one* optimizing object; see, e.g., [15–23, 25, 27, 29, 30].

Interestingly, if we consider all the cases when there are *two* equally good optimizing objects, such an algorithm is no longer possible; see, e.g., [21–23, 30]. In this sense, the case of uniqueness is the most general case we can consider if we want our problems to be algorithmically solvable.

*Comment.* There is also some evidence that even when the algorithms for the multi-optima case are possible, in general, algorithms corresponding to the single-optimum case are more efficient; see, e.g., [2]; see, however, [33].

**Conclusion.** Thus, if we want to actually *create* a beautiful artistic reflection of a given real-life object or situation, a natural idea is to impose additional restrictions that would make the optimal reflection unique. This is exactly what Sartre described as one of the main Baudelaire's ideas. Thus, this idea is indeed explained.

**Acknowledgements** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes). It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

## References

1. Beeson, M.J.: Foundations of Constructive Mathematics. Springer, N.Y. (1985)
2. Beigel, R., Buhrman, H., Fortnow, L.: NP might not be as easy as detecting unique solutions. In: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing (STOC), pp. 203–208 (1998)
3. Birkhoff, G.D.: A mathematical approach to aesthetics. *Scientia* **50**, 133–146 (reprinted in Ref. [8], vol. 3, pp. 320–333)

4. Birkhoff, G.D.: A mathematical theory of aesthetics. Rice Institute Pamphlet, vol. 19, pp. 189–342 (1932) (reprinted in Ref. [8], vol. 3, pp. 382–535)
5. Birkhoff, G.D.: Aesthetic Measure. Harvard University Press, Cambridge, Massachusetts (1933)
6. Birkhoff, G.D.: Three public lectures on scientific subjects. Rice Institute Pamphlet, vol. 28, pp. 1–76 (1941) (reprinted in Ref. [8], vol. 3, pp. 755–777)
7. Birkhoff, G.D.: Mathematics of aesthetics. In: Newman, J.R. (ed.) The World of Mathematics, vol. 4, pp. 2185–2208. Simon and Schuster, New York (1956)
8. Birkhoff, G.B.: Collected Mathematical Papers. Dover, New York (1960)
9. Birkhoff, G.: Mathematics and psychology. SIAM Rev. **11**(4), 429–467 (1969)
10. Bishop, E.: Foundations of Constructive Analysis. McGraw-Hill (1967)
11. Bishop, E., Bridges, D.S.: Constructive Analysis. Springer, N.Y. (1985)
12. Bridges, D.S.: Constructive Functional Analysis. Pitman, London (1979)
13. Bridges, D.S., Vita, S.L.: Techniques of Constructive Analysis. Springer, New York (2006)
14. Jaynes, E.T., Bretthorst, G.L.: Probability Theory: The Logic of Science. Cambridge University Press, Cambridge, UK (2003)
15. Kohlenbach, U.: Theorie der majorisierbaren und stetigen Funktionale und ihre Anwendung bei der Extraktion von Schranken aus inkonstruktiven Beweisen: Effektive Eindeutigkeitsmodule bei besten Approximationen aus ineffektiven Eindeutigkeitsbeweisen, Ph.D. dissertation, Frankfurt am Main (1990)
16. Kohlenbach, U.: Effective moduli from ineffective uniqueness proofs. An unwinding of de La Vallée Poussin’s proof for Chebycheff approximation. Annals Pure Appl. Logic **64**(1), 27–94 (1993)
17. Kohlenbach, U.: Applied Proof Theory: proof Interpretations and their Use in Mathematics. Springer, Berlin-Heidelberg (2008)
18. Kreinovich, V.: Uniqueness implies algorithmic computability. In: Proceedings of the 4th Student Mathematical Conference, pp. 19–21. Leningrad University, Leningrad (1975) (in Russian)
19. Kreinovich, V.: Reviewer’s remarks in a review of D.S. Bridges. In: Constructive Functional Analysis, Pitman, London (1979); *Zentralblatt für Mathematik*, vol. 401, pp. 22–24 (1979)
20. Kreinovich, V.: Categories of space-time models, Ph.D. dissertation, Novosibirsk, Soviet Academy of Sciences, Siberian Branch, Institute of Mathematics (1979) (in Russian)
21. Kreinovich, V.: Philosophy of Optimism: notes on the Possibility of using algorithm theory when describing historical processes, Leningrad Center for New Information Technology “Informatika”. Technical Report, Leningrad (1989) (in Russian)
22. Kreinovich, V.: Physics-motivated ideas for extracting efficient bounds (and algorithms) from classical proofs: beyond local compactness, beyond uniqueness. In: Abstracts of the Conference on the Methods of Proof Theory in Mathematics, p. 8. Max-Planck Institut für Mathematik, Bonn, Germany (2007)
23. Kreinovich, V., Lakeyev, A., Rohn, J., Kahl, P.: Computational Complexity and Feasibility of Data Processing and Interval Computations. Kluwer, Dordrecht (1998)
24. Kreinovich, V., Longpré, L., Koshelev, M.: Kolmogorov complexity, statistical regularization of inverse problems, and Birkhoff’s formalization of beauty. In: Mohamad-Djafari, A. (ed.) Bayesian Inference for Inverse Problems, Proceedings of the SPIE/International Society for Optical Engineering, vol. 3459, pp. 159–170, San Diego, CA (1998)
25. Kreinovich, V., Villaverde, K.: Extracting computable bounds (and algorithms) from classical existence proofs: Girard domains enable us to go beyond local compactness. Int. J. Intell. Technol. Appl. Stat. (IJITAS) **12**(2), 99–134 (2019)
26. Kushner, B.A.: Lectures on constructive mathematical analysis. American Mathematical Society Providence, Rhode Island (1984)
27. Lacombe, D.: Les ensembles récursivement ouvert ou fermés, et leurs applications à l’analyse récursive. Comptes Rendus de l’Académie des Sci. **245**(13), 1040–1043 (1957)
28. Li, M., Vitányi, P.: An Introduction to Kolmogorov Complexity and Its Applications. Springer, Berlin, Heidelberg, New York (2008)

29. Lifschitz, V.A.: Investigation of constructive functions by the method of fillings. *J. Sov. Math.* **1**, 41–47 (1973)
30. Longpré, L., Kreinovich, V., Gasarch, W., Walster, G.W.:  $m$  solutions good,  $m - 1$  solutions better. *Appl. Math. Sci.* **2**(5), 223–239 (2008)
31. Pour-El, M., Richards, J.: *Computability in Analysis and Physics*. Springer, New York (1989)
32. Sartre, J.-P.: *Baudelaire*. New Direction Publ, New York (1967)
33. Valiant, L.G., Vazirani, V.V.: NP is as easy as detecting unique solutions. *Theor. Comput. Sci.* **47**, 85–93 (1986)

# **Applications to Biosciences**

# Selfish Gene Theory Explains Oedipus Complex



Olga Kosheleva and Vladik Kreinovich

**Abstract** Sigmund Freud famously placed what he called Oedipus complex at the center of his explanation of psychological and psychiatric problems. Freud's analysis was based on anecdotal evidence and intuition, not on solid experiments—as a result, for a long time, many psychologists dismissed the universality of Freud's findings. However, lately, experiments seem to confirm that indeed men, on average, select wives who resemble their mothers, and women select husbands who resemble their fathers. In this paper, we provide a possible biological explanation for this observational phenomenon.

## 1 Oedipus Complex: A Brief Reminder

**What is Oedipus complex.** From his experience with patients, Sigmund Freud discovered that several of his male patients experienced subconscious hostility towards their fathers and subconscious sexual feelings towards their mothers; see, e.g., [4, 5]. Since his attention was attracted to these unexpected feelings, he started searching for such feelings in other patients and found such feeling in most of his patients—as well as in several healthy folks whom he analyzed. So, he came to a conclusion that such feelings are universal, starting with early childhood.

Freud called such feelings Oedipus complex [4], after the legendary King Oedipus who killed his own father—not knowing that this was his father, and married his own mother—again, not knowing that this lady was his mother.

---

O. Kosheleva

Department of Teacher Education, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA

e-mail: [olgak@utep.edu](mailto:olgak@utep.edu)

V. Kreinovich (✉)

Department of Computer Science, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA

e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

Psychologists also discovered similar feelings in females: they have a subconscious hostility towards their mothers and subconscious sexual feelings towards their fathers.

**Modern attitude.** Freud's ideas were based on anecdotal evidence—as many other ideas in the psychology of his time. Later, psychology became more of an experimental science, many anecdotal-based ideas turned out to be not fully supported by the evidence. As a result, many psychologists summarily dismissed these ideas—and the universality of Oedipus complex was one of the ideas that many psychologists dismissed.

However, later experimental studies provided support to the Oedipus complex idea: for example, it turned out that, statistically, a person's wife is more similar in appearance to the person's mother than an average woman from his region; see, e.g., [1, 7, 8].

**How can we explain this experimental evidence.** Freud himself provided a sociological explanation—that in ancient times, sometimes, the only way for young people to get food and women was to kill their fathers, and that we still keep, to some extent, this murderous instinct, just like many we keep many other aspects of behavior—like fight or flight body reaction to dangers—even though they mostly make no sense at present.

This explanation is as speculative as the original idea of the Oedipus complex. It is desirable to have a more solid explanation for the Oedipus complex phenomenon. Such an explanation is provided in this paper.

## 2 Our Explanation

**Selfish gene theory: the basis for our explanation.** In our explanation, we will use the *selfish gene* theory; see, e.g., [3, 9]. According to this theory, each gene wants to spread as much as possible.

**From selfish gene theory to an explanation: an idea.** From the viewpoint of the selfish gene theory, what is the best partner for a man so that their children preserve as many of the man's genes as possible?

**Seemingly ideal spouse.** In general, in a child, half of the genes come from the father, and half from the mother. So, the ideal situation when all 100% of the genes are preserved in all the children is when the wife is genetically identical to the husband. In this case, children get exactly the same genes as both parents.

**Problem with this seemingly ideal idea.** But how does the body know about the genes? The only way for a body to see how close are the genes is to rely on the fact that people with similar genes look similar and have other similar characteristics such as smell, voice, etc., i.e., have similar *phenotype*—observable appearance.

The problem is that there is a big difference in appearance between men and women, so even when a man and a woman have similar genes, they look differently.

Thus, when a man selects a wife—and a woman selects a husband—they cannot rely on similarity in appearance to decide which potential partner is the closest to them genes-wise.

**We need to look for similarity between people of the same sex.** To make such a decision, a man needs to have some other pattern (not himself) to whom to compare his future wife—this must be a woman who has as many genes in common as this man.

**Two options.** Here, we have two options.

- First, since a man inherits half of his genes from his mother, so the mother has 50% genes in common with her son—we mean the genes that vary from person to person; of course, the vast majority of the genes are common to all of us—these are the genes that make human beings and no fish or apes.
- Another female relative who, in principle, shares half of the genes with a man is his sister. All other relatives share 25% and less of the genes.

From this viewpoint, a man should look for a wife who resembles either his mother or his sister.

**Mother or sister.** Shall he select a mother or a sister?

- With a mother, there is a guarantee that she shares 50% of the genes.
- In contrast, for a sister, the actual percentage may be lower. It is known that even in monogamous animals, there is a significant percentage of children whose father is different from the permanent partner; see, e.g., [2, 6].

From this viewpoint, on average, a sister shares fewer genes. Thus, a mother is a more reliable pattern to whom men can compare future wives.

**This explains the experimental Oedipus effect for men.** So, our conclusion is that from the biological viewpoint, it is beneficial for men to look for wives who resemble their mothers—this is exactly what the experiments show.

**What about women?** Similar arguments show that for a woman, the best chance of preserving as many genes as possible in their children is to select a husband who is similar to their father.

**But why now take Oedipus's example literally?** But why do we say “similar” to a person's mother or father? Why not identical?

At first glance, it may look like that from this viewpoint, the best idea is to simply have joint children with your own parents, exactly as Oedipus himself did. The reason why we cannot replace “similar” with “identical” in our conclusions is that, as is well known, such an incest (and even more remote incest) is very damaging to the genes.

In addition to good genes, there are also not so good genes causing all kinds of diseases. Many of these genes only become active when they are inherited from both parents. When spouses are not previously related, the probability that both have a copy of such a damaging gene is low. However, when they are closely related and their

genes are similar, the probability becomes high—and diseases start. Experiments on animals—when practitioners try to match close relatives to come up with the fastest horse or the most productive cow—show that incest (which for animals is called inbreeding) indeed leads to a wide spread of damaging fluctuations and degeneration; see, e.g., [10].

This is a simple biological explanation of why marrying your own parent is a undesirable pathology, but marrying someone who somewhat resembles you own parent is a healthy common practice.

**Acknowledgements** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

## References

1. Bereczkei, T., Gyuris, P., Weisfeld, G.E.: Sexual imprinting in human mate choice. In: Proc. R. Soc. London. Ser. B: Biol. Sci. **271**(1544), 1129–1134 (2004)
2. Black, J.: Partnerships in Birds: the Study of Monogamy. Oxford University Press, Oxford, UK (1996)
3. Dawkins, R.: The Selfish Gene. Oxford University Press, Oxford, UK (2016)
4. Freud, S.: A special type of choice of object made by men. In: The Collected Works of Sigmund Freud. Pergamon Media, Oxford, UK (2015)
5. Freud, S.: The Interpretation of Dreams. Wiley, Chichester, UK (2020)
6. Jones, M.G.W., Techow, N.M.S.M., Ryan, P.G.: Dalliances and doubtful dads: what determines extra-pair paternity in socially monogamous wandering albatrosses? *Behav. Ecol. Sociobiol.* **66**, 1213–1224 (2012)
7. Khan, M.: Basic Freud: psychoanalytic Thought for the 21st Century, pp. 59–60. Basic Books, New York (2002)
8. Petter, O.: We seek romantic partners who look like our parents, finds study. *The Independent* (2017)
9. Williams, G.C.: *Adaptation and Natural Selection: a Critique of Some Current Evolutionary Thought*. Princeton University Press, Princeton, New Jersey (2018)
10. Wolf, A.P., Durham, W.H.: *Inbreeding, Incest, and the Incest Taboo: the State of Knowledge at the Turn of the Century*. Stanford University Press, Stanford, California (2004)



# **Applications to Education**

# How to Teach Advanced Highly Motivated Students: Teaching Strategy of Iosif Yakovlevich Verebeichik



Olga Kosheleva and Vladik Kreinovich

**Abstract** The paper describes and explains the teaching strategy of Iosif Yakovlevich Verebeichik, a successful mathematics teacher at special mathematical high schools—schools for students interested in and skilled in mathematics. The resulting strategy seems counterintuitive and contrary to all the pedagogical advice. Our explanation is not complete: it worked well for this teacher, but others who tried to follow seemingly the same strategy did not succeed. How he made it work, how can others make it work—this is still not clear. In the words of Verebeichik himself, while mathematics itself is a science, teaching mathematics is an art, which cannot be reduced to a few recommendations.

## 1 Introduction

**Who was Vereberichik.** Iosif Yakovlevich Verebeichik was a teacher of mathematics in St. Petersburg, Russia, who taught in special *mathematical high schools* where a special emphasis was made on mathematics, most of the time in School No. 30, where one of the authors (VK) studied under his guidance. Among mathematics teachers from such schools, he was one of the most successful—his students regularly won prizes at local and national olympiads for high school students, and after graduation, were regularly accepted into highly competitive university programs; see, e.g., [1].

Not only he was successful in teaching students mathematics, he also managed to make them feel good. Most of his students adored him—although not everybody. For most of his students, he was their favorite teacher.

---

O. Kosheleva

Department of Teacher Education, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA

e-mail: [olgak@utep.edu](mailto:olgak@utep.edu)

V. Kreinovich (✉)

Department of Computer Science, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA

e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

19

M. Ceberio and V. Kreinovich (eds.), *Decision Making Under Uncertainty*

*and Constraints*, Studies in Systems, Decision and Control 217,

[https://doi.org/10.1007/978-3-031-16415-6\\_3](https://doi.org/10.1007/978-3-031-16415-6_3)

**What was the secret of his success?** Many folks—his students, other teachers, journalists—often asked him the same question: what is the secret of his success as a teacher? How can other teachers become more successful? Vebereichik was willing to help others, he gladly allowed interested teachers to attend his classes—and many teachers benefited from this experience. However, he was not able to formulate what exactly he is doing differently.

In replies to such a question, he always emphasized that teaching is an art—just as it is difficult to explain why some music affects us and some does not, it is not easy to explain why some teaching ideas work better. As a result, in contrast to many other successful teachers, he did not leave a description of his teaching strategy.

**But why?** Now that many of his students became teachers themselves—in schools, in universities, etc.—our minds go back to Verebeichik’s success. We all ask ourselves the same questions: What was his secret? How can we use his teaching techniques in our own teaching?

**Why now?** This year, we celebrated the 100th anniversary of Verebeichik’s birth. Many of his students shared their memories about him. Naturally, the question of why was raised again and again.

We think we found some explanation—at least the explanation for some of the features of his teaching that we all remember. In this paper, we tried to provide this explanation.

This is a first attempt, maybe others can dig deeper and find other explanations of these and other features of his teaching—we would welcome that.

## 2 How Mathematics (And Other Disciplines) is Usually Taught: A Brief Reminder

**Why do we need this reminder.** In order to understand what Verebeichik did, let us recall how mathematics (and other high school disciplines) is usually taught.

**Classwork and homework.** Most material is studied in class. Usually, for each topic:

- the teacher describes the main ideas,
- then the teacher shows, in detail, how to solve typical problems,
- after that, the teacher asks students to solve similar problems in class.

After that, other similar problems are assigned as homeworks. The amount of assigned homework is reasonable, so that students can maintain a healthy work-life balance.

There is usually a textbook that describes all this in detail. So, if something is not clear, students can always look into the textbook and clarify their understanding.

**Tests and quizzes.** To gauge the student knowledge, students take several tests and quizzes. Usually, the problems given on these tests and quizzes are similar to what the students saw in class and on the homeworks.

**How homeworks, tests, and quizzes are graded.** Usually, if a student worked reasonably hard, this student gets a perfect grade (A in the US system, which corresponds to 5 in the Russian grading system). Students who do not work as hard as required get corresponding lower grades.

**Praise, praise, and praise.** In accordance with pedagogical advice, students are always praised for what they have done, criticisms are limited to a necessary minimum and packaged in the most nice way—e.g., “sandwiched” better two positive statements.

### 3 What Verebeichik Wanted and What He Therefore Did

**Specifics of Verebeichik’s students.** The above-described traditional approach to teaching works well for many students. However, the students in mathematical school are different from the average students: they clearly have better abilities to do mathematics. In their previous schools, they easily got As in math without making too much effort.

**What Verebeichik wanted.** The main objective of Verebeichik—and of other math teachers in the mathematical school—was to motivate the students to work harder, to unveil their full potential in math. Not all the students became professional mathematicians, but all the students learned much more in this high school than their friends in regular schools.

From the viewpoint of this objective, let us look again at how classes are usually taught, and let us analyze what needs to be changed to better motivate the students in mathematical school. Interestingly, we arrive at exactly the techniques that Verebeichik used.

**Classwork.** In a regular class, when explaining a new material, the teacher explains, in detail, how to solve several typical problems. To make students think harder, instead of explaining the solution to such a problem, a natural idea is to have students come up with a solution—Socratic way, with a series of hints.

Another trick is: once the main ideas are clear, instead of explaining all the details to the class, let the students themselves come up with these details on their own—as a result, practically no solution was explained in detail in class (but of course, a detailed solution is required on the homeworks!).

**Homeworks.** In a regular class, the amount of homework is reasonable, to maintain the work-life balance. Of course, for students who have special math abilities, this “reasonable” amount is larger than for a general student. But how larger? All students in the class have higher math skills, but in this, they are not equal:

- some of them are future (or even past) winners of national olympiads,
- some are simply somewhat better in math than an average high school student.

If we set up the amount of homework based on the students who are somewhat better, then the best students would not reveal their full potential. No matter where we place the threshold, if there are students who can easily do this amount of homework, these students will not reveal their full potential.

A natural solution—which, without the above explanation, sounds very counter-intuitive—is to assign an *unreasonable* amount of homework, so that no student will be able to do all of it.

**Do we need a textbook?** Following a textbook makes studying easier. So, a way to make students work harder is to follow some unusual path to each topic, a path which is not reflected in any well-designed textbooks.

**Tests and quizzes.** In traditional pedagogy, tests and quizzes contain:

- a reasonably small number of problems, and
- these problems are similar to what was studied in class and what was done in the homework.

To make it more challenging, natural ideas are:

- to give a high number of problems, and
- to make sure that these problems are somewhat different from what was previously studied.

*Comment.* Of course, when we drastically increase the number of problems, the time needed for grading also increases—or, if it is an oral exam, the time for asking questions and listening for the answers also increases. For grading written exams and for asking questions on oral exams, students from a higher class (or alumni) who have already studied this topic are asked to help.

They help willingly, first, because they were similarly helped by other students, and two, because this way, they recall this material and learn it better.

**How homeworks, tests, and quizzes are graded.** In the usual teaching practice, a student who works reasonably hard gets an excellent (A) grade. Of course, in a mathematical school, where students' abilities are higher, the threshold for A should be higher. But here we encounter the same problem as with determining the amount of homeworks to assign: whatever threshold we set, students who can easily solve that many problems on the text will not reveal their full potential.

A solution is the same as with homework: make this threshold unrealistically high, so that most (or even all) students will get at most B (which is 4 in the Russian system).

*Comment.* Of course, it is desirable not to ruin the students' Grade Point Average (GPA)—which is important for admission to universities, etc. So, this tough grading is only done in the beginning; after that, students already get in the habit of working hard.

**Praise?** In the traditional teaching practice, a teacher tries to praise the students as much as possible. One of the reasons for this is that praise provides an additional motivation for students—in addition to grades.

The problem with this is that once a student is praised for his/her achievement, this student is less motivated to do better. To avoid this slow-down, and to make sure that all the students work as hard as they can, praise is minimized, and criticisms become more bare.

## 4 Let Us Summarize

**Summary.** According to our analysis, the best strategy for a teacher in mathematical school is as follows:

- instead of explaining the topic, use Socratic method: give hints so that the students themselves come up with the ideas;
- for all examples, provide the main ideas, but never all the details;
- assign an unreasonably large amount of homeworks, so that no student will be able to do all of them;
- select a way of presenting each topic which is not described in detail in any textbook;
- on each test and quiz, assign many problems, and make sure that they are somewhat different from the types of the problems the students had earlier;
- grade the homeworks, tests, and quizzes in such a way that practically no one gets a perfect grade.

**Are we serious?** When formulated this way, what we have described is a monster teacher who violates all known pedagogical principles. Based on this description, students should hate this teacher, and the school principal should fire him/her on the spot. And still we loved Vereberichik—and, by the way, hated some teachers who tried to follow seemingly the same approach. Why?

The only answer we can give here is to repeat Vebereichik's statement that teaching is largely art: there was something in Vereberichik's personality that allowed us to accept his teaching style—while other teachers who lacked this “something” were not that successful. What is this “something”—maybe someone can find out.

**Acknowledgements** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

## Reference

1. Kreinovich, V., (ed.), Special Section in memory of I. Ya. Verebeichik. *J. Uncert. Syst.* **1**(4), 243–290 (2007); In: *Memoriam*, p. 243

# Why 70/100 Is Satisfactory? Why Five Letter Grades? Why Other Academic Conventions?



Christian Servin, Olga Kosheleva, and Vladik Kreinovich

**Abstract** Why 70/100 is usually a threshold for a student's satisfactory performance? Why there are usually only five letter grades? Why the usual arrangement of research, teaching, and service is 40-40-20? We show that all these arrangements—and other similar academic arrangements—can be explained by two ideas: the Laplace Indeterminacy Principle and the seven plus minus two law.

## 1 Why 70/100 Is Satisfactory?

**Formulation of the problem.** In the standard US teaching arrangement, about 70 points out of 100 means a satisfactory grade—less than that is failing.

A similar proportion works well outside the academic world: e.g., at Google, if you have fulfilled 70% of your annual goals, this is considered to be a satisfactory performance.

Since this arrangement is actively used for a long time, it probably reflects the intuitive idea of a satisfactory learning level. But a natural question remains: how can we explain this empirical fact—that namely 70/100 is the satisfactory threshold?

**What is satisfactory: intuitive idea.** Crudely speaking, satisfactory means that the amount of the course material that the student knows is (significantly) larger than

---

C. Servin

Information Technology Systems Department, El Paso Community College (EPCC), 919 Hunter Dr., El Paso, TX 79915-1908, USA

e-mail: [cservin1@epcc.edu](mailto:cservin1@epcc.edu)

O. Kosheleva

Department of Teacher Education, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA

e-mail: [olgak@utep.edu](mailto:olgak@utep.edu)

V. Kreinovich (✉)

Department of Computer Science, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA

e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

25

M. Ceberio and V. Kreinovich (eds.), *Decision Making Under Uncertainty*

*and Constraints*, Studies in Systems, Decision and Control 217,

[https://doi.org/10.1007/978-3-031-16415-6\\_4](https://doi.org/10.1007/978-3-031-16415-6_4)



the amount of the course material that the student does not know. Equivalently, the amount of the course material that the student does not know is (much) smaller than the amount of the course material that the student knows.

**We need to formalize this idea.** If we did not have the imprecise words “significantly” and “much”, the formalization would be very straightforward: the proportion  $k$  of the course material that the student knows should be larger than the proportion  $d = 1 - k$  of the course material that the student does not know:  $d < k$ . This would mean that the threshold would be 50/100. However, while, e.g., 0.51 is larger than  $1 - 0.51 = 0.49$ , one cannot say that it 0.51 significantly larger than 0.49.

Yes, 0.49 is smaller than 0.51, but, intuitively, 0.49 is not a meaningful representative of numbers which are smaller than 0.51. If you ask a person to name a typical representative of numbers which are smaller than 0.51, it is highly improbable that this person will select a value 0.49. So, what is the typical representative of numbers smaller than a given one?

**Analysis of the problem.** In general, once we have a number  $k$ , what is a typical representative of all the non-negative numbers which are smaller than  $k$ ?

To answer this question, let us first note that while from the purely mathematical viewpoint, there are infinitely many numbers on the interval  $[0, k]$ , in practice, there is usually some small amount  $h$  such that values whose difference is smaller than  $h$  are indistinguishable. For example, for grades scaled from 0 to 100, it is usually 1 point or, sometimes, 0.1 points.

In this case, we have only finitely many possible smaller values:  $0, h, 2h, 3h, \dots$ , all the way to the largest value  $n \cdot h$ , where  $n \approx k/h$ . For example, if  $h = 1$ , then for grades smaller than 70, we have 70 different possible values  $0, 1, 2, 3, \dots$ , all the way to 69. For  $h = 0.1$ , we get possible values  $0, 0.1, 0.2, 0.3, \dots$ , all the way to 69.9.

When we say that some value  $t$  is a “typical” representation of all these values, what we mean that this typical value should be kind of close to all possible values, i.e., that we should have  $t \approx 0, t \approx h, t \approx 2h, t \approx 3h, \dots, t \approx n \cdot h$ . In other words, the tuple  $(t, t, t, t, \dots, t)$  formed by the left-hand sides of these approximate equalities should be close to the tuple  $(0, h, 2h, 3h, \dots, n \cdot h)$  formed by the right-hand sides.

Tuples of real numbers can be naturally represented as points in the corresponding multi-D space, and thus, the distance

$$d((t, t, t, t, \dots, t), (0, h, 2h, 3h, \dots, n \cdot h)) = \sqrt{(t - 0)^2 + (t - h)^2 + (t - 2h)^2 + (t - 3h)^2 + \dots + (t - n \cdot h)^2} \quad (1)$$

between the corresponding points is the natural measure of closeness between the tuples. The closer the tuples, the more typical is the value  $t$ . Thus, we need to select the value  $t$  for which the distance (1) is the smallest possible.

A non-negative expression (1) is the smallest if and only if its square

$$d^2((t, t, t, \dots, t), (0, h, 2h, 3h, \dots, n \cdot h)) = (t - 0)^2 + (t - h)^2 + (t - 2h)^2 + (t - 3h)^2 + \dots + (t - n \cdot h)^2 \tag{2}$$

is the smallest. Differentiating this expression with respect to the unknown  $t$  and equating the resulting derivative to 0, we conclude that

$$2 \cdot (t - 0) + 2 \cdot (t - h) + 2 \cdot (t - 2h) + 2 \cdot (t - 3h) + \dots + 2 \cdot (t - n \cdot h) = 0. \tag{3}$$

Dividing both sides of this equality by 2 and moving all free terms to the right-hand side, we get

$$(n + 1) \cdot t = 0 + h + 2h + 3h + \dots + n \cdot h = (0 + 1 + 2 + 3 + \dots + n) \cdot h. \tag{4}$$

It is known that

$$0 + 1 + 2 + 3 + \dots + n = \frac{n \cdot (n + 1)}{2},$$

hence the equality (4) takes the form

$$(n + 1) \cdot t = \frac{n \cdot (n + 1)}{2} \cdot h,$$

and thus,

$$t = \frac{n \cdot h}{2}.$$

Since  $n \cdot h \approx k$ —and the difference between these two value is of order  $h$ , i.e., negligible, we conclude that  $t \approx k/2$ .

In other words, among all the values which are smaller than  $k$ , the typical value is

$$t = \frac{k}{2}. \tag{5}$$

*Comment.* In the above argument, we implicitly assumed that all possible values  $0, h, 2h, 3h, \dots$ , are equally possible. This assumption makes sense—since we have no reason to assume that some of these values are more probable than others. Such an argument is known as *Laplace Indeterminacy Principle*. It is a particular case of a very successful more general argument of this type known as the Maximum Entropy Approach; see, e.g., [2].

**Resulting formalization leads to approximately 70/100 threshold for Satisfactory.** Let us apply the above description (5) to our problem. Our description of satisfactory is that the proportion  $d = 1 - k$  of the course material that a student does not know is much smaller than the proportion  $k$  of the course material that the

student knows. It is reasonable to select, as a threshold for this property, a “typical” smaller-than- $k$  value, i.e.,  $k/2$ .

The condition that  $1 - k = k/2$  leads to  $k = 2/3 = 0.66\dots$ , i.e., indeed to approximately 70%.

## 2 Why 40-40-20 Proportion for Research, Teaching, and Service: First Explanation

**Formulation of the problem.** In many universities, it is recommended that faculty spend 40% of their time on research, 40% on teaching, and 20% on service. Again, the fact that this arrangement is widely accepted means that it corresponds to the intuitive ideas and is empirically reasonable. How can we explain this empirical fact?

**Intuitive idea.** Intuitively, the idea is that we should spend equal time on research and teaching, and less time on service.

**Let us formalize this idea.** The proportion  $r$  of time spent on research should be equal to the proportion  $t$  of time spent on teaching, and should be larger than the proportion of time  $s$  spent on service. Equality is straightforward:  $r = t$ . In line with the above general description, it is reasonable to formalize the fact that the proportion  $s$  is smaller than the proportion  $r = t$  as  $s = r/2$ .

**This formalization leads exactly to the 40-40-20 arrangement.** Let us show that the above formalization explains the above arrangement. Indeed, from  $r + t + s = 1$ ,  $t = r$ , and  $s = r/2$ , we conclude that  $2r + r/2 = 2.5r = 1$ , hence  $r = 0.4$ . Thus,  $t = r = 0.4$  and  $s = r/2 = 0.2$ , which is exactly the current arrangement.

## 3 Why 40-40-20 Proportion for Research, Teaching, and Service: Second Explanation

**Seven plus minus two law.** Our second explanation is based on the well-known “seven plus minus two” law (see, e.g., [3, 4]), according to which we naturally divide everything into  $7 \pm 2$  clusters—into how many depends on the person. Because of this, a person who divides everything into 9 clusters will not pay serious attention to 1/9-th of the time, a person who divides everything into 5 clusters will not pay serious attention to any activity that takes less than 1/5-th of the overall time, etc.

**Resulting explanation.** The main objectives of a university are teaching and research, service is clearly not that important—but we still want people to do service, otherwise the university will not function smoothly—serve on committees, develop curricula, etc. We do not want faculty to spend too much time on service, but we want them to take it seriously.

Thus, it is reasonable to select for the service, the smallest possible proportion that would still be taken seriously by everyone, no matter whether they divide everything into 5 or into 9 clusters. Thus, we need the smallest number which is larger than all the corresponding thresholds  $1/9$ ,  $1/8$ ,  $1/7$ ,  $1/6$ , and  $1/5$ . One can easily see that this smallest non-negligible number is exactly  $1/5 = 20\%$ , which is exactly how much time is allocated to service.

If we consider research and reaching to be equally important, then the remaining time  $1 - 0.2 = 0.8$  should be equally divided between these two activities, into two equal parts of 40 and 40%. So, we indeed get an explanation for the 40-40-20 arrangement.

## 4 Why 50-30-20 Proportion for Research Universities: Two Explanations

**What we want to be explained.** In many research universities, the usual proportion is different: 50% for research, 30% for teaching, and 20% for service. How can we explain this arrangement?

**First explanation.** The main idea behind this arrangement is that a faculty should spend less time on teaching than on research, and less time on service than on teaching. In our notation, this means that we should have  $s < t$  and  $t < r$ .

According to our formalization, this implies that  $t = r/2$  and  $s = t/2$  (hence  $s = r/4$ ). Thus, the condition that  $r + t + s = 1$  implies that

$$r + r/2 + r/4 = (7/4) \cdot r = 1,$$

hence  $r = 4/7 \approx 0.57$ ,  $t = r/2 = 2/7 \approx 0.29$ , and  $s = t/2 = 1/7 \approx 0.14$ . The resulting 57-29-14 arrangement is indeed close to 50-30-20.

**Second explanation.** Let us see what seven plus minus two law implies in this situation. For service, we still want to the smallest non-negligible proportion, i.e., 20%. The difference from the previous case is that instead of allocating equal time to research and teaching, we allocate more time to research.

Teaching is important, so a reasonable idea is to allocate to teaching the largest possible time for which the difference between teaching and research time should be significant to everybody. As we have mention, the smallest non-negligible difference is 20%. So, we have  $t + r = 1 - 0.2 = 0.8$  and  $r - t = 0.2$ . This implies exactly  $r = 0.5$ ,  $t = 0.3$ , and  $s = 0.2$ —exactly the 50-30-20 arrangement.

## 5 Why Five Letter Grades

**What we want to explain.** In the US system, number of points is transformed into one of five “letter grades”—A (excellent), B (good), C (satisfactory), D (sometimes passable), and F (fail). Letter grades are usually the only thing that does into the student’s transcript.

In Russia—where two of us are from—we have a different system, but also 5 main grades. Why five?

*Comment.* At our university, periodically, faculty raise the need to have a more specific scale, with the possibility to have  $A-$ ,  $B+$ , and other combination of grades. However, every time, a significant proportion of faculty objects, and the motion does not pass.

In Russia, we had such an plus-minus option, we could even have two pluses like  $5++$  for a really outstanding performance, and  $3--$  for an almost failing one. However, these pluses and minuses did not go into an official transcript and were not taken into account when computing the average grade.

**Natural explanation.** We want the difference between letter grades to be clearly understood by everyone, irrespective of whether they divide everything into 5, 7, or 9 clusters. This means that we must have no more than 5 grades—otherwise, if we had 6 or more letter grades, the difference between some of these grades would not be clear to those who divided everything into 5 clusters. This explains why we normally use 5 letter grades.

*Comment.* A similar fact is true for musical scales. Traditionally, many cultures had different scales, some have 5 notes (*pentatonic scales*), the traditional Western scale has 7 notes—which corresponds to the most frequent number of 7 clusters, and practically all the scales have between 5 and 9 notes—in full agreement with the seven plus minus two law.

## 6 Why Excellent Is Usually Close to 90

**Idea.** Excellent means that there may be some minor faults in the student’s knowledge of the course material, but overall, no one should be able to notice any major fault, irrespective of whether this person divides everything into 5 or 9 clusters.

**Resulting explanation.** To be un-noticeable to a person who divides everything into  $c$  clusters, the proportion  $d$  of the course material that the student does not know should be smaller than  $1/c$ —the smallest amount seriously recognizable by this person. Thus, excellent knowledge means that the part  $d$  that the student does not know should be smaller than all possible values  $1/5$ ,  $1/6$ ,  $1/7$ ,  $1/8$ , and  $1/9$ . This is equivalent to requiring that  $d < 1/9$  and that  $k = 1 - d > 8/9 \approx 0.89$ . This is indeed very close to the usual 90/100 threshold for “excellent” (A).

## 7 How to Allocate Grades to Tests, Homeworks, etc.

**Idea.** The overall grade comes from adding grades for different tests, assignments, etc. Let us use the above ideas to decide how many points out of 100 to allocate to each test, to the final exam, to different assignments, etc. We will illustrate this idea on two examples.

**First example: a regular undergraduate class.** We have three tests (also known as midterm exams), homeworks, and a final exam. Intuitively, we should assign similar number of points  $t_1 = t_2 = t_3$  to each of the three tests, and approximately the same number of points to the homeworks  $h \approx t_i$ , but definitely the final exam is more important, so the number of points  $f$  allocated to the final exam should be larger:

$$t_i < f.$$

Similarly to what we did earlier, we interpret  $t_i < f$  as  $t_i = f/2$ , i.e., as  $f = 2t_i$ . Thus, the fact that the sum of all the points is 100 means that

$$t_1 + t_2 + t_3 + h + f = 4t_i + 2t_i = 6t_i = 100.$$

This implies that  $t_1 = t_2 = t_3 = h = 100/6 \approx 17$  and  $f = 2 \cdot (100/6) \approx 33$ .

It is usually more convenient to use round numbers of points, i.e., numbers divisible by 5. For 17, the closest such value is 15, and for 33, it is 35. However, if we take  $t_1 = t_2 = t_3 = h = 15$  and  $f = 35$ , the overall maximum grade is  $4 \cdot 15 + 35 = 95 < 100$ . To make it 100, we need to increase one of the allocations by 5. Which one we increase? We want to keep all tests equally important, so we cannot increase one of these allocations, we should increase either  $h$  or  $f$ . Which one?

- If we increase  $h$  from 15 to 20, the difference between the new value 20 and the original value  $\approx 17$  is  $\approx 3$ .
- If we increase  $f$  from 35 to 40, the difference between the new value 40 and the original value  $\approx 33$  is  $\approx 7$ .

So, the smallest deviation from the original arrangement is when we increase  $h$ . Thus, we arrive at the following arrangement—that many of our faculty actually use in such situations:

- each of the three tests is worth 15 points,
- all the homeworks are worth 20 points, and
- the final exam is worth 35 points.

**Second example: a regular graduate class.** We have three tests, homeworks, a project, and a final exam. This time, all three tests and homeworks are equally important just as in the previous example, a project is more important than any of them, and the final exam is the most important. So, we still have  $t_1 = t_2 = t_3 = h$ . Since

the project is more important, we allocate the number of points to it which is larger than  $t_i$ . According to our arrangement, this means  $t_i = p/2$ , i.e.,  $p = 2t_i$ . Similarly, the condition that  $p < f$  leads to  $p = f/2$ , i.e., to  $f = 2p$  and thus, to  $f = 4t_i$ . The condition that these allocations add up to 100 leads to

$$4t_i + p + f = 4t_i + 2t_i + 4t_i = 10t_i = 100,$$

i.e., to  $t_i = 10$ . So,  $p = 2t_i = 20$  and  $f = 4t_i = 40$ . Thus, in this case:

- each of the three tests is worth 10 points,
- all the homeworks are worth 10 points,
- the project is worth 20 points, and
- the final exam is worth 40 points.

This is close to the arrangement that we came up with empirically.

**What if we have a different number of tests.** In the undergraduate case, if we have  $T$  tests, then;

- each of the tests is worth  $100/(T + 3)$  points,
- all the homeworks are worth  $100/(T + 3)$  points, and
- the final exam is worth  $200/(T + 3)$  points.

In the graduate case, if we have  $T$  tests, then:

- each of the three tests is worth  $100/(T + 7)$  points,
- all the homeworks are worth  $100/(T + 7)$  points,
- the project is worth  $200/(T + 7)$  points, and
- the final exam is worth  $400/(T + 7)$  points.

**Acknowledgements** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes), and by the AT&T Fellowship in Information Technology.

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478, and by a grant from the Hungarian National Research, Development and Innovation Office (NRDI).

## References

1. Hewitt, M.: Musical Scales of the World. The Note Tree, Gloucester, UK (2013)
2. Jaynes, E.T., Bretthorst, G.L.: Probability Theory: the Logic of Science. Cambridge University Press, Cambridge, UK (2003)
3. Miller, G.A.: The magical number seven plus or minus two: some limits on our capacity for processing information. Psychol. Rev. **63**, 81–97 (1956)
4. Milner, P.M.: Physiological Psychology. Holt, Rinehart and Winston, New York (1970)

# Shall We Ignore All Intermediate Grades?



Christian Servin, Olga Kosheleva, and Vladik Kreinovich

**Abstract** In most European universities, the overall student's grade for a course is determined exclusively by this student's performance on the final exam. All intermediate grades—on homework, quizzes, and previous texts—are, in effect, ignored. This arrangement helps gauge the student's performance by the knowledge that the student shows at the end of the course. The main drawback of this approach is that some students do not start studying until later, thinking that they can catch up and even get an excellent grade—and this hurts their performance. To motivate students to study hard throughout the semester, most US universities estimate the overall grade for the course as a weighted average of the grade on the final exam and of all intermediate grades. In this paper, we show that even when a student is already motivated, to accurately gauge the student's level of knowledge it is important to take intermediate grades into account.

## 1 Formulation of the Problem

**Two systems of grading.** In most countries, in most universities, in most courses, students have intermediate tests and quizzes, homeworks, labs, most of which are graded. At the end of the course, there is usually a final exam which is also graded.

---

C. Servin

Computer Science and Information Technology Systems Department, El Paso Community College (EPCC), 919 Hunter Dr., El Paso, TX 79915-1908, USA  
e-mail: [cservin1@epcc.edu](mailto:cservin1@epcc.edu)

O. Kosheleva

Department of Teacher Education, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA  
e-mail: [olgak@utep.edu](mailto:olgak@utep.edu)

V. Kreinovich (✉)

Department of Computer Science, University of Texas, 500 W. University, El Paso, TX 79968, USA  
e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)



In the US, usually, the overall grade for the course is estimated by combining the grade for the final exam and all intermediate grades—most frequently, by taking the weighted average. In contrast, in most European countries, the overall grade for the course is the grade on the final exam, with intermediate grades serving only on a pass-fail basis—to be able to take the final exam, a student needs to have a satisfactory average on all intermediate exams.

**Why the difference: pro and contra.** The ultimate goal of the grade is that it should reflect the knowledge that the student acquired after taking the course. From this viewpoint, it seems to make sense to use the European system: if the student did not do perfectly well on the intermediate exams, but eventually learned the material perfectly, this student should get a perfect grade.

The downside of this approach is that some students procrastinate and only start studying much later, thinking that they still have a chance to learn the material and get the perfect grade. Sometimes they succeed, but often they don't: they get a low grade or even fail the class. This problem did not bother people in the past, when a relatively small number of people could get higher education. Students were accepted only after very competitive final exams, so if a student does not want to study hard, well, good riddance, there are plenty of practically as good students eager to take this student's place.

However, nowadays, when jobs not requiring education are more and more automated, societies need highly educated people to survive in the global competition. Universities accept a large number of people, and not all of them are prepared to work hard. So we need to motivate them to study—and the US system clearly motivates students to start studying from the very beginning, since otherwise their not so good intermediate grades will affect their final grade for the class.

From this motivational viewpoint, some version of a US system is preferred.

**What we do in this paper.** In this paper, we show—somewhat unexpectedly—that even if we have perfect motivations, and we are willing to gauge a student by the knowledge he/she attained after the course, we still need to take intermediate grades into account.

## 2 Analysis of the Problem and the Resulting Conclusion

**Main idea.** A typical US final exam lasts for 2 h and 45 min. The final exam is supposed to be comprehensive, covering all main topics that were studied in the course. It is possible to cover many things in this time, but clearly not everything that was taught during the semester.

If a student correctly solved 7 problems out of 10, but did not do well on intermediate assignments, then maybe this student's degree of knowledge is less than 70%, he/she just got lucky by the fact that questions on the exam—which were reasonably randomly selected from all possible questions—were mostly from the parts of the material that this student knew. On the other hand, if, in addition to correctly solving

7 problems out of 10 on the final exam, the student also had a similar satisfactory grade for all intermediate assignments, we are much more confident that this student indeed knows at least 70% of the material.

Thus, to accurately gauge the student's degree of knowledge, it is necessary to also take into account this student's intermediate grades.

**How to take intermediate grades into account?** The above qualitative argument shows that it is desirable to take intermediate grades into account. A natural next question is how exactly to take the intermediate grades into account when computing the overall grade for the course.

**How the grade on the final exam is usually computed.** To answer this question, let us first recall how the grade on the final exam is usually computed.

Each question on the final exam usually consists of several parts (explicit or implicit "sub-questions"), and the grade for this question is determined by how many of these parts the student answered correctly. For example, if an assignment is to apply a multi-stage algorithm, the instructor will check whether each of the steps is correctly performed.

The grade for the final exam is then obtained by adding the grades for all the questions. From this viewpoint, the grade for the final exam is determined by the number of correctly answered sub-questions.

In general, if on the final grade, out of  $s$  sub-questions, the student correctly answered  $n$  of them, then the student gets a fraction

$$\tilde{p} = \frac{n}{s} \quad (1)$$

of the maximum possible score.

**What do we want to estimate and how can we estimate it based on the final exam.**

A natural measure of the student's knowledge is the proportion  $p$  of all possible sub-questions to which the student knows the correct answer. This means that for each randomly selected sub-question, the probability that the student knows the correct answer to this sub-question is equal to the proportion  $p$ .

The actual number  $n$  of sub-questions that the student answered correctly on the final exam can be obtained by adding  $s$  independent 0-1-valued random variables  $v_i$  describing whether the  $i$ -th sub-question was answered correctly. For each of these variables,  $v_i$  the mean value is equal to

$$E[v_i] = 1 \cdot p + 0 \cdot (1 - p) = p, \quad (2)$$

and the variance is equal to

$$E[(v_i - E[v_i])^2] = p \cdot (1 - p)^2 + (1 - p) \cdot (0 - p)^2 = p \cdot (1 - p) \cdot (1 - p + p) =$$

$$p \cdot (1 - p); \quad (3)$$

see, e.g., [1] for this and following formulas.

For the sum  $n$  of several independent random variables, the mean is equal to the sum of the means, and the variance is equal to the sum of the variances, so  $E[n] = p \cdot s$  and  $V[n] = p \cdot (1 - p) \cdot s$ .

During the 2 hours and 45 minutes we can ask a lot of sub-questions, so the number  $s$  is reasonably large. It is known that the probability distribution of the sum of a large number of small independent random variables is close to Gaussian—this is a consequence of the Central Limit Theorem (and the main reason why normal distributions are ubiquitous). Thus, we can conclude that the number  $n$  of correctly answered sub-questions is normally distributed, with mean  $E[n] = p \cdot s$  and standard deviation  $\sigma[n] = \sqrt{p \cdot (1 - p) \cdot s}$ .

The difference  $n - E[n] = n - p \cdot s$  is normally distributed, with 0 mean and standard deviation  $\sigma = \sqrt{p \cdot (1 - p) \cdot s}$ . For large  $n$ , the difference  $n - p \cdot s$  is small, so  $n \approx p \cdot s$  and thus,

$$p \approx \tilde{p} \stackrel{\text{def}}{=} \frac{n}{s}, \quad (4)$$

hence  $\sigma[n] \approx \tilde{\sigma} \stackrel{\text{def}}{=} \sqrt{\tilde{p} \cdot (1 - \tilde{p}) \cdot s}$ . So, once we know the number  $n$  of sub-questions that the student has correctly answered on the final exam, we can conclude that the value  $p \cdot s$  is normally distributed with mean  $n$  and standard deviation  $\tilde{\sigma}$ .

Thus, the actual (unknown) grade  $p$  is also normally distributed, with mean

$$\tilde{p} = \frac{n}{s} \quad (5)$$

and standard deviation

$$\sigma \approx \sqrt{\frac{\tilde{p} \cdot (1 - \tilde{p})}{s}}. \quad (6)$$

**How to take into account intermediate grades: idea.** In addition to the  $s$  sub-questions that form the final exam, the student also answered several sub-questions before that, as part of intermediate tests, quizzed, homeworks, etc. Let us denote the overall number of such sub-questions by  $S$ , and the overall number of those of these sub-questions that the student answered correctly by  $N$ .

This does not necessarily mean that this is how much the student knows now, at the time of the final exam: the student may have learned what he or she missed earlier. What we can conclude, however, is that the student's degree of knowledge is at least as large as what can be concluded from this student's intermediate grades.

In the worst case scenario, when the student did not learn anything since previous exams, this student's degree of knowledge is normally distributed with the mean

$$\tilde{P} = \frac{N}{S} \quad (7)$$

and standard deviation

$$\Sigma = \sqrt{\frac{\tilde{P} \cdot (1 - \tilde{P})}{S}}. \quad (8)$$

Thus, with high certainty, we can conclude that this actual degree of the student's knowledge is located on the  $k$ -sigma interval  $[\tilde{P} - k \cdot \Sigma, \tilde{P} + k \cdot \Sigma]$ , where  $k$  depends on the desired degree of certainty: for  $k = 3$ , we get the degree of certainty 99.9%, for  $k = 6$ , we get the degree of certainty  $1 - 10^{-8}$ , etc.

This degree of knowledge could only increase, so we can conclude that the degree of knowledge cannot be smaller than the value

$$\underline{P} = \tilde{P} - k \cdot \Sigma = \tilde{P} - k \cdot \sqrt{\frac{\tilde{P} \cdot (1 - \tilde{P})}{S}}. \quad (9)$$

This leads is to the following recommendation.

**How to take into account intermediate grades: recommendation.** Let us assume that out of  $s$  sub-questions on the final exam, the student answered  $n$  sub-questions correctly. Let us also assumed that out of  $S$  sub-questions asked before the final exam, the student answered  $N$  sub-questions correctly. Then, the actual student's degree of knowledge  $p$  can be described by a normal distribution with mean

$$\tilde{p} = \frac{n}{s} \quad (10)$$

and standard deviation

$$\sigma \approx \sqrt{\frac{\tilde{p} \cdot (1 - \tilde{p})}{s}} \quad (11)$$

restricted to values

$$p \geq \tilde{P} - k \cdot \sqrt{\frac{\tilde{P} \cdot (1 - \tilde{P})}{S}}, \quad (12)$$

where we denoted

$$\tilde{P} = \frac{N}{S} \quad (13)$$

Thus, a natural measure of the student's knowledge is the mean value of this restricted normal distribution. The larger the intermediate grade  $\tilde{P}$ , the larger the restricting lower bound on  $p$  and thus, the larger the resulting mean. So, we indeed take into account the intermediate grades when estimating the overall grade for the class.

**Acknowledgements** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

## Reference

1. Sheskin, D.J.: Handbook of Parametric and Non-Parametric Statistical Procedures. Chapman & Hall/CRC, London, UK (2011)

# Why $\infty$ is a Reasonable Symbol for Infinity



Olga Kosheleva and Vladik Kreinovich

**Abstract** The fact that  $\infty$  is actively used as a symbol for infinity shows that this symbol is probably reasonable in this role, but why? In this paper, we provide a possible explanation for why this is indeed a reasonable symbol for infinity.

## 1 Formulation of the Problem

**Fact.** In mathematics, we use the symbol  $\infty$  for infinity.

**History.** This symbol was first used to describe infinity in 1655, by John Wallis in his book [6], on p. 4 of the section “De Sectionibus Conicis, Nova Methodo Expositis”; see also [1, 5].

Interestingly, this symbol was, at first, not universally accepted. For example, Leonard Euler, one of the most famous 18th century mathematicians (and probably the most productive mathematician of all ages), used a different symbol—similar to the current symbol  $\sim$  for similarity; see, e.g., Euler’s paper [2] published in the Proceedings of Russian Academy of Sciences. But:

- in spite of Euler’s authority and fame as a mathematician, it was not his symbol that was eventually accepted as the symbol for infinity,
- it was the symbol proposed by a much less known and much less authoritative Wallis.

---

O. Kosheleva

Department of Teacher Education, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA

e-mail: [olgak@utep.edu](mailto:olgak@utep.edu)

V. Kreinovich (✉)

Department of Computer Science, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA

e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

39

M. Ceberio and V. Kreinovich (eds.), *Decision Making Under Uncertainty*

*and Constraints*, Studies in Systems, Decision and Control 217,

[https://doi.org/10.1007/978-3-031-16415-6\\_6](https://doi.org/10.1007/978-3-031-16415-6_6)

**A natural question.** Why was the current infinity symbol adopted and other symbols not? Why is this a reasonable symbol for infinity—because if it was not, another more reasonable symbol would replace it.

**What we do in this paper.** In this paper, we provide a possible explanation.

## 2 Our Explanation

**What is a natural way to represent infinity: a question.** How can we represent infinity—i.e., a process without end?

**First natural idea.** The first natural idea that comes to mind when we think about infinity is a straight (or curved) line.

So why not use a straight line to represent infinity?

**Problem with this idea.** In each sheet of paper, we only have a limited space to put symbols in. As a result, it is not possible to place the whole straight line.

And if we cut it off and only draw a segment of the straight line, this segment does not have any association with infinity.

**We need a closed curve.** Since we cannot represent an infinite motion in which the body moves farther and farther from the original point, the next natural idea is to represent a never-ending motion that is confined to a limited space.

Of course, in a limited space, we can only represent a limited part of the infinite trajectory. To make sure that this part indeed represents the never-ending motion, we need to make sure that the trajectory does not end abruptly, that it is clear how it continues. The only way to do that is to make sure that the trajectory goes back to one of its previous points, i.e., in mathematical terms, that the trajectory is a *closed cycle*, a *closed curve*.

**Which closed curve should we select?** There are many different closed curves, with different number of self-intersections.

Which one should we select?

**Second natural idea: let us select the simplest curve.** A natural idea is to select the simplest of the closed curves, i.e., a closed curve without self-intersections.

**Problem with this idea.** The problem with this idea is that this is exactly a symbol 0 for zero—a closed curve with no self-intersections.

**Final idea: let us select the next simplest curve.** Since we cannot select the simplest closed curve, with no self-intersections, a natural idea is to select the next simplest curve, with exactly one self-intersection.

**This is exactly the usual infinity symbol.** This is exactly the usual infinity symbol!

Thus, this symbol is indeed explained.

### 3 Real-Life Examples of an $\infty$ -Like Trajectory

Trajectories that form a closed curve with exactly one self-intersection are common. Let us give a few examples.

**Astronomy.** If we show how the position of the Sun in the sky—as seen from a fixed location on Earth at the same time of day—varies over the course of a year, we will end up with an  $\infty$ -shaped trajectory.

This fact was already known to the ancient Greeks, this is why this trajectory is known by its Greek name—*analemma*. Claudius Ptolemy, the most famous astronomer of the ancient times—whose system was used all the way until Copernicus—even had a book titled *Analemma*; see, e.g., [4].

**Chaos.** Now everyone has heard about chaos and chaotic systems, i.e., systems for which long-term prediction is not possible—since a tiny uncertainty in the original position will eventually lead to huge uncertainty in the future state.

Historically the first such system—a simplified version of a weather system—was discovered by Edward Lorenz in the 1960s and is, because of this, known as the *Lorenz system*. Its trajectories resemble the  $\infty$  symbol; see, e.g., [3].

**Space flights.** This was the shape of the trajectories of all the missions during the 1960s Apollo missions to the Moon.

The fact that the selected trajectory was a closed curve made perfect sense: it made sure that even if the major engine fails near the Moon, the spaceship would return, by itself, to the near-Earth part of the orbit from which it started—and thus, be able to make a safe landing.

Out of all closed-curve trajectories, other considerations led to the selection of the trajectory with a single self-intersection; see, e.g., [7].

**Acknowledgements** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

## References

1. Cajori, F.: A History of Mathematical Notations. Cosimo Classics, New York (2019)
2. Euler, L.: *Variae observationes circa series infinitas*. *Commentarii Academiae Scientiarum Petropolitanae* **9**(1744), 160–188 (in Latin). <http://eulerarchive.maa.org/docs/originals/E072.pdf>
3. Gleick, J.: *Chaos: making a New Science*. Penguin Books, New York (2008)
4. Neugebauer, O.: *A History of Ancient Mathematical Astronomy*. Springer, Berlin and New York (1975)
5. Scott, J.F.: *The Mathematical Work of John Wallis, D.D., F.R.S. (1616–1703)*. American Mathematical Society, Providence, Rhode Island (1981)
6. Wallis, J.: *Pars Prima*, London (1655) (in Latin)
7. Woods, W.D.: *How Apollo Flew to the Moon*. Springer, New York (2011)



# What Is $1/0$ from the Practical Viewpoint: A Pedagogical Note



Olga Kosheleva and Vladik Kreinovich

**Abstract** What is  $1/0$ : Students are first taught—in elementary school—that it is undefined, then—in calculus—then it is infinity. In both cases, the answer is usually provided based on abstract reasoning. But what about the practical meaning? In this paper, we show that, depending on the specific practical problem, we can have different answers to this question: in some practical problems, the correct answer is that  $1/0$  is undefined, in others, the correct answer is that  $1/0 = 0$ —and there are probably other practical problems where we can have different answers. Bottom line: there is no universal answer, the correct answer depends on what practical problem we are considering.

## 1 Formulation of the Problem

**What is  $1/0$ : what students learn.** In elementary school, students learn that you cannot divide by 0. This makes sense: by definition, the ratio  $a/b$  is a number that, multiplied by  $b$ , gives  $a$ . Of course, no matter what number you multiply by  $b = 0$ , you always get 0, so you will never get  $a = 1$ .

Later, the student learn that in calculus,  $1/0$  is infinity—since 0 is the limit of, e.g., a sequence  $1/n$ , we can thus interpret  $1/0$  as the limit of values  $1/(1/n) = n$ , i.e., infinity.

**Problem.** From the purely mathematical viewpoint, both answers make sense—as well as many other facts and results from mathematics. However, there is a difference between this mathematical fact and other mathematical facts: many other facts

---

O. Kosheleva

Department of Teacher Education, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA

e-mail: [olgak@utep.edu](mailto:olgak@utep.edu)

V. Kreinovich (✉)

Department of Computer Science, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA

e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

43

M. Ceberio and V. Kreinovich (eds.), *Decision Making Under Uncertainty*

and *Constraints*, Studies in Systems, Decision and Control 217,

[https://doi.org/10.1007/978-3-031-16415-6\\_7](https://doi.org/10.1007/978-3-031-16415-6_7)

makes perfect sense in practical applications. For example,  $8/4 = 2$  means that if we equally divide 8 apples between 4 students, each student gets 2 apples. Such practical examples help students better understand the corresponding mathematical facts and results. In contrast,  $1/0$  seems to be a purely mathematical exercise. Usually, no practical examples are provided to explain the meaning of this ratio. This makes studying this idea too abstract and thus, more complicated to many students.

**What we do in this paper.** In this paper, we provide one of the possible practical meanings of this ratio, and explain, for this practical example, what will the ratio  $1/0$  mean in this particular case.

Of course, this is just one possible practical example. We are sure that there can be many other practical examples, and in many of them, the meaning of  $1/0$  will be different.

## 2 Practical Problem

**General description of the situation.** Let us start with the general description of a practical problem that corresponds to computing  $1/c$  for any real number  $c$ .

This problem is related to signal propagation. It is well known that as a signal travels—be it by wire or through air—its amplitude decreases. As a result, when the sender sends a signal of amplitude  $a$ , the receiving agent receives a signal of smaller amplitude  $r = c \cdot a$ , for some value  $c < 1$ .

To reconstruct the original signal, the receiving agent thus needs to *amplify* the received signal, i.e., in precise terms, to multiply it by some constant  $C > 1$ .

**What is the problem.**

- We know the coefficient  $c < 1$  that describes how much the original signal decreased.
- We want to find the amplification coefficient  $C$  that allows us to reconstruct the original signal.

## 3 Idealized Setting

**Description of the ideal case.** Let us first consider the ideal situation when there is no noise, and the only change in the original signal is that its amplitude decreases, from  $a$  to  $c \cdot a$ .

**What is the proper amplification coefficient.** In this case:

- we receive the signal  $r = c \cdot a$ , and
- we multiply it by  $C$ , getting  $C \cdot r = C \cdot c \cdot a$ .

We want to make sure that for all signals  $a$  sent by the sender, the resulting signal  $C \cdot c \cdot a$  is equal to exactly  $a$ , i.e., that  $C \cdot c \cdot a = a$ .

In particular, for  $a = 1$ , we get  $C \cdot c = 1$ , so  $C = 1/c$ . One can easily check that for this amplification coefficient  $C = 1/c$ , and for every sender's signal  $a$ , we indeed have  $C \cdot c \cdot a = (C \cdot c) \cdot a = 1 \cdot a = 1$ .

Thus, this practical problem provides a practical interpretation for  $1/c$ .

**In this case, what is 1/0?** In this interpretation, the ratio  $1/0$  is simply not defined: if  $c = 0$ , then, no matter what amplification coefficient  $C$  we select, we will never get  $C \cdot c = 1$ .

**This is exactly what kids learn in school.** True, this is exactly what kids learn, that  $1/0$  is not defined. So far, nothing new, nothing interesting.

But remember that we consider an idealized case, when we assume that there is no noise. In practice, there is always some noise. What happens to this practical problem in this more realistic setting?

## 4 Realistic Setting

**Realistic setting: general idea.** Let us now take into account that, in addition to being multiplied by a coefficient  $c < 1$ , the signal also gets corrupted by noise  $n$ . In other words, the received signal  $r$  is equal to

$$r = c \cdot a + n, \tag{1}$$

where  $n$  denotes the noise, i.e., the additional change in the signal.

In this case, after amplification, you do not get the original signal, you get a signal

$$s = C \cdot r = C \cdot c \cdot a + C \cdot n, \tag{2}$$

which is different from  $a$  even when  $C \cdot c = 1$ .

The goal is to find the amplification coefficient  $C$  for which the amplified signal  $s$  is the closest to the original signal  $a$ .

**Realistic setting: details.** We do not know the value of the noise  $n$ —if we knew it, we could simply subtract this known value from the received signal  $r$  and thus, eliminate the effect of the noise. From the mathematical viewpoint, this means that  $n$  is a random variable.

Natural characteristics of a random variable  $n$  are its mean value  $E[n]$  and its variance  $V \stackrel{\text{def}}{=} E[(n - E[n])^2]$ —or, equivalently, its standard deviation  $\sigma \stackrel{\text{def}}{=} \sqrt{V}$  for which  $V = \sigma^2$ ; see, e.g., [1]. While we do not know the exact value of the noise, based on the previous experiences, we can estimate both the mean and the standard deviation.

The additive random noise can be both positive and negative. A priori, there is no reason to believe that positive values are more probable or negative values are more probable, so it make sense to assume that both are equally probable, and that the mean value of the noise is 0. Let us denote the standard deviation of noise by  $\sigma_n$ .

Similarly, we do not know what will be the signal that the sender will be sending— if we knew, there would be no need to send anything. Thus, the signal can also be viewed as a random variable. We also do not have any reason to believe that positive values of the signal will be more or less probable than negative values. So, it also makes sense to assume that the mean value of the signal is 0. Let us denote the standard deviation of the signal by  $\sigma_a$ .

**How do we gauge which coefficient  $C$  is better.** We are interested in minimizing the reconstruction error, i.e., the difference  $d \stackrel{\text{def}}{=} s - a$  between the reconstructed signal  $s$  and the original signal  $a$ . Due to (2), we get the following expression for this error:

$$d = (C \cdot c - 1) \cdot a + C \cdot n. \quad (3)$$

Since the mean values of  $a$  and  $n$  are both 0s  $E[a] = E[n] = 0$ , the mean value of their linear combination  $d$  is also 0:  $E[d] = 0$ . It is therefore reasonable to gauge the value  $d$  by its variance  $V[d] = E[d^2]$ . Due to (3), we have

$$E[d^2] = (C \cdot c - 1)^2 \cdot E[a^2] + 2(C \cdot c - 1) \cdot C \cdot E[a \cdot n] + C^2 \cdot E[n^2]. \quad (4)$$

Signal  $a$  and noise  $n$  are clearly independent, so  $E[a \cdot n] = E[a] \cdot E[n] = 0 \cdot 0 = 0$ . Thus, the formula (4) takes the form

$$E[d^2] = (C \cdot c - 1)^2 \cdot \sigma_a^2 + C^2 \cdot \sigma_n^2. \quad (5)$$

We want to find the amplification  $C$  that minimizes this expression.

**The resulting optimal value of the amplification coefficient.** Differentiating the expression (5) with respect to  $C$  and equating the derivative to 0, we conclude that

$$2(C \cdot c - 1) \cdot c \cdot \sigma_a^2 + 2C \cdot \sigma_n^2. \quad (6)$$

If we divide both sides by 2 and move all the terms not containing  $C$  to the other side, we get

$$C \cdot (c^2 \cdot \sigma_a^2 + \sigma_n^2) = c \cdot \sigma_a^2, \quad (7)$$

hence the optimal amplification coefficient  $C$  is equal to

$$C = \frac{c \cdot \sigma_a^2}{c^2 \cdot \sigma_a^2 + \sigma_n^2}. \quad (8)$$

Of course, this value depends on the noise level. When the noise is small  $\sigma_n \approx 0$ , the value  $C$  is close to the limit value of this expression when  $\sigma_n \rightarrow 0$ :

$$C \approx C_{\lim} = \lim_{\sigma_n \rightarrow 0} \frac{c \cdot \sigma_a^2}{c^2 \cdot \sigma_a^2 + \sigma_n^2}. \tag{9}$$

What is this limit?

**What if  $c \neq 0$ .** In this case, both numerator and denominator have definite limits, so

$$C_{\lim} = \frac{c \cdot \sigma_a^2}{c^2 \cdot \sigma_a^2} = \frac{1}{c}. \tag{10}$$

Thus, this practical problem indeed provides a natural practical interpretation for the value  $1/c$ —at least when  $c \neq 0$ .

**But what if we take  $c = 0$ ?** In this case, the problem also makes sense, but the limit is different: here for all  $\sigma_n$ , we have

$$C = \frac{0}{\sigma_n^2} = 0,$$

and thus,

$$C_{\lim} = \lim_{\sigma_n \rightarrow 0} 0 = 0. \tag{11}$$

So, in this case, a practical problem leads to an unexpected conclusion that  $1/0 = 0$ .

**Conclusion.** On the example of two practical problems, we got two different answers to the question of what is  $1/0$ : that it is undefined, and that it is equal to 0. We are sure that there may be other practical problems in which the answer is  $1/c$  for  $c \neq 0$  and for  $c = 0$ , we get a different value.

Bottom line: what is  $1/0$  depends on the specific practical problem, we cannot always rely on abstract arguments.

**Acknowledgements** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes). It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

## Reference

1. Sheskin, D.J.: Handbook of Parametric and NonParametric Statistical Procedures. Chapman & Hall/CRC, London, UK (2011)

# Historical Diversity Through Base-10 Representation of Mayan Math



Julian Viera and Olga Kosheleva

**Abstract** This paper attempts to engage teacher education professionals (the field of teacher education) in a discussion about a novel approach to teaching mathematical operations using a base-10 representation of the Mayan number system. The base-10 representation was developed by Luis Fernando Magaña (Magaña, L. F. (1990). *Las matemáticas y los mayas*. Ciencias (019).) as an auxiliary tool to teach elementary children in Yucatan, Mexico. The Mayan vigesimal (base-20) system represents a divergent historical perspective from hegemonic European treatises of the origins of mathematics. We use a base-10 representation for mathematical operations in an elementary methods class so that pre-service teachers begin to develop culturally responsive pedagogy to critique discourse of power.

**Keywords** Mathematics education · Culturally responsive education · Mayan math

## 1 Introduction




It is common for K12 students to have learned about or read about European mathematicians or their contributions to mathematics. Discoveries of many laws and theorems attributed to Europeans, such as Newton's laws of motion, Pythagorean Theorem, and L'Hopital's Rule, are well known by students. Even whole subjects are credited to Europeans, such as Newton inventing calculus. Yet [2] found evidence that mathematical scholars disagreed and finally concluded that calculus was developed by both Newton and Leibniz. More recently, Bressoud [1] posited that Indian mathematicians were close to developing calculus. What is scarce is the history of early American mathematical achievements.

---












J. Viera  
Berea College, Berea, USA  
e-mail: [vierajrj@bereda.edu](mailto:vierajrj@bereda.edu)

O. Kosheleva (✉)  
University of Texas at El Paso, El Paso, USA  
e-mail: [olgak@utep.edu](mailto:olgak@utep.edu)

**Table 1** Mayan representations for one, five and zero

Mayan Symbol	Numerical representation
	one
	five
	zero

**Table 2** Mayan base-10 representation as developed by L. F. Magaña

Developed by L. F. Magana					
0	1	2	3	4	5
					
	6	7	8	9	10
					

One ancient American civilization that appeared in Mexico about 1800 B.C. was the Olmec. The Olmec were responsible for a 365-day calendar, the tracking of planets, and developed several systems of writings [9]. From the Olmec arose a mathematically advanced culture in Mesoamerica, who built large stone architecture and developed the concept of zero with the “first recorded zero in the Americas” occurring in a Maya carving from 357 A.D. [9, p. 22]. The Maya also developed a base-20 numbering system using three symbols, a dot or bean, a bar, and a seashell to represent zero (Table 1) [3, 5, 6, 8].

They employed these three symbols to calculate the movement of planets and developed their calendar. The Maya numbering system is a position-based system similar to the Hindu-Arabic system used in the U.S. Each digit or symbol has a value based on its position in the representation of the number [7]. In this paper, we will use a base-10 representation of the Mayan system (Table 2).

## 2 Base-10 Representation

The Mayan number system utilized a vertical representation for numbers. The number 212 is represented as two dots in the uppermost row, then one dot, followed by two dots in the lowest level row (Table 3).

The position, or place-value, is like the Hindu-Arabic system in that there is a 2 in the ones-place, a 1 in the tens place, and a 2 in the hundredths place. Using this method

**Table 3** Mayan base-10 representation for the number 212

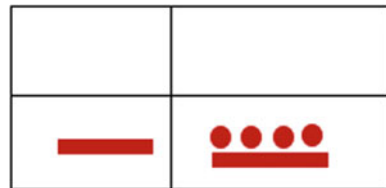
● ●	$2 \times 10^2$
●	$1 \times 10^1$
● ●	$2 \times 10^0$

for numerical operations allows us to introduce a culturally relevant pedagogy for our students. We will demonstrate addition, subtraction, and multiplication using the base-10 Mayan representation.

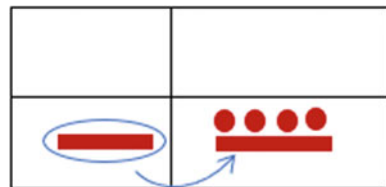
### 3 Addition

Using manipulatives and multiple representations is one of the five process standards designated by NCTM [11]. Using manipulatives such as Cuisenaire rods, the addition of  $5 + 9$  is evaluated as shown in Figs. 1, 2, 3, 4 and 5. First, the numbers 5 and 9 are represented in the first grid. Next, combine the rods into one frame. Using place-value and regrouping, we replace bars with dots in Figs. 4 and 5. In these last two steps, pre-service teachers can see the hierarchical aspect of regrouping. In this example, two bars in one level, 10-ones in this example, are regrouped as one dot in the next higher level. The two bars, 10-ones, become one in the tens place. Thus, the sum of five and nine is 14 (Fig. 5).

**Fig. 1** .



**Fig. 2** .



**Fig. 3** .

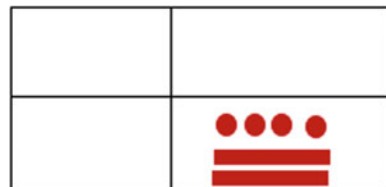
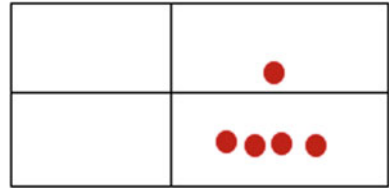




Fig. 4 .



Fig. 5 .



Three-digit addition can be modeled and evaluated to emphasize place-value and regrouping. The following example is  $212 + 89$ . The two numbers are written vertically as per the Mayan system. As depicted in the previous example, the dots and bars are combined into one grid column. Next, convert the five dots into two bars. Regroup the two bars as one dot in the next highest level or tens place. Regrouping follows the NCTM standards for elementary addition; understand the place-value structure of the base-ten number system and be able to represent and compare whole numbers and decimals [11]. The five dots in Fig. 10 are converted to two bars in Fig. 11, then regrouped as one dot in the next level. Figures 11 and 12 show that when two bars are regrouped, and there are no dots left in the tens place, the seashell, which represents zero, is used as a place holder. The final answer is 301 (Fig. 12).

Fig. 6 .

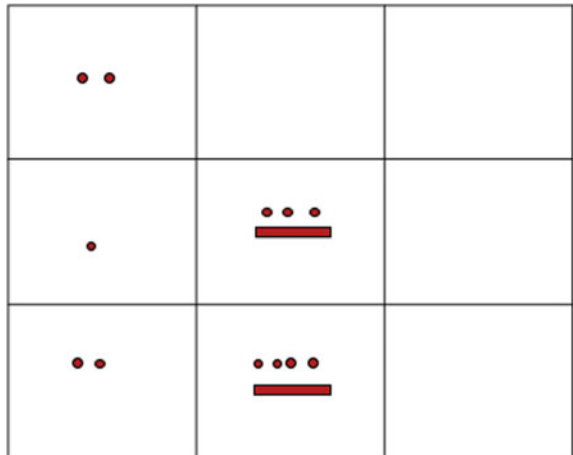


Fig. 7 .



Fig. 8 .

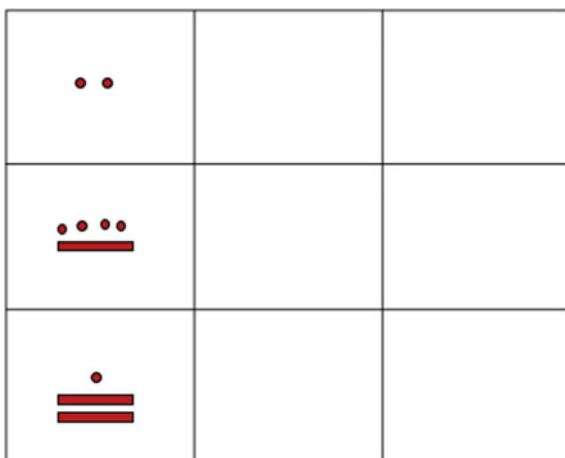


Fig. 9 .

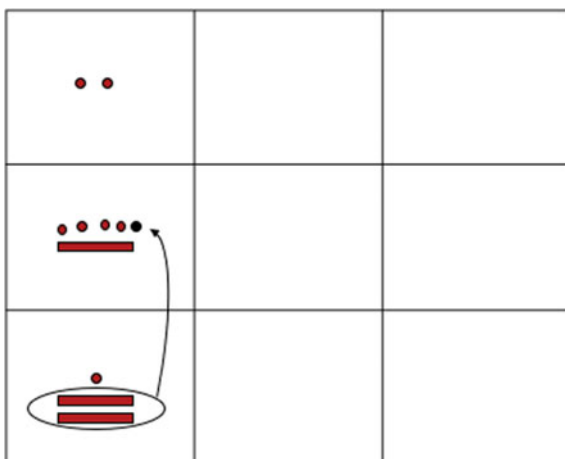


Fig. 10 .



Fig. 11 .

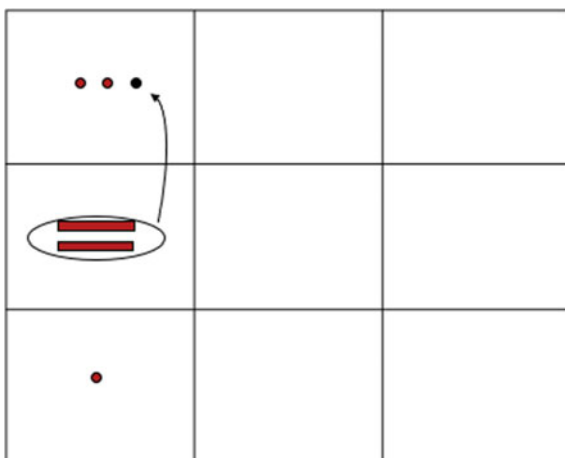
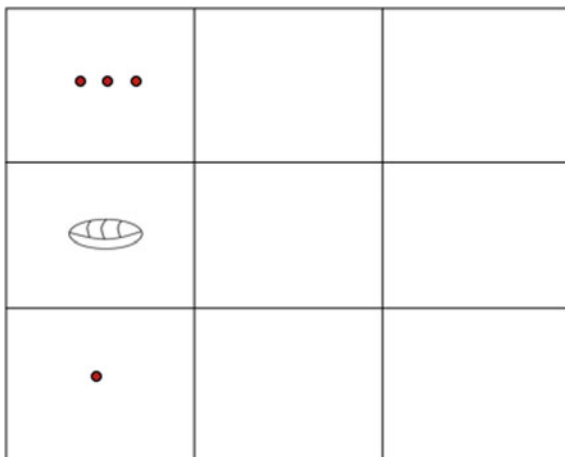


Fig. 12 .



### 4 Subtraction

The comparison model fits best with the Mayan base-10 model for subtraction. In the following example, we will evaluate  $645 - 227$ . For subtraction, the two numbers are written as vertical representations. By comparing each column, dots and bars are eliminated without calculation. The objective is to eliminate all dots and bars from one column to get to the difference between the two original columns or numbers. In Fig. 15, we compare the two columns again and consider how to get dots into an empty frame to be eliminated. Regrouping one dot from the tens to have 10-ones in the ones-place as in Fig. 16, we can then eliminate all remaining dots. Figure 18 depicts the difference between 645 and 227 is 418.

Fig. 13 .

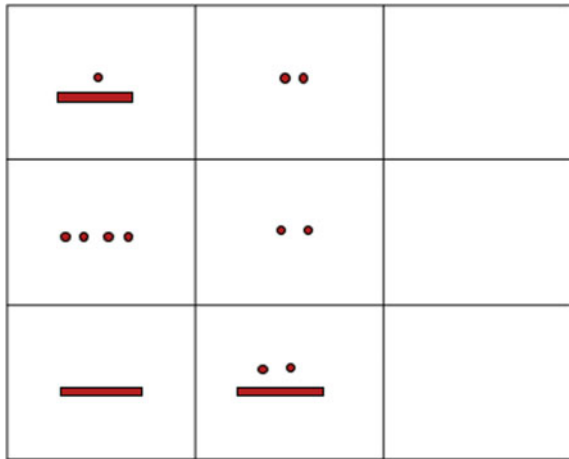


Fig. 14 .

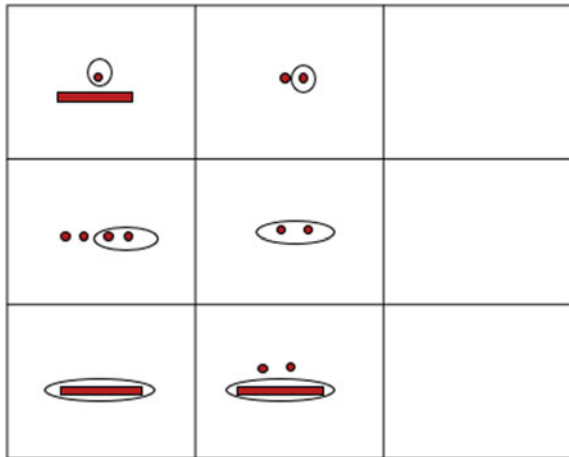


Fig. 15 .

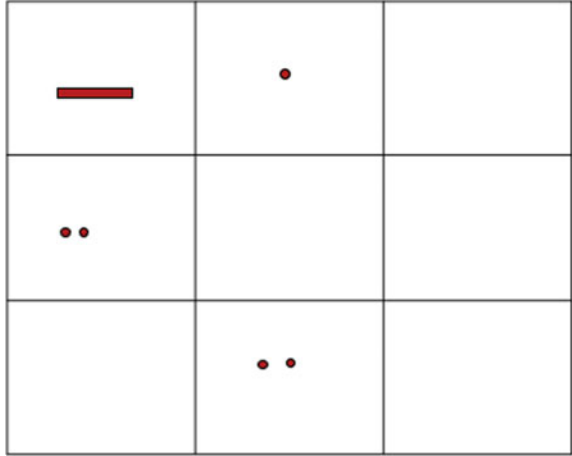


Fig. 16 .

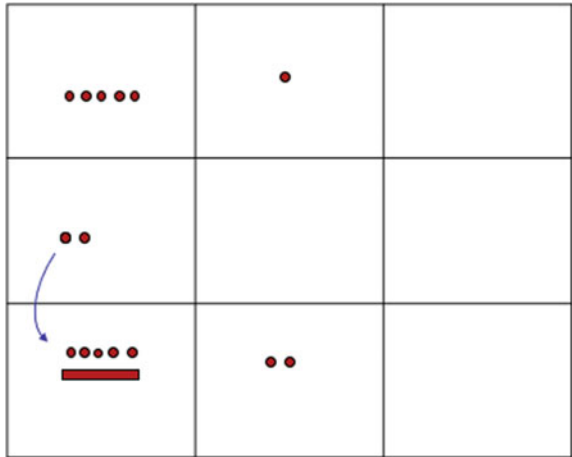


Fig. 17 .









		
		
		

Fig. 18 .

### 5 Multiplication

Our last example is multiplication. In multiplication, the numbers are placed on the outer edges of the grid, Fig. 19. In this model-based representation, we see that we are looking for one group of twos in each frame in column one, Fig. 20. Students can see that we have one group of 2-tens and one group of 2-ones. The diversity of this representation is that this process can be done by looking at the rows first. Using the first row as a starting point, we would have two groups of 1-tens and two groups of 4-ones (see Fig. 21. In either case, we know that by grouping our dots, we can fill in the frames in column two and row two in the same manner, Fig. 22). The product of these two numbers is drawn in the diagonal. We combine all dots into the diagonal, as shown in Fig. 23. Notice that the combined dots were groups of tens maintaining our place-values. In Fig. 24, two bars represent 10-tens, so 10-tens must be regrouped into the hundredths place, and any group of five dots has been converted into a bar. Figure 26 shows the solution to this product, 308.

Fig. 19 .

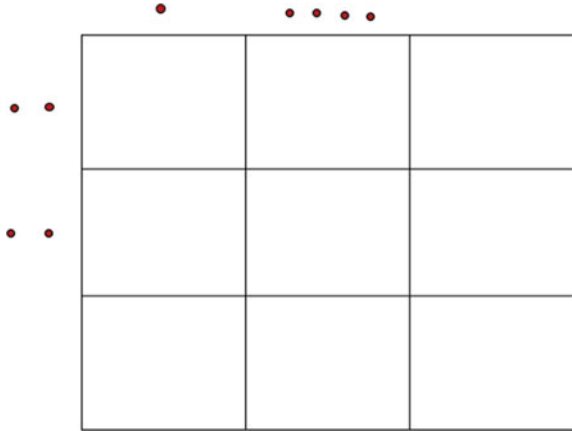


Fig. 20 .

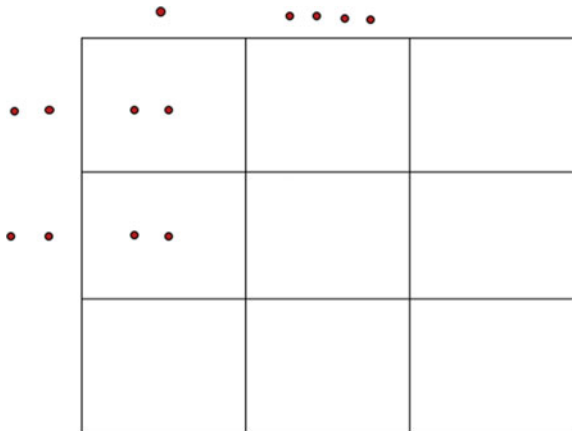


Fig. 21 .

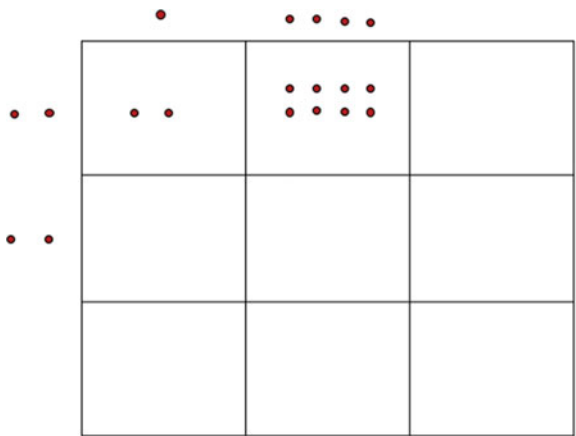


Fig. 22 .

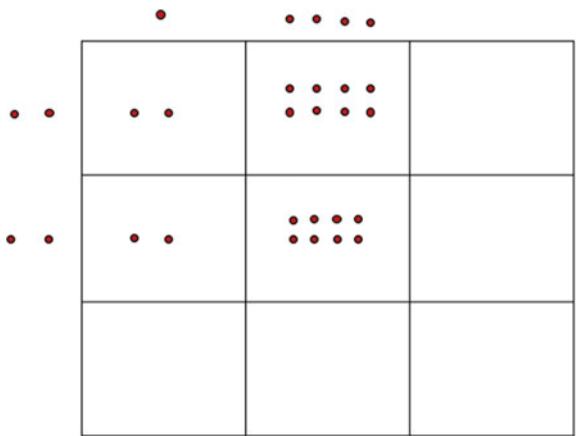


Fig. 23 .

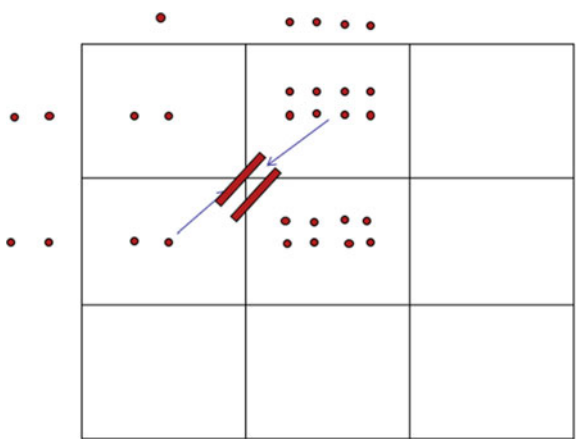




Fig. 24 .

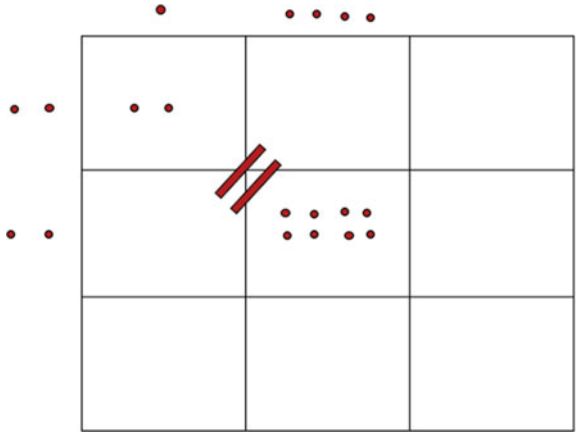


Fig. 25 .

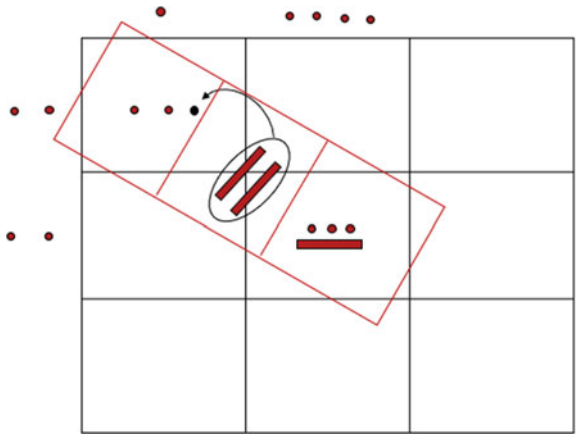
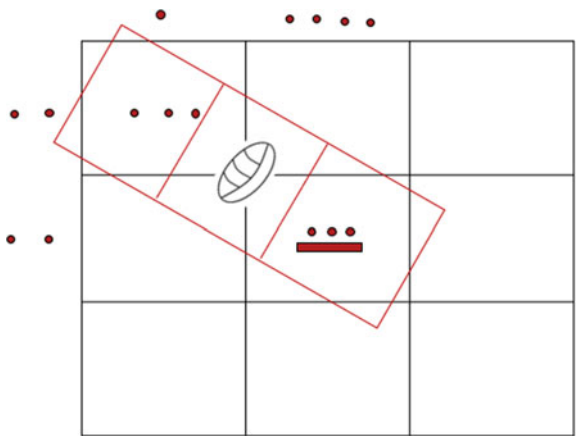


Fig. 26 .



The operations shown above are lessons in an elementary methods class presented throughout the course. We developed these lessons for an elementary methods class taught at a private liberal arts college in eastern Kentucky. Seventeen percent of the students were Latinx, 8% were African-American, 1% other, and 71% white. These students knew about Pythagoras, Newton, and other European mathematicians and historical figures. They were not aware of the mathematical contributions of the Mayans and other ancient American civilizations. One pre-service student noted: “I liked learning the Mayan-inspired lessons because it enhanced regrouping and place value, something important to know. Learning the symbols and understanding base-ten better was achieved through Mayan math.” This comment shows that students were making cognitive connections with place value and regrouping through a historical lesson.

## 6 Discussion

McGee and Hostetler [10] posit that teachers should draw on historical and contemporary narratives to position social justice in mathematics education. Another student from the elementary methods class commented: “I was able to put myself in the shoes of students who are coming into school for the first time and do not know what math is. I also learned why we regroup in more depth because you have to in ding Mayan math.” This pre-service teacher empathized with what a student might feel when learning mathematics for the first time.

The importance of diversity in our pre-service teacher education programs has become crucial in the light of the recent history in the United States through the lens of social justice. Developing culturally relevant pedagogy is rooted in social justice education and implementing the culturally diverse history of mathematics. Culturally responsive teachers can remove the standard narrative of European mathematical authority and liberate their students from oppressive educational practices and ideologies [4]. As the one presented here, historical math lessons can bridge learning mathematics better with developing culturally responsive teachers.

## References

1. Bressoud, D.: Was calculus invented in India? *Coll. Math. J.* **33**(1), 2–13 (2002)
2. Cajori, F.: Who was the first inventor of the calculus? *Am. Math. Mon.* **26**(1), 15–20 (1919)
3. Carmack, R.M.: A historical anthropological perspective on the Mayan civilization. *Soc. Evolut. Hist.* **2**(1) (2003)
4. Gay, G.: *Culturally Responsive Teaching: theory, Research, and Practice*. Teachers College Press (2018)
5. Ladson-Billings, G.: *The Dreamkeepers: successful Teachers of African American Children*. Jossey-Bass, San Francisco, CA (1994)
6. Magaña, L.F.: *Las matemáticas y los mayas*. *Ciencias* (019) (1990)

7. Magaña, L.F.: To learn mathematics: Mayan mathematics in base 10. In: International Conference on Education and New Learning (2010)
8. Magaña, L.F.: The ludic and powerful Mayan mathematics for teaching. *Procedia Soc. Behav. Sci.* **106**, 2921–2930 (2013)
9. Mann, C.C.: 1491: new Revelations of the Americas Before Columbus. Alfred A. Knopf Incorporated (2005)
10. McGee, E.O., Hostetler, A.L.: Historicizing mathematics and mathematizing social studies for social justice: A call for integration. *Equity & Excellence in Education* **47**(2), 208–229 (2014)
11. National Council of Teachers of Mathematics.: Principles and Standards for School Mathematics. Reston, VA, Author (2000)

# Why Base-20, Base-40, and Base-60 Number Systems?



Sean R. Aguilar, Olga Kosheleva, and Vladik Kreinovich

**Abstract** Historically, to describe numbers, some cultures used bases much larger than our usual base 10, namely, bases 20, 40, and 60. There are explanations for base 60, there is some explanation for base 20, but base 40—used in medieval Russia—remains largely a mystery. In this paper, we provide a possible explanation for all these three bases, an explanation based on the natural need to manage large groups of people. We also speculate why different cultures used different bases.

## 1 Formulation of the Problem

**Historical facts.** In the ancient times, in addition to our usual base-10 number system and to systems with a smaller or similar-size base, some cultures used number systems with much larger bases:

- Babylonians used the 60-based system (see, e.g., [3, 6, 7]). We still divide an hour into 60 min, a minute into 60 s—this idea originated with the ancient Babylonians.
- Ancient Romans used the base-20 system. This can still be traced to how numbers are named in modern French: for example, 80 is *quatre-vingts*, meaning four-twenties, and 96 is *quatre-vingt-seize*, meaning four-twenties-sixteen. A similar

---

S. R. Aguilar

Department of Computer Science, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA

e-mail: [sraguilar4@miners.utep.edu](mailto:sraguilar4@miners.utep.edu)

O. Kosheleva

Department of Teacher Education, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA

e-mail: [olgak@utep.edu](mailto:olgak@utep.edu)

V. Kreinovich (✉)

Department of Computer Science, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA

e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

20-based system—with 20 divided into four 5s—was used by the Mayans and by the Aztecs; see, e.g., [3–7]

- An unusual 40-based system was used in medieval Russia. For example, to describe the (large) number of churches in the medieval Moscow, the Russian chronicle says that there were 40 of 40s (sorok sorokov), i.e.,

$$40 \cdot 40 = 1600.$$

**But why?** A natural question is: why these bases and not others?

**There are answers to some “why” questions, but not to all of them.** There is a good explanation of why 60: this is the number that has unusually many divisors: it is divisible by 2, 3, 4, 5, 6, 10, 12, 15, and 20. So:

- 1/3 of a usual 60 min hour is a whole number of minutes,
- 1/4 of an hour is a whole number of minutes, etc.

This would not have been possible if we divided an hour into 100 min; see, e.g., [3, 6, 7].

There is a similar partial explanation of base 20; see, e.g., [1]. However, there is no similar explanation for selecting 40. Moreover, from the viewpoint of the above explanation of the base-60 system, the values 20 and 40 are not good at all: for example, if the Romans selected 24 or 30 instead of 20, they would have had many more divisors.

**What we do in this paper.** In this paper, we provide a possible explanation for all three number bases—an explanation based on analyzing practical problems that ancient and medieval folks faced.

## 2 Analysis of the Problem and the Resulting Explanation

**Practical problem: management.** Ancient and medieval civilizations had many activities involving large groups of people: from army to construction. The possibilities to undertake big construction projects—e.g., in irrigation or in building a protective fortress—and to have a strong army to make peaceful life possible, these possibilities are one of the main reasons why civilizations appeared in the first place.

When you have a large group of people involved in a certain activity, it is important to manage them properly.

- This problem is not as acute in the army, where the soldiers are trained to follow orders—and thus, to be managed.
- However, effective management is crucial in civilian projects, when most workers do not have special training in following orders. These workers need to be organized, and there is a need to have managers (overseers) for overseeing the organized groups of workers.

When the overall number of workers is very large, it is not enough to simply organize workers in groups—there will still be too many groups. So we need to combine groups into groups of higher level—in other words, we need to have a hierarchical organization.

**Let us start at the lowest level of the hierarchy.** On the lowest level of the hierarchy, we need to combine workers into working groups. How many people can one boss effectively oversee? To answer this question, we need to take into account that, according to psychology, there is a “seven plus minus two” law, according to which a person can only keep between  $7 - 2 = 5$  and  $7 + 2 = 9$  ideas in mind; how many depends on the person:

- some can only keep 5,
- some can keep 9;

see, e.g., [2, 8–10].

- So, to make sure that any person can serve as a supervisor of such lower-level group, we need to make sure that this group contains no more than 5 people—otherwise people who can only keep 5 ideas in their mind at the same time will not be able to effectively supervise this group.
- On the other hand, everyone can keep 5 ideas, so it will be a waste of resources to make these primary groups with fewer than 5 folks.

Thus, the ideal size of the primary group is 5.

*Comment.* This argument shows that it is reasonable to expect base-5 number systems. Such systems have actually been used by several cultures; see, e.g., [4].

**Second level of the hierarchy.** As we have mentioned earlier, even if we divide thousands of workers into groups of 5, we will get many groups. So, to effectively supervise these primary groups, we need to combine them into secondary groups.

How many primary groups should we combine into a secondary one? It is much more difficult to be a boss of bosses than simply a low-level boss of people. Because of this increased difficulty, the number of primary groups combined into a secondary group should be smaller than 5—the number of people in each primary group. So, we have 3 options:

- we can have 4 groups of 5, making up 20—which explains the base-20 system; actually, the Mayans explicitly considered 20 as 4 groups of 5;
- we can have 3 groups of 5, making up 15; historically, there is no direct evidence of base-15 systems, but there is an indirect evidence: e.g., Russia used to have 15-kopec coins, a very unusual nomination;
- we can have 2 groups of 5, making up 10; this is our usual decimal system; its representation as two groups of 5 can be seen, e.g., in the design of the abacus; see, e.g., [3, 5].

**Third level.** On the next level, it is even more difficult to manage, so the number of secondary groups that form a ternary group must be smaller than the number of primary groups in a secondary group. Here:

- For  $10 = 2 \cdot 5$ , there is no possibility to have fewer than 2 secondary groups.
- For  $15 = 3 \cdot 5$ , the only option is having 2 groups of 15 together, making it  $2 \cdot 15 = 30$ . There does not seem to be any evidence of any culture using base-30 number systems.
- For  $20 = 4 \cdot 5$ , we have two options:
  - having 3 groups of 20, making it  $3 \cdot 20 = 60$ ; and
  - having 2 groups of 20, making it  $2 \cdot 20 = 40$ .

The last two options provide an explanation of why 60 and 40 were used as bases.

**Why 60 in Babylon, 40 in Russia, and 20 in Europe: brainstorming.** The above arguments explain why 20, 40, and 60 were used as bases, but do not explain why different systems appeared in different countries—this requires going beyond mathematics, to history. We are not historians, but we can try to speculate.

Our speculation is based on the natural idea that the more obedient people are, the less they rebel, the easier it is to control them, and thus, the larger ternary groups can be formed.

From this viewpoint:

- Babylonia was ruled by mighty rulers for several centuries, so it could perform a control of the largest number of 20-size groups supervised by one person: 3. This explains why the corresponding value  $3 \cdot 20 = 60$  was used in Babylonia.
- Medieval Russia was also ruled with a heavy hand, but there were still many riots and uprisings. So, it could afford only the smaller number of 20-size groups supervised by one person: 2. This explains why the corresponding value  $2 \cdot 20 = 40$  was used in Russia.
- Finally, the Roman Empire was the site of many uprisings and revolts. This kind of explains why even combining two 20-size groups under one person was difficult—and this is why the ancient Romans only used base-20 system.

**Acknowledgements** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

## References

1. Bokati, L., Kosheleva, O., Kreinovich, V.: How can we explain different number systems? In: Ceberio, M., Kreinovich, V. (eds.) *How Uncertainty-Related Ideas Can Provide Theoretical Explanation for Empirical Dependencies*. Springer, Cham, Switzerland
2. Bokati, L., Kreinovich, V., Katz, J.: Why 7 plus minus 2? A possible geometric explanation. *Geoinformatics* **30**(1), 109–112 (2021)
3. Boyer, C.B., Merzbach, U.C.: *A History of Mathematics*. Wiley, New York (1991)
4. Heath, T.L.: *A Manual of Greek Mathematics*. Dover, New York (2003)
5. Ifrah, G.: *The Universal History of Numbers: from Prehistory to the Invention of the Computer*. John Wiley & Sons, Hoboken, New Jersey (2000)
6. Kosheleva, O.: Mayan and Babylonian arithmetics can be explained by the need to minimize computations. *Appl. Math. Sci.* **6**(15), 697–705 (2012)
7. Kosheleva, O., Villaverde, K.: *How Interval and Fuzzy Techniques Can Improve Teaching*. Springer, Cham, Switzerland (2018)
8. Miller, G.A.: The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* **63**(2), 81–97 (1956)
9. Reed, S.K.: *Cognition: theories and application*. Wadsworth Cengage Learning, Belmont, California (2010)
10. Trejo, R., Kreinovich, V., Goodman, I.R., Martinez, J., Gonzalez, R.: A realistic (non-associative) logic and a possible explanations of  $7 \pm 2$  law. *Int. J. Approx. Reason.* **29**, 235–266 (2002)



# Why Chomsky Normal Form: A Pedagogical Note



Olga Kosheleva and Vladik Kreinovich

**Abstract** To simplify the design of compilers, Noam Chomsky proposed to first transform a description of a programming language—which is usually given in the form of a context-free grammar—into a simplified “normal” form. A natural question is: why this specific normal form? In this paper, we provide an answer to this question.

## 1 Formulation of the Problem: Why Chomsky Normal Form?

**How programming languages are usually described.** The usual way to describe a programming language is by introducing special auxiliary notions. For example:

- The notion of a *digit* can be described as 0, 1, ..., or 9.
- An *unsigned integer* can be described as either a digit, or a digit followed by an integer.
- An *if-then statement* can be described as the word *if* followed by an opening parenthesis, a condition, a closing parenthesis, and a statement.

One way to describe this in precise terms is by using *context-free grammars*; see, e.g., [1]. In this description, we separate:

- symbols that will appear in the resulting program; such symbols are called *terminal*, and
- symbols describing auxiliary notions—like integer or digit—that will *not* appear in the final program; these symbols are called *variables*.

---

O. Kosheleva

Department of Teacher Education, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA

e-mail: [olgak@utep.edu](mailto:olgak@utep.edu)

V. Kreinovich (✉)

Department of Computer Science, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA

e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

In a formal description, terminal symbols are usually described by small letters, and variables by capital letters. In terms of such symbols, the above informal descriptions are written down as *rules*. For example:

- The fact that 0 is a digit ( $D$ ) can be written as  $D \rightarrow 0$ .
- Similarly, the fact that 1, ..., 9 are digits can be written as

$$D \rightarrow 1, \dots, D \rightarrow 9.$$

In general, a rule  $S \rightarrow r_1 \dots r_k$  means that if we have a combination of texts corresponding to  $r_1, \dots, r_k$ , then this combination is of type  $S$ . For example:

- The above description of an unsigned integer ( $I$ ) can be reformulated as the rules

$$I \rightarrow D \text{ and } I \rightarrow DI.$$

- The description of an if-then statement ( $T$ ) can be reformulated into the rule

$$T \rightarrow \text{if}(C)S,$$

where  $C$  means a condition and  $S$  is a statement.

In a description of a programming language, we start with a notion of a program—and in general, we start with some variable which is called *starting variable*. Then, we can repeatedly use the rules to replace each notion with its clarification—until we get to a text that only includes terminal symbols.

For example, to show that 2021 is an unsigned integer, we can start with  $I$  and then:

- first, we apply the rule  $I \rightarrow DI$ ;
- we then apply the rule  $D \rightarrow 2$  to replace  $D$  with 2, and the rule  $I \rightarrow DI$  to get

$$I \rightarrow DI \rightarrow 2DI;$$

- we apply the rule  $D \rightarrow 0$  to replace  $D$  with 0, and the rule  $I \rightarrow DI$  to get

$$I \rightarrow DI \rightarrow 2DI \rightarrow 20DI;$$

- we apply the rule  $D \rightarrow 2$  to replace  $D$  with 2, and the rule  $I \rightarrow DI$  to get

$$I \rightarrow DI \rightarrow 2DI \rightarrow 20DI \rightarrow 202D;$$

- finally, we apply the rule  $D \rightarrow 1$  to get

$$I \rightarrow DI \rightarrow 2DI \rightarrow 20DI \rightarrow 202D \rightarrow 2021.$$

**Why do we need this formal description?** But why do we need to translate a clear and understandable natural-language description into a barely understandable formal one?

The answer becomes clear if we take into account that the whole idea of a programming language is that:

- we write a program, and
- the computer will automatically translate it into executable code and implement it.

Unfortunately, computers do not understand natural language well. So, to have an automatic way of designing a compiler based on the description of the programming language, we need to translate the original description into a precise language—a language that a computer can understand.

**Need for a normal form.** There exist such “compiler compilers” that automatically produce a compiler based on the description of a programming language. Probably the best known is yacc—short of Yet Another Compiler Compiler—which is part of a usual Unix setting.

The problem is that context-free grammars can be very complicated, with long and complex rules. It is therefore desirable to be able to describe the original language in a simplified (“normal”) form.

**Chomsky normal form.** The first such simplified form was produced by Noam Chomsky, the famous linguist and the author of many concepts actively used in programming languages [1]. He showed that every context-free grammar can be transformed into a simplified form, in which only three types of rules are allowed:

- a rule  $S \rightarrow \varepsilon$ , where  $S$  is a starting variable, and  $\varepsilon$  means an empty string;
- rules of the type  $V \rightarrow a$ , where  $V$  is a variable and  $a$  is a terminal symbol; and
- rules of the type  $A \rightarrow BC$ , where  $A$ ,  $B$ , and  $C$  are variables.

**Why Chomsky normal form?** A natural question is: why these three types of rules?

In this paper, we provide an answer to this question.

## 2 Analysis of the Problem and the Resulting Explanation

**Let us restrict the length of the right-hand sides.** The longer the right-hand side of the rule, the more complex this rule. Thus, to make the description simpler, it is desirable to restrict the lengths of the right-hand sides.

The fact that every context-free grammar can be transformed into Chomsky normal form—in which every rule has right-hand side of length at most 2—shows that it is possible to have a normal form in which all these lengths are bounded by 2.

Can we bound it further, to 1 or 0? Not really: if we only have rules in which the length of the right-hand side is 0 or 1, i.e., in which the right-hand side is either an empty string or a single symbol, then, since we start with a single symbol, we will

only get one-letter words—and many programs have more than one letter. So, we do need rules in which the right-hand side has length 2.

**What rules with 2-symbol right-hand sides can we have?** Each of the two symbols on the right-hand side of a rule can be either a terminal symbol  $a, b, \dots$ , or a variable  $A, B, \dots$ . Thus, we can have four possible types of such rules:  $A \rightarrow bc$ ,  $A \rightarrow bC$ ,  $A \rightarrow Bc$ , and  $A \rightarrow BC$ .

For simplicity, it is desirable to restrict ourselves to rules of only one of these four types. Which type should we choose so that we will still be able to describe any context-free grammar in this form?

Suppose first that we only allow rules of the type  $A \rightarrow bc$ . All other rules—with right-hand sides of length 0 or 1—do not increase the length of the word. So, using rules  $A \rightarrow bc$  is the only way to get words longer than one symbol. Thus, we can get some 2-symbol words. However, these words do not contain variables, so we cannot apply any rules to make them longer. Thus, with this type of rules, we will only get 2-letter words, not enough to describe all possible programming languages.

What if we only allow rules of the type  $A \rightarrow bC$ ? It is known that such rules correspond to finite automata—every finite automaton can be represented as such a grammar if:

- we assign, to each state of this automaton, a variable, and
- we transform each transition  $a \xrightarrow{b} c$  into a rule  $A \rightarrow bC$ .

It is known that not all context-free grammars can be described by finite automata; see, e.g., [1]. So this restriction also does not allow us to represent all possible context-free languages.

Similarly, if we only allow rules of the type  $A \rightarrow Bc$ , then the resulting language consists of reverses of all the words obtained by using reversed rules  $A \rightarrow cB$ . The language of reverses is thus obtainable by a finite automaton—and hence, the original language too. So, selection of these rules also does not allow us to represent all possible context-free languages.

The only remaining case is rules of the type  $A \rightarrow BC$ , which is exactly what we have in Chomsky normal form.

**What rules with 1-symbol right-hand sides can we have?** The symbol in the right-hand side is either a terminal symbol or a variable. So, rules with a single symbol in the right-hand side are either of the form  $V \rightarrow A$  or of the form  $V \rightarrow a$ .

For simplicity, it is desirable to restrict ourselves to rules of only one of these two types. Which type should we choose so that we will still be able to describe any context-free grammar in this form?

If we only allow rules of the type  $V \rightarrow A$ , then we will never be able to introduce any terminal symbols at all. Thus, if we restrict ourselves to this type of rules, we will never be able to generate any program at all.

The only remaining case is rules of the type  $V \rightarrow a$ , which is exactly what we have in Chomsky normal form. And we need such rules—otherwise, we will not be able to get any programs at all.

**What rules with empty right-hand sides can we have?** All these rules have the form  $A \rightarrow \varepsilon$ , for some variable  $A$ . There is only one fixed variable: the starting variable. So, the only way to limit these rules is:

- either to allow these rules only for the starting variable,
- or to allow such rules only for all other variables.

In the second case, we may have many such rules, while in the first case, either one such rule or none. So the first restriction—to the starting variable  $A$ —is simpler.

This is exactly what Chomsky normal form allows. And we may need such rule—since by only using rules of the type  $A \rightarrow BC$  and  $V \rightarrow a$ —none of which decreases the length—we will never get an empty string, while some concepts in programming languages can be empty strings.

**Conclusion.** So, we explained why Chomsky normal form is used—because it is the simplest possible normal form.

**Acknowledgements** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD2034030 (CAHSI Includes).

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

## Reference

1. Sipser, M.: Introduction to the Theory of Computation. Cengage Learning, Boston, Massachusetts (2012)

# How to Best Write Research Papers: Basic English? Sophisticated English?



Martine Ceberio, Christian Servin, Olga Kosheleva, and Vladik Kreinovich

**Abstract** Instructors from English department praise our students when they use the most sophisticated grammatical constructions and the most appropriate (often rarely used) words—as long as this helps better convey all the subtleties of the meaning. On the other hand, we usually teach the students to use the most primitive Basic English when writing our papers—this way, the resulting paper will be most accessible to the international audience. Who is right? In this paper, we analyze this question by using a natural model—inspired by Zipf’s law—and we conclude that to achieve the largest possible effect, the paper should be written on an intermediate level—not too primitive, not too sophisticated (actually, on the level of the middle school).

## 1 Formulation of the Problem

**Tension between English classes and what we teach.** There seems to be a systemic tension between what our students learn in their English classes and what we teach them when describing how to write scientific papers:

---

M. Ceberio

Department of Computer Science, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA

e-mail: [longpre@utep.edu](mailto:longpre@utep.edu)

C. Servin

Computer Science and Information Technology Systems Department, El Paso Community College (EPCC), 919 Hunter Dr., El Paso, TX 79915-1908, USA

e-mail: [cservin1@epcc.edu](mailto:cservin1@epcc.edu)

O. Kosheleva

Department of Teacher Education, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA

e-mail: [olgak@utep.edu](mailto:olgak@utep.edu)

V. Kreinovich (✉)

Department of Computer Science, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA

e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

75

M. Ceberio and V. Kreinovich (eds.), *Decision Making Under Uncertainty and Constraints*, Studies in Systems, Decision and Control 217,

[https://doi.org/10.1007/978-3-031-16415-6\\_11](https://doi.org/10.1007/978-3-031-16415-6_11)

- In English classes, use of rare words and complex constructions is strongly encouraged if it provides a more adequate description of the message. If the English teacher says that the essay was written on the 5-th grade level, this is *not* a compliment, such an essay would not get an Excellent grade.
- On the other hand, when we teach students, we tell them to write in Basic English—since science is an international endeavor, and many foreign researchers do not know English that well to know rare words and rare constructions.

**Problem.** So, what is the optimal level? If we write in too complex a language, we will miss most of the audience, and the impact of the paper will be small. On the other hand, if we write in too simple a language, we do not convey many subtleties of the meaning—and thus, decrease the impact as well.

**What we do in this paper.** In this paper, we analyze this problem, and we show what is the optimal level of language complexity.

## 2 Towards Formulating the Problem in Precise Terms

**Levels of complexity.** Even native speakers of English are not born with the knowledge of all the language's words and constructions, they acquire it as they study. This provides a natural scale for the language complexity used by linguists: we can be on the level of corresponding to the average language level of kindergarten students, we can be on the level of the 1st grade, ..., level of the 12th grade, of the 1st year of college, etc. Overall, there are about 20 different levels, all the way to PhD level.

For simplicity, we will simply mark them by numbers from 1 to 20, so that Level 1 corresponds to the most basic use of language, and Level 20 to the most sophisticated use of the language.

**How widely spread are different levels.** Clearly, many folks around the world have a very basic knowledge of English—and are thus on Level 1, a little fewer are on Level 2, ..., all the way to very complex Level 20 on which there is a small minority. How many people are on each level?

A reasonable idea is to use Zipf's law (see, e.g., [1, 3, 4]) for estimating the relative number of people on each level. This law was first observed in linguistics, where it turned out that if we sort all the words from a language in the reverse order of their frequencies  $f_i$ , so that

$$f_1 \geq f_2 \geq f_3 \geq \dots,$$

then we have

$$f_n = \frac{c}{n}. \quad (1)$$

for some constant  $c$ . So, the second most frequent word is twice less frequent than the most frequent one, the third most frequent word is three times less frequent, etc.

It turned out that a similar formula (1) is ubiquitous not only in linguistics, it is ubiquitous in many other application areas (see, e.g., [2, 5])—and there are good explanations for its ubiquity; see, e.g., [1, 3].

Because of this ubiquity, it makes sense to apply this law to our situation as well, and to assume that the number of people of the  $i$ -th level of knowledge is proportional to  $1/i$ .

**What is the impact of different readers.** The overall impact of a paper comes from combining the impacts on different readers. Intuitively, it is clear that the most sophisticated—thus, the most learned—readers can provide the largest impact, both in terms of the effect on their own work and in terms of them spreading the word around, while readers who have just started doing research will have, on average, the smallest impact.

Here, readers on the last— $n$ -th level ( $n = 20$ ) have the largest impact, readers on the  $(n - 1)$ -st level have a slightly smaller impact, etc., all the way to people on the 1st level who have, on average, the smallest impact. It makes sense to use Zipf’s law to describe how this impact decreases: folks on the  $n$ -th level have the highest impact  $I$ , folks on the next  $(n - 1)$ -th level have impact  $\frac{I}{2}$ , folks on the  $(n - 2)$ -nd level have the impact  $\frac{I}{3}$ , etc., and, in general, folks on level  $i$  have the impact  $\frac{I}{n + 1 - i}$ .

The overall impact-per-unit-of-information of all the folks on level  $i$  can be obtained if we multiply the number of people on this level—which is proportional to  $\frac{1}{i}$ —and the impact of each of these folks, which is proportional to  $\frac{1}{n + 1 - i}$ . Thus, this overall impact  $I_i$  is proportional to the product

$$I_i \sim \frac{1}{i \cdot (n + 1 - i)}. \tag{2}$$

**How much information is conveyed on each level.** A big portion of information can be conveyed already on the very first Level 1. If we allow Level 2, then an additional portion of the original information can be conveyed, etc., and if we go from Level  $n - 1$  to Level  $n$ , a few very subtle places can finally be conveyed. Intuitively, as we go to a higher and higher level, the portion of new information conveyable by this new level decreases. It is therefore reasonable to us Zipf’s law to describe these portions as well: if we denote the portion that can be conveyed on Level 1 by  $p$ , then the new portion whose conveyance becomes possible on Level 2 is approximately equal to  $\frac{p}{2}$ , the new portion whose conveyance has become possible on Level 3 is approximately equal to  $\frac{p}{3}$ , etc.

So, if we use Level  $k$  to write our paper, then the portion of information conveyed by this paper can be obtained by adding up all the portions corresponding to Levels 1 through  $k$  and is, thus, proportional to the sum



$$1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{k}. \quad (3)$$

**So what is the overall impact of the paper: towards the final formula.** If we write a paper on Level  $k$ , then the portion of information that we convey is limited by folks on this level or higher. The overall impact-per-piece of information of all these folks can be obtained by adding the impacts (2) corresponding to Levels  $k$  through  $n$ :

$$\frac{1}{k \cdot (n + 1 - k)} + \frac{1}{(k + 1) \cdot (n - k)} + \dots + \frac{1}{n \cdot 1}. \quad (4)$$

Thus, the overall effect  $E$  of the paper can be obtained by multiplying the amount (3) of conveyed information and the impact (4) per piece of information:

$$E = \left(1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{k}\right) \cdot \left(\frac{1}{k \cdot (n + 1 - k)} + \frac{1}{(k + 1) \cdot (n - k)} + \dots + \frac{1}{n \cdot 1}\right). \quad (5)$$

**What we will do.** We will find the level  $k$  for which the effect  $E$  of the paper is the largest.

### 3 So Which Level Is Optimal: Towards the Answer

**Simplification.** To simplify the expression (3), let us introduce a special notation for the first factor in the expression (5):

$$S_k \stackrel{\text{def}}{=} 1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{k}. \quad (6)$$

The second factor in the expression (5) can also be represented in terms of the values  $S_i$  if we take into account that for every  $i$ , we have

$$\frac{1}{i} + \frac{1}{n + 1 - i} = \frac{n + 1}{i \cdot (n + 1 - i)}.$$

Thus,

$$\frac{1}{i \cdot (n + 1 - i)} = \frac{1}{n + 1} \cdot \left(\frac{1}{i} + \frac{1}{n + 1 - i}\right).$$

So, the sum (4) can be reformulated as

$$\frac{1}{n+1} \cdot \left( \frac{1}{k} + \dots + \frac{1}{n} + \frac{1}{1} + \dots + \frac{1}{n+1-k} \right) = \frac{1}{n+1} \cdot (S_n - S_{k-1} + S_{n+1-k}).$$

So, the expression (5) takes the form

$$E = \frac{1}{n+1} \cdot S_k \cdot (S_n - S_{k-1} + S_{n+1-k}).$$

Maximizing this expression is equivalent to maximizing the same expression but multiplied by  $n + 1$ . So, we arrive at the following conclusion.

**Resulting simplified problem.** To find the optimal level  $k$ , we must maximize the expression

$$M_k \stackrel{\text{def}}{=} S_k \cdot (S_n - S_{k-1} + S_{n+1-k}), \tag{7}$$

where  $S_i$  is described by the formula (6).

**Examples.** When we write on the most basic level, we get  $S_1 = 1$ ,  $S_n \approx 3$  and thus,

$$M_1 \approx 6.$$

When we write on the most sophisticated level, we get

$$M_n = S_{20} \cdot \frac{1}{n} \approx 3.0 \cdot \frac{1}{20} = 0.15.$$

Computations show that the value  $M_k$  is the largest for  $k = 5$ , in which case  $M_k \approx 8.4$ . This effect is 40% higher than when writing on the most primitive Level 1, and more than 50 times higher than writing on the most sophisticated level.

**Discussion.** Of course, Zipf’s law is only approximately true, so the actual optimal level may be  $k = 4$  or  $k = 6$ . However, in all these cases, we can make the following conclusion.

**Conclusion.** To achieve the largest possible effect, a research paper must be written on the level  $k \approx 5$ , crudely speaking corresponding to the middle school. This will drastically increase the effect in comparison with using the most sophisticated level.

*Comment.* In other words, in an argument between us and folks from the English department, both are wrong: if we want maximal efficiency, we should not use the most primitive level and we should not use the most sophisticated level. Instead, we should use an appropriate level in between. A consolation for us is that since this optimal Level 5 is closer to the most primitive Level 1 than to the most sophisticated Level 20, we were kind of closer to the truth.)

**Acknowledgements** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

## References

1. Cervantes, D., Kosheleva, O., Kreinovich, V.: Why Zipf's law: a symmetry-based explanation. *Int. Math. Forum* **13**(6), 255–258 (2018)
2. Kosheleva, O., Kreinovich, V.: Zipf's law and  $7 \pm 2$  principle lead to a possible explanation of Daniel's law. *Int. Math. Forum* **9**(8), 391–396 (2014)
3. Kosheleva, O., Kreinovich, V., Autcharyapanikul, K.: Commonsense explanations of sparsity, Zipf law, and Nash's bargaining solution. In: Thach, N.N., Ha, D.T., Nguyen, D.T., Kreinovich, V. (eds.) *Prediction and Causality in Econometrics and Related Topics*. Springer, Cham, Switzerland
4. Mandelbrot, B.: *The Fractal Geometry of Nature*. Freeman, San Francisco, California (1983)
5. Zapata, F., Kosheleva, O., Kreinovich, V.: Several years of practice may not be as good as comprehensive training: Zipf's law explains why. *Math. Struct. Model.* **54**, 145–148 (2020)

# **Applications to Engineering**

# How to Select Typical Objects



Mariana Benitez, Jeffrey Weidner, and Vladik Kreinovich

**Abstract** In many practical situations, we have a large number of objects, too many to be able to thoroughly analyze each of them. To get a general understanding, we need to select a representative sample. For us, this problem was motivated by the need to analyze the possible effect of an earthquake on building in El Paso, Texas. In this paper, we provide a reasonable formalization of this problem, and provide a feasible algorithm for solving thus formalized problem.

## 1 Formulation of the Problem

**General problem.** We have a large number of objects  $N$ . Each object is characterized by the values of  $q$  quantities. Let us denote the value of the  $j$ -th quantity for the  $i$ -th object by  $v_{ij}$ . Then, the object  $i$  is characterized by a tuple

$$v_i = (v_{i,1}, \dots, v_{i,q}).$$

We can only thoroughly process  $n \ll N$  objects. We therefore want to select  $n$  out of  $N$  objects so that the resulting sample of  $n$  objects be the most representative; see, e.g., [2].

**Case study.** We are interested in possible effect of an earthquake on buildings in El Paso, Texas—a potentially seismic area in which, however, earthquakes have been very rare. There are many thousands of buildings in El Paso, it is not realistic to thoroughly analyze each of them. So, we need to select a feasible-to-analyze sample.

---

M. Benitez · J. Weidner · V. Kreinovich (✉)  
University of Texas at El Paso, 500 W. University El Paso, Texas 79968, USA  
e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

M. Benitez  
e-mail: [mбенitez3@miners.utep.edu](mailto:mбенitez3@miners.utep.edu)

J. Weidner  
e-mail: [jweidner@utep.edu](mailto:jweidner@utep.edu)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023  
M. Ceberio and V. Kreinovich (eds.), *Decision Making Under Uncertainty and Constraints*, Studies in Systems, Decision and Control 217,  
[https://doi.org/10.1007/978-3-031-16415-6\\_12](https://doi.org/10.1007/978-3-031-16415-6_12)

For this problem, each building is characterized by 4 parameters: occupancy, age (i.e., equivalently, year of construction), number of stories, and height.

## 2 Main Idea and How We Can Implement It

In the general case, we want to make sure that each object is similar to one of the selected objects. How can we describe this similarity? In general, the  $q$  quantities have different effect on the properties that we want to analyze: a difference of one unit in one quantity may affect this property much more than a difference in 1 unit in some other quantity.

For example, in our case study, the 1 year difference in the building's age will have practically no effect on the building's stability against a strong earthquake, but a difference in 1 story can drastically change this stability—e.g., if we consider the difference between 1-story and 2-story buildings.

To take this into account, it makes sense to “equalize” these quantities. For example, if the effect of adding 1 story is roughly equivalent to the effect of adding  $w$  years to the age, this means that adding  $s$  stories is equivalent to adding  $w \cdot s$  years. We can estimate similar “weights” for other quantities, so that for the correspondingly equalized quantities

$$e_{i,j} \stackrel{\text{def}}{=} w_j \cdot q_{i,j} \quad (1)$$

the unit change in each of these quantities has approximately the same effect on the property of interest. In the following text, we will assume that the values of the weights have been found, and that the values of the quantities have already been equalized. In these terms, each object  $i$  is characterized by the tuple

$$e_i = (e_{i,1}, \dots, e_{i,q}).$$

In geometric terms, each tuple  $e_i$  can be represented as a point in a  $q$ -dimensional space. So, to describe the degree of dissimilarity between the two objects  $i$  and  $i'$  characterized by the tuples  $e_i = (e_{i,1}, \dots, e_{i,q})$  and  $e_{i'} = (e_{i',1}, \dots, e_{i',q})$ , it is reasonable to take the distance between these two  $q$ -dimensional points, i.e. the value

$$d(e_i, e_{i'}) \stackrel{\text{def}}{=} \sqrt{\sum_{j=1}^q (e_{i,j} - e_{i',j})^2}.$$

Our goal is to select, among  $N$  given objects  $1, \dots, N$ ,  $n$  typical objects  $t(1), \dots, t(n)$ . Once we have selected them, then, for each object  $i$ , as its approximate representation, we will take the typical object  $t(n(i))$  which is the closest to the object  $i$ , i.e., for which the distance to the  $i$ -th object is the smallest:

$$d(e_i, e_{t(n(i))}) = \min_{k=1, \dots, n} d(e_i, e_{t(k)}).$$

In general, the distance is the smallest if and only if the square of the distance is the smallest, so

$$d^2(e_i, e_{t(n(i))}) = \min_{k=1, \dots, n} d^2(e_i, e_{t(k)}). \quad (1)$$

We want to make sure that for each object  $i$  and for each (equalized) quantity  $j$ , the values of this quantity for the original object  $i$  and for the approximating typical object  $t(n(i))$  be close, i.e., that we should have  $e_{i,j} \approx e_{t(n(i)),j}$ . In other words, we want to make sure that following approximate equalities hold:

$$\begin{aligned} e_{1,1} &\approx e_{t(n(1)),1}, \dots, e_{1,q} \approx e_{t(n(1)),q}, \\ &\dots \\ e_{N,1} &\approx e_{t(n(N)),1}, \dots, e_{N,q} \approx e_{t(n(N)),q}. \end{aligned}$$

We want these approximate equalities to be as accurate as possible. This means that the distance between the tuple

$$\ell = (e_{1,1}, \dots, e_{1,q}, \dots, e_{N,1}, \dots, e_{N,q})$$

formed by all the left-hand sides and the tuple

$$r = (e_{t(n(1)),1}, \dots, e_{t(n(1)),q}, \dots, e_{t(n(N)),1}, \dots, e_{t(n(N)),q})$$

formed by all the right-hand sides should be as small as possible. As we have mentioned, the distance is the smallest if and only if the square of the distance is the smallest. Thus, we must select the typical values  $t_1, \dots, t_n$  for which the value

$$\begin{aligned} &(e_{1,1} - e_{t(n(1)),1})^2 + \dots + (e_{1,q} - e_{t(n(1)),q})^2 + \\ &\dots + \\ &(e_{N,1} - e_{t(n(N)),1})^2 + \dots + (e_{N,q} - e_{t(n(N)),q})^2 \end{aligned}$$

is the smallest possible. The sum

$$(e_{1,1} - e_{t(n(1)),1})^2 + \dots + (e_{1,q} - e_{t(n(1)),q})^2$$

of the first  $q$  terms in this expression is simply the square  $d^2(e_1, e_{t(n(1))})$  of the distance between the tuples  $e_1$  and  $e_{t(n(1))}$ . Similarly, the sum of the next  $q$  terms is the square  $d^2(e_2, e_{t(n(2))})$  of the distance between the tuples  $e_2$  and  $e_{t(n(2))}$ , etc. So, the overall expression that we want to minimize has the form

$$\sum_{i=1}^N d^2(e_i, e_{t(n(i))}).$$

In view of the formula (1), this expression takes the form

$$\sum_{i=1}^N \min_k d^2(e_i, c_k), \quad (2)$$

where we denoted  $c_k \stackrel{\text{def}}{=} e_{t(k)}$ .

Minimizing this expression is exactly the problem solved by k-means clustering (see, e.g., [1]), where each  $c_k$  is called the center of the  $k$ -th cluster. The only difference between the k-means and our problem is that:

- in the k-means clustering, we can take any point  $c_k$ , while
- in our problem,  $c_k$  must be one of the original points  $e_i$ .

Thus, after we apply the k-means clustering algorithm and get the resulting values  $c_k$ , then, for each  $k$ , we must find the point  $t(k)$  which is the closest to  $c_k$ :

$$d(e_{t(k)}, c_k) = \min_i d(e_i, c_k).$$

So, we arrive at the following algorithm.

### 3 Resulting Algorithm

We start with  $N$  objects  $i = 1, \dots, N$  characterized by tuples  $v_i = (v_{i,1}, \dots, v_{i,q})$ . Among these objects, for some pre-defined value  $n$ , we want to select  $n$  most representative ones. To do this, we use the following algorithm:

- first, for each of  $q$  quantities  $j = 1, \dots, q$ , we find the “equalizing” weight  $w_j$ , i.e., the weight such that the effect of adding 1 unit to quantity  $j$  is equivalent to the effect of adding  $w_j$  units to the quantity 1;
- then, we use the weights  $w_j$  to equalize all the values  $v_{i,j}$  into the values  $e_{i,j} = w_j \cdot v_{i,j}$ ; this way, we get  $N$  tuples  $e_i = (e_{i,1}, \dots, e_{i,q})$ ;
- next, we apply the k-means algorithm to these  $N$  tuples and find the centers  $c_1, \dots, c_n$  of the corresponding clusters;
- finally, for each  $k$  from 1 to  $n$ , we find the original tuple closest to this  $c_k$ , i.e., the tuple  $e_{t(k)}$  for which the distance  $d(e_{t(k)}, c_k)$  is the smallest possible.

As the resulting “most representative” set of  $n$  objects, we select the objects

$$t(1), \dots, t(n).$$



*Comment.* In addition to “typical” objects, we may also want to select one or more extreme objects—to make sure that we do not miss the objects for which the effect is expected to be the largest.

For example, in the earthquake-analysis case, in which the effect increases with an increase in each of the values  $v_{i,j}$ , we may want to consider the building with the largest possible value of the corresponding weighted sum  $\sum_j w_j \cdot v_{i,j}$ .

**Acknowledgements** This work was supported in part by the National Science Foundation grants:

- 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and
- HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported:

- by the AT&T Fellowship in Information Technology, and
- by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

The authors are thankful to Michael Beer from Leibniz University, Hannover, for valuable discussions.

## References

1. Bezdek, J.C.: Elementary Cluster Analysis: four Basic Methods that (Usually) Work. River Publishers, Gistrup, Denmark (2021)
2. Salomon, J., Broggi, M., Kruse, S., Weber, S., Beer, M.: Resilience decision-making for complex systems. ASCE-ASME J. Risk Uncert. Eng. Syst. Part B: Mech. Eng. **6**, 020901-1 (2020)

# Why Homogeneous Membranes Lead to Optimal Water Desalination: A Possible Explanation



Julio Urenda, Martine Ceberio, Olga Kosheleva, and Vladik Kreinovich

**Abstract** A recent experiment has shown that out of all possible biological membranes, homogeneous ones proved the most efficient water desalination. In this paper, we show that natural symmetry ideas lead to a theoretical explanation for this empirical fact.

## 1 Formulation of the Problem

**Membranes.** One of the most efficient desalinization techniques is the use of biological membranes.

**What was believed and what turned out.** Traditionally, researchers believed that the efficiency of a membrane is determined by the average values of relevant quantities—such as the average density of the proteins forming the membrane.

It was known that the knowledge of all these average values enables us to only approximately estimate the membrane's efficiency: two membranes with the same average values of the corresponding quantities may have somewhat different efficiencies.

A recent paper used innovative nano-imaging and nano-manipulating techniques to analyze and control the nano-structure of different membranes. The resulting analysis shows that the difference between efficiencies of different membranes with the same average values of the relevant quantities can be explained by the fact that dif-

---

J. Urenda · M. Ceberio · O. Kosheleva · V. Kreinovich (✉)  
University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA  
e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

J. Urenda  
e-mail: [jcurenda@utep.edu](mailto:jcurenda@utep.edu)

M. Ceberio  
e-mail: [mceberio@utep.edu](mailto:mceberio@utep.edu)

O. Kosheleva  
e-mail: [olgak@utep.edu](mailto:olgak@utep.edu)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023  
M. Ceberio and V. Kreinovich (eds.), *Decision Making Under Uncertainty and Constraints*, Studies in Systems, Decision and Control 217,  
[https://doi.org/10.1007/978-3-031-16415-6\\_13](https://doi.org/10.1007/978-3-031-16415-6_13)

ferent membranes have different degrees of homogeneity: specifically, homogenized membranes can be 50% more efficient than the usual non-homogeneous ones [1].

**How this phenomenon is explained.** According to [1], the newly observed phenomenon can be explained by the fact that density fluctuations are detrimental to water transport. This paper also mentions that there may be other factors affecting this phenomenon.

**What we do in this paper.** In this paper, we provide a general symmetry-based explanation for the observed phenomenon.

## 2 Towards a Precise Formulation of the Problem

**What does optimal mean: general reminder.** In simple situations, when the quality of an alternative can be described by a single number, optimization is usually straightforward: we select the alternative for which this number is the largest. In many practical cases, however, the situation is more complicated, we have several different numerical characteristics that need to be taken into account.

What is common in all such cases is that we should be able to decide, given two alternatives  $a$  and  $b$ :

- whether the alternative  $a$  is better; we will denote it by  $b < a$ ,
- or the alternative  $b$  is better:  $a < b$ ,
- or these two alternatives are of the same quality to the users; we will denote it by

$$a \sim b.$$

We can combine these two relations into a single preference relation  $a \leq b$  meaning that either  $b$  is better than  $a$  or  $b$  has the same quality as  $a$ . Once we know this combined relation:

- we can reconstruct  $a \sim b$  as  $(a \leq b) \& (b \leq a)$ , and
- we can reconstruct  $a < b$  as  $(a \leq b) \& (b \not\leq a)$ .

Clearly,  $a \leq a$  for all  $a$ , i.e., the relation  $\leq$  must be reflexive. Also, if  $a \leq b$  and  $b \leq c$ , then we should have  $a \leq c$ , i.e., the relation  $\leq$  should be transitive.

**Preference relation should be final.** In general, for a given preference relation, we may have several different alternatives which are optimal—in the sense that they are better than (or of the same quality as) any other alternative. For example, we may have several different membranes that are all equally efficient in terms of water desalination. In this case, we can use this non-uniqueness to optimize something else—e.g., select the membrane with the lowest cost or with the longest expected life.

From the mathematical viewpoint, this means that we replace the original preference relation  $\leq$  with a new one  $\leq'$  in which  $a \leq' b$  if and only if:

- either  $a < b$ ,
- or  $a \sim b$  and  $a \leq_1 b$  for the additional criterion  $\leq_1$ .

If after that, we still have several optimal alternatives, we can use this non-uniqueness to optimize something else, etc., until we finally get a *final* preference relation—for which there is only one optimal alternative.

**Preference relation should be invariant with respect to natural symmetries.** In many practical situations, there are natural symmetries, i.e., natural transformations with respect to which the physical situation does not change. For example, if I drop a pen, it will fall down with the acceleration of  $9.81 \text{ m/s}^2$ . If I move to another location and repeat the same experiment, I get the same result. In this sense, the situation does not change with shift. Similarly, if I rotate myself by  $90^\circ$  and repeat the experiment, I get the same result, so the situation is invariant with respect to rotations too.

For the membrane, a natural transformation is shift: if we move from one location of the membrane to another one, nothing should change—since all the related processes are local.

If there is a transformation  $T$  that does not change the physical situation, then it is reasonable to require that it should not change our preference relation: i.e., if we had  $a \leq b$  for some alternatives  $a$  and  $b$ , then for the transformed alternatives  $Ta$  and  $Tb$ , we should also have  $Ta \leq Tb$ .

Now, we are ready to formulate the problem in precise terms.

### 3 Formulation of the Problem in Precise Terms and the Resulting Explanation

**Definition 1** Let  $A$  be a set. Its elements will be called *alternatives*.

- By a *preference relation* on the set  $A$ , we mean a reflexive and transitive binary relation  $\leq$ .
- We say that an alternative  $a_{\text{opt}}$  is *optimal* with respect to a preference relation  $\leq$  if  $a \leq a_{\text{opt}}$  for all  $a \in A$ .
- We say that a preference relation is *final* if there exists exactly one alternative which is optimal with respect to this relation.

**Definition 2** Let  $T : A \rightarrow A$  be an invertible transformation.

- We say that an alternative  $a$  is *T-invariant* if  $T(a) = a$ .
- We say that the preference relation  $\leq$  is *T-invariant* if for every two alternatives  $a$  and  $b$ ,  $a \leq b$  if and only if  $T(a) \leq T(b)$ .

**Proposition 1** *For every final  $T$ -invariant preference relation, the optimal alternative  $a_{\text{opt}}$  is also  $T$ -invariant.*

**Corollary** In our case, since the physical situation does not change with shift, it is reasonable to assume that the preference relation should also be invariant with respect to shift. Thus, due to Proposition, we conclude that the optimal membrane should also not change if we shift from one point to another. Since every two locations can be transformed into each other by an appropriate shift, this means that the values of all the corresponding quantities—including density—should be the same at all the locations. In other words, this means that the optimal membrane should be homogeneous, which is exactly what the experiments show. Thus, we have indeed showed that natural symmetry requirements explain the latest experimental results.

**Proof** The main idea of this proof first appeared in [2].

Let  $\leq$  be a final and  $T$ -invariant preference relation, and let  $a_{\text{opt}}$  be the alternative which is optimal with respect to this relation. This means that  $a \leq a_{\text{opt}}$  for all  $a \in A$ . In particular, we have  $T^{-1}(a) \leq a_{\text{opt}}$ , where  $T^{-1}$  denotes the inverse function to  $T(a): b = T^{-1}(a)$  if and only if  $a = T(b)$ .

Then, due to  $T$ -invariance, we conclude that  $T(T^{-1}(a)) \leq T(a_{\text{opt}})$ , i.e., that  $a \leq T(a_{\text{opt}})$ . This is true for all  $a$ , so, by definition of an optimal alternative, the alternative  $T(a_{\text{opt}})$  is optimal. However, the preference relation  $\leq$  is final. This means that there exists only one optimal alternative. Therefore,  $T(a_{\text{opt}}) = a_{\text{opt}}$ . Thus, the optimal alternative  $a_{\text{opt}}$  is indeed  $T$ -invariant.

The proposition is proven.

**Acknowledgements** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

## References

1. Culp, T.E., Khara, B., Brickey, K.P., Geitner, M., Zimudzi, T.J., Wilbur, J.D., Jons, S.D., Roy, A., Paul, M., Ganapathysubramanian, B., Zydny, A.L., Kumar, M., Gomez, E.D.: Nanoscale control of internal inhomogeneity enhances water transport in desalination membranes. *Science* **371**(6524), 72–75 (2021). <https://doi.org/10.1126/science.abb8518>
2. Nguyen, H.T., Kreinovich, V.: *Applications of Continuous Mathematics to Computer Science*. Kluwer, Dordrecht (1997)

# Fault Detection in a Smart Electric Grid: Geometric Analysis



Hector Reyes, Dillon Trinh, and Vladik Kreinovich

**Abstract** The main idea behind a smart grid is to equip the grid with a dense lattice of sensors monitoring the state of the grid. If there is a fault, the sensors closer to the fault will detect larger deviations from the normal readings than sensors that are farther away. In this paper, we show that this fact can be used to locate the fault with high accuracy.

## 1 What Is a Smart Electric Grid

The main idea is to set up a lattice of sensors that would monitor the electric grid; see, e.g., [1]. Based on the measurement results provided by the sensors:

- we would get a good picture of the current state of the grid, and
- we would be able to effectively control it.

---

H. Reyes · D. Trinh · V. Kreinovich (✉)

Department of Computer Science, University of Texas at El Paso, El Paso, TX 79968, USA

e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

H. Reyes

e-mail: [hareyes2@miners.utep.edu](mailto:hareyes2@miners.utep.edu)

D. Trinh

e-mail: [dgtrinh@miners.utep.edu](mailto:dgtrinh@miners.utep.edu)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023  
M. Ceberio and V. Kreinovich (eds.), *Decision Making Under Uncertainty and Constraints*, Studies in Systems, Decision and Control 217,  
[https://doi.org/10.1007/978-3-031-16415-6\\_14](https://doi.org/10.1007/978-3-031-16415-6_14)



### 2 How the Grid of Sensors Can Detect Faults

Each sensor measures characteristics of the electric current at its location. Each fault affects all the sensors, some more, some less.

By observing the changes in the sensor signals, we can detect the existence of the fault. We can also get some information of the fault's location.

Sensors closer to the fault's location will detect a stronger change in their measurements results than sensors which are further away. Thus, by comparing the measurement results of the two sensors, we can decide whether the fault is:

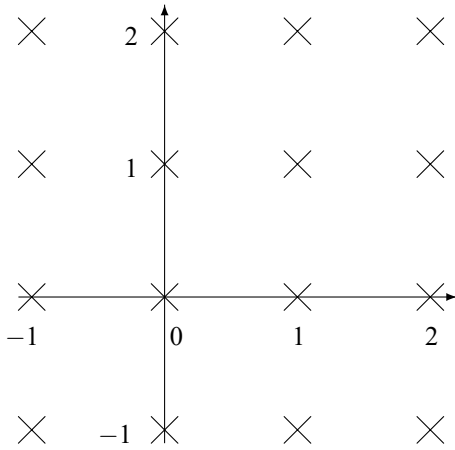
- closer to the first sensor or
- closer to the second sensor.

### 3 Let Us Describe This Situation in Precise Terms

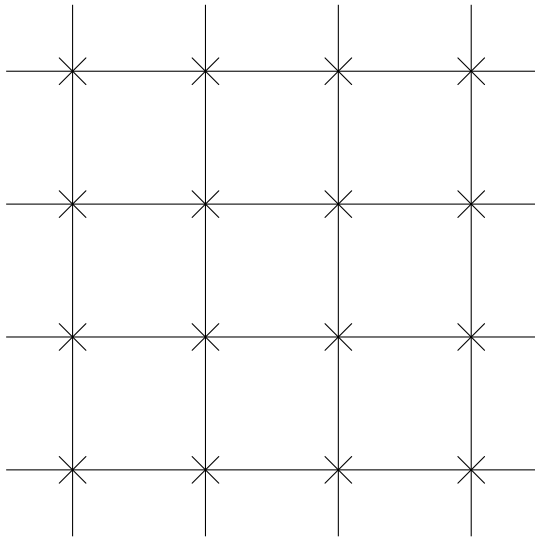
Let us consider the case when the sensors form a (potentially infinite) rectangular lattice. For simplicity of analysis, let us select a coordinate system in which:

- the location of one the sensors is the starting point  $(0, 0)$ , and
- the distance between the closest sensors is used as a measuring unit.

In this coordinate system, sensors are located at all the points  $(a, b)$  with integer coordinates.

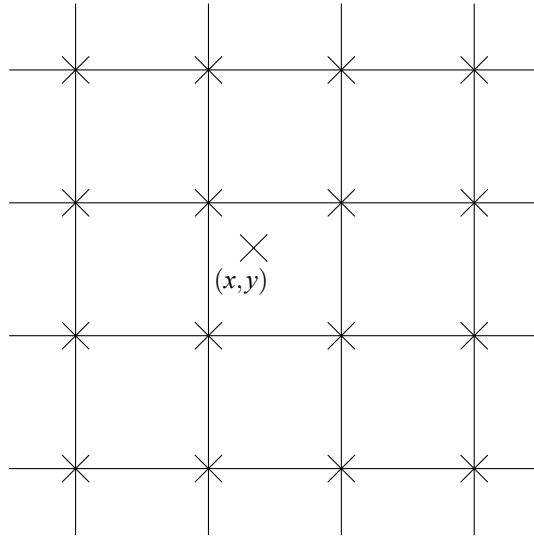


These sensors divide the plane into squares  $[a, a + 1] \times [b, b + 1]$ .



Each spatial location  $(x, y)$  is in one of these squares:



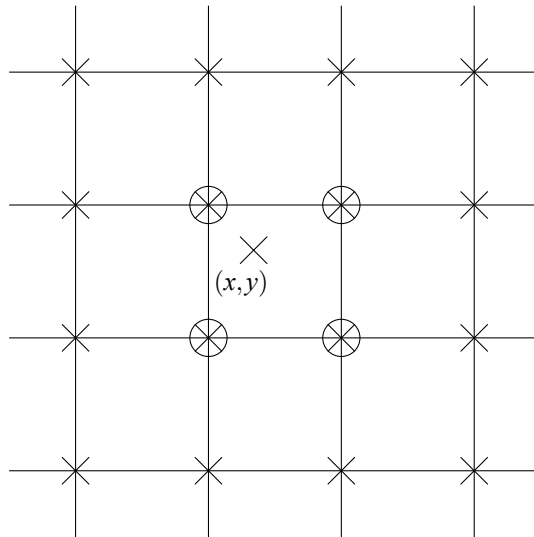


One can easily check that:

- for each spatial location within a square,
- the vertices  $(a, b)$ ,  $(a, b + 1)$ ,  $(a + 1, b)$ , and  $(a + 1, b + 1)$  of this square are the closest grid points.

Thus:

- by finding the 4 sensors at which the disturbance signal is the strongest,
- we can find the square that contains the location of the fault.



### 4 Research Question

Can we determine the location of the fault more accurately than “somewhere in the square”?

### 5 Our Answer

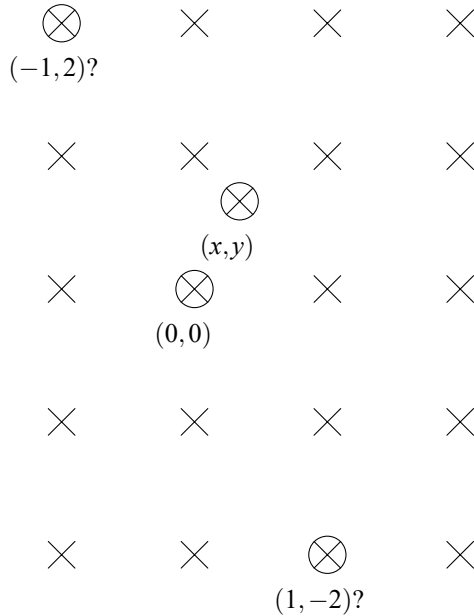
We show that, in principle:

- by using the lattice of sensors,
- we can locate the fault with any desired accuracy.

Indeed, without losing generality, let us assume that the square containing the fault is the square  $[0, 1] \times [0, 1]$ . In other words, we know that the coordinates  $(x, y)$  of the fault satisfy the inequalities  $0 \leq x \leq 1$  and  $0 \leq y \leq 1$ .

For each pair of positive integers  $(p, q)$ , we can check whether

- the sensor at  $(p, -q)$  gets a stronger signal than
- the sensor at  $(-p, q)$ .



The first sensor’s signal is stronger if and only if:

- the squared distance  $d^2(f, s_1) = (x - p)^2 + (y - (-q))^2$  between the fault  $f$  and the first sensor  $s_1$  is smaller than
- the squared distance  $d^2(f, s_2) = (x - (-p))^2 + (y - q)^2$  to the second sensor.

One can check that  $d^2(f, s_1) < d^2(f, s_2)$  if and only if  $q \cdot y < p \cdot x$ , i.e., if and only if

$$\frac{y}{x} < \frac{p}{q}.$$

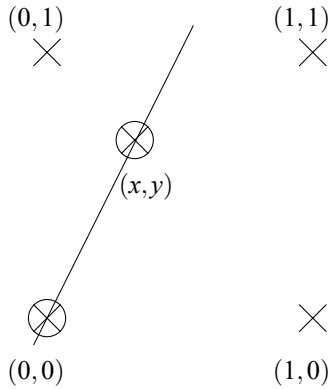
A real number can be uniquely determined if we know:

- which rational numbers  $p/q$  are smaller than this number and
- which are larger.

Thus:

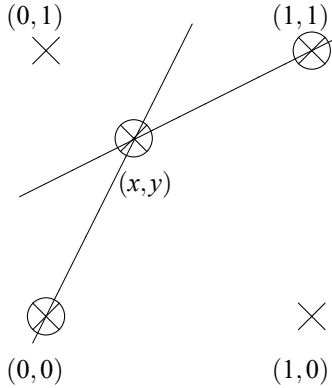
- by comparing signals from different sensors,
- we can determine the ratio  $r \stackrel{\text{def}}{=} y/x$  with any given accuracy.

Hence, we can determine the line  $y = r \cdot x$  going through  $(0, 0)$  that contains the fault:



Similarly, we can find a straight line going through the point  $(1, 1)$  that contains the fault. Thus:

- the fault’s location can be uniquely determined
- as the intersection of these two straight lines.



**Acknowledgements** This work was supported in part by the National Science Foundation grants:

- 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and
- HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported:

- by the AT&T Fellowship in Information Technology, and
- by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

The authors are thankful to all the participants of the 26th Annual UTEP/NMSU Workshop on Mathematics, Computer Science, and Computational Science (El Paso, Texas, November 5, 2021) for valuable discussions.

## Reference

1. Momoh, J.: Smart Grid: Fundamentals of Design and Analysis. IEEE Press, Piscataway, New Jersey (2012)

# **Applications to Geosciences**

# Why Geological Regions?



Daniela Flores, Olga Kosheleva, and Vladik Kreinovich

**Abstract** In most practical applications, we approximate the spatial dependence by smooth functions. The main exception is geosciences, where, to describe, e.g., how the density depends on depth and/or on spatial location, geophysicists divide the area into regions on each of which the corresponding quantity is approximately constant. In this paper, we provide a possible explanation for this difference.

## 1 Formulation of the Problem

In many practical problems, we want to describe how the value of some quantity  $q$  depends on the 2D or 3D spatial location  $x$ . This can be the description:

- of an electromagnetic field or
- of the state of the atmosphere

In most such situations, we use smooth (differentiable) functions to describe the dependence  $q(x)$ . However, in geological sciences, the usual description consists of dividing the spatial area into *geological regions*. These are zones in each of which the value  $q$  is assumed to be constant.

So why, in geosciences, this different approximating approach is more successful?

---

D. Flores · V. Kreinovich (✉)

Department of Computer Science, University of Texas at El Paso, El Paso, Texas 79968, USA

e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

O. Kosheleva

Department of Teacher Education, University of Texas at El Paso, El Paso, Texas 79968, USA

e-mail: [olgak@utep.edu](mailto:olgak@utep.edu)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

M. Ceberio and V. Kreinovich (eds.), *Decision Making Under Uncertainty*

and *Constraints*, Studies in Systems, Decision and Control 217,

[https://doi.org/10.1007/978-3-031-16415-6\\_15](https://doi.org/10.1007/978-3-031-16415-6_15)

## 2 Our Idea

In general, a natural way to describe an unknown function is to select an orthonormal basis  $e_1(x), e_2(x), \dots$ . Then, each function  $q(x)$  can be represented as

$$q(x) = \sum_{i=1}^{\infty} c_i \cdot e_i(x),$$

where  $c_i = \int q(x) \cdot e_i(x) dx$ . So, with any desired accuracy, we can approximate the function  $q(x)$  as

$$q(x) \approx \sum_{i=1}^n c_i \cdot e_i(x),$$

for a sufficiently large  $n$ .

In practice, we only know approximate values  $\tilde{q}(x) \approx q(x)$ . So we get

$$\tilde{q}(x) \approx \sum_{i=1}^n \tilde{c}_i \cdot e_i(x),$$

where  $\tilde{c}_i = \int \tilde{q}(x) \cdot e_i(x) dx$ .

We want to select the basis  $e_i(x)$  for which this approximation is as accurate as possible. How can we measure this accuracy?

## 3 How Can We Measure Approximation Accuracy: Usual Case

How can we measure approximation accuracy? This depends on the application.

In weather prediction, we are not trying to predict the temperature or the wind speed at every single location in the city. Understandably:

- some areas will be more windy, some less windy,
- some slightly warmer, some slightly colder.

What we want to predict is average temperature over some area, average wind speed, etc.

In such situations, a reasonable measure of accuracy is the usual “average” (mean square) difference  $\int (q(x) - \tilde{q}(x))^2 dx$ .

## 4 Geosciences are Different

In contrast, in geosciences, we are usually interested in specific locations.

- It is useless to learn that on average, the area contains some oil. We want to know where exactly is this oil.
- It makes sense to predict the weather in Southern California in general. However, it would be useless to just say that this is a seismic zone. We want to know which areas are more vulnerable to future earthquakes.

In all these cases, we want to make sure that the value  $q(x)$  at each location  $x$  is accurately approximated, with some accuracy  $\varepsilon > 0$ .

## 5 The Resulting Explanation: Formulation of the Result

We want to make sure that the sum of the terms  $\tilde{c}_i \cdot e_i(x)$  approximates the sum of the terms  $c_i \cdot e_i(x)$ . It is reasonable to require that each term  $\tilde{c}_i \cdot e_i(x)$  is as close to the corresponding ideal term  $c_i \cdot e_i(x)$  as possible.

In other words, we want to minimize the worst-case approximation error

$$A \stackrel{\text{def}}{=} \max_{x, q(x), \tilde{q}(x)} |\tilde{c}_i \cdot e_i(x) - c_i \cdot e_i(x)|.$$

Here:

- we denoted  $c_i = \int q(x) \cdot e_i(x) dx$  and  $\tilde{c}_i = \int \tilde{q}(x) \cdot e_i(x) dx$ , and
- the maximum is taken over all the functions  $q(x)$  and  $\tilde{q}(x)$  for which, for all  $x$ , we have

$$|\tilde{q}(x) - q(x)| \leq \varepsilon.$$

It turns out that the smallest value of this worst-case approximation error  $A$  is attained when the function  $e_i(x)$  is piece-wise constant.

This explains why such an approximation—corresponding to geological regions—is indeed very effective in geosciences.

## 6 Proof

We want to minimize the expression

$$A \stackrel{\text{def}}{=} \max_{x, q(x), \tilde{q}(x)} |\tilde{c}_i \cdot e_i(x) - c_i \cdot e_i(x)|.$$



Here,  $\tilde{c}_i \cdot e_i(x) - c_i \cdot e_i(x) = \Delta c_i \cdot e_i(x)$ , where

$$\Delta c_i \stackrel{\text{def}}{=} \tilde{c}_i - c_i = \int \Delta q(x) \cdot e_i(x) dx \text{ and } \Delta q(x) \stackrel{\text{def}}{=} \tilde{q}(x) - q(x).$$

Thus,

$$A = \max_{x, \Delta q(x)} |\Delta c_i \cdot e_i(x)| = \max_{x, \Delta q(x)} (|\Delta c_i| \cdot |e_i(x)|).$$

The only condition on  $\Delta q(x)$  is that  $|\Delta q(x)| \leq \varepsilon$ .

The maximized expression  $|\Delta c_i| \cdot |e_i(x)|$  is the product of two terms:

- the term  $|\Delta c_i|$  only depends on  $\Delta q(x)$ , and
- the term  $|e_i(x)|$  only depends on  $x$ .

Thus,

$$A = \left( \max_{\Delta q(x)} |\Delta c_i| \right) \cdot \left( \max_y |e_i(y)| \right).$$

The largest value of the sum  $\Delta c_i = \int \Delta q(x) \cdot e_i(x) dx$  is attained when each term  $\Delta q(x) \cdot e_i(x)$  is the largest.

- When  $e_i(x) \geq 0$ , maximum is attained when  $\Delta q(x)$  is the largest  $\Delta q(x) = \varepsilon$ , then  $\Delta q(x) \cdot e_i(x) = \varepsilon \cdot e_i(x)$ .
- When  $e_i(x) \leq 0$ , maximum is attained when  $\Delta q(x)$  is the smallest  $\Delta q(x) = -\varepsilon$ , then  $\Delta q(x) \cdot e_i(x) = -\varepsilon \cdot e_i(x)$ .

In both cases, the largest value is equal to  $\varepsilon \cdot |e_i(x)|$ . Thus:

$$\max_{\Delta q(x)} |\Delta c_i| = \max_{\Delta q(x)} \left| \int \Delta q(x) \cdot e_i(x) dx \right| = \int \varepsilon \cdot |e_i(x)| dx = \varepsilon \cdot \int |e_i(x)| dx.$$

So,

$$A = \varepsilon \cdot \left( \int |e_i(x)| dx \right) \cdot \max_y |e_i(y)|.$$

Minimizing  $A$  is equivalent to minimizing

$$J \stackrel{\text{def}}{=} \frac{A}{\varepsilon} = \left( \int |e_i(x)| dx \right) \cdot \max_y |e_i(y)|.$$

The functions  $e_i(x)$  are orthonormal, so

$$\int e_i^2(x) dx = \int |e_i(x)| \cdot |e_i(x)| dx = 1.$$

For each  $x$ , we have  $|e_i(x)| \leq \max_y |e_i(y)|$ . So:

$$1 = \int |e_i(x)| \cdot |e_i(x)| dx \leq \int \left( \max_y |e_i(y)| \right) \cdot |e_i(x)| dx =$$

$$\max_y |e_i(y)| \cdot \int |e_i(x)| dx = J.$$

If at least for one  $x$ , we have  $|e_i(x)| \cdot |e_i(x)| < \left( \max_y |e_i(y)| \right) \cdot |e_i(x)|$ , then  $1 < J$ .

The smallest possible value  $J = 1$  is therefore attained if for all  $x$ , we have:

$$|e_i(x)| \cdot |e_i(x)| = \left( \max_y |e_i(y)| \right) \cdot |e_i(x)|. \quad (1)$$

- If  $|e_i(x)| = 0$ , the equality (1) is satisfied.
- If  $|e_i(x)| \neq 0$ , then we can divide both side of the equality (1) by  $|e_i(x)|$  and get

$$|e_i(x)| = \max_y |e_i(y)|.$$

So, for each  $x$ , the value of  $e_i(x)$  is:

- either equal to 0,
- or equal to  $\pm \max_y |e_i(y)|$ .

Thus, the optimal function  $e_i(x)$  is indeed piecewise-constant. The statement is proven.

*Comment.* Ideas of this proof are similar to the ideas from [1].

**Acknowledgements** This work was supported in part by the National Science Foundation grants:

- 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and
- HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported:

- by the AT&T Fellowship in Information Technology, and
- by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

The authors are thankful to all the participants of the 26th Annual UTEP/NMSU Workshop on Mathematics, Computer Science, and Computational Science (El Paso, Texas, November 5, 2021) for valuable discussions.

## Reference

1. Brito, A.E., Kosheleva, O.: Interval + Image = Wavelet: for image processing under interval uncertainty, wavelets are optimal. *Reliable Comput.* **4**(3), 291–301 (1998)

# **Applications to Machine Learning**

# Why, in Deep Learning, Non-smooth Activation Function Works Better Than Smooth Ones



Daniel Cruz, Ricardo Godoy, and Vladik Kreinovich

**Abstract** Since in the physical world, most dependencies are smooth (differentiable), traditionally, smooth functions were used to approximate these dependencies. In particular, neural networks used smooth activation functions such as the sigmoid function. However, the successes of deep learning showed that in many cases, non-smooth activation functions like  $\max(0, z)$  work much better. In this paper, we explain why in many cases, non-smooth approximating functions often work better—even when the approximated dependence is smooth.

## 1 Formulation of the Problem

**The world is mostly smooth.** In the physical world, most dependencies are smooth (differentiable)—phase transitions and explosions are a few exceptions; see, e.g., [4, 9].

**Because of this, we usually try smooth models.** Because most real-life dependencies are smooth, a reasonable idea is to fit data with smooth dependencies.

In particular, this applies to machine learning, especially to neural networks. In a neural network, we intertwine linear combinations

$$y = c_0 + c_1 \cdot x_1 + \dots + c_n \cdot x_n$$

and non-linear steps, where the input signal  $z$  is transformed into an output  $s(z)$  for some non-linear functions  $s(z)$ . This non-linear function is known as the *activation function*.

---

D. Cruz · R. Godoy · V. Kreinovich (✉)

Department of Computer Science, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA

e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

D. Cruz

e-mail: [djcruz@miners.utep.edu](mailto:djcruz@miners.utep.edu)

R. Godoy

e-mail: [rgodoy@miners.utep.edu](mailto:rgodoy@miners.utep.edu)

Any linear function is, of course, always differentiable. So, to make sure that everything is differentiable, we need to make sure that the activation function is smooth. This is exactly what researchers did in the neural networks until the last decade, when they used smooth activation functions, e.g., the sigmoid function

$$s(z) = \frac{1}{1 + \exp(-z)};$$

see, e.g., [1].

**Surprisingly, a not-everywhere-smooth function works much better.** However, now it turned out that much better results are obtained when we use non-smooth activation functions such as rectified linear function

$$s(z) = \max(0, z)$$

which is not differentiable at the point  $z = 0$ ; see, e.g., [6].

**But why?** Why are non-smooth functions better—even if we approximate a smooth dependence?

Some explanations for why rectified linear activation function works better are possible; see, e.g., [5, 7]. However, this explanation is purely mathematical, it does not provide a clear explanation of why non-smooth functions work better than smooth ones.

## 2 Our Explanation

**Decision making—the ultimate goal of science and engineering.** One of the ultimate goals of all human activities is to make decisions. This is why we predict weather: we want to decide what to wear tomorrow. This is why we study nuclear physics—we want to find new isotopes for medical applications, new ways to generate and store energy, etc.

From this viewpoint, instead of going into technical details and analyzing how a function can be approximated, let us start with this ultimate goal, let us start with decision making.

We will show that already in the simplest case of decision making, we can find that non-smooth approximations are more efficient.

**The simplest case of decision making: majority rule.** Most decisions affect several people. Therefore, when making a decision, we need to take into account the effect of different possible decisions on different people.

Usually, different people are affected differently: e.g., when we build a plant, people living near this plant are affected much more than people living reasonably far way from this plant. In many real-life decision making, we need to take this difference into account.

Let us consider the simplest case of decision making, when all people are affected equally. In this case, the decision on whether to accept a certain proposal or not is usually decided by the majority rule (also known as voting): if the majority votes for, the proposal is accepted.

**Let us describe the majority rule in precise terms.** For simplicity, let us assume that no one abstains, that everyone votes yes or no. Let us denote the result of the  $i$ -th person's vote by  $x_i$ :

- if the  $i$ -th person voted “yes”, we take  $x_i = 1$ , and
- if the  $i$ -th person voted “no”, we take  $x_i = -1$ .

The majority rules  $y = f(x_1, \dots, x_n)$  means that:

- if most people voted for, then we should take  $y = 1$ ; and
- if most people voted against, then we should take  $y = -1$ .

This is the function that we want to approximate.

**How can we come up with a smooth approximation?** The usual way to approximate a dependence by a smooth function is to use the fact that sufficiently smooth functions can be expanded in Taylor series. So, we can take the first few terms in the corresponding series, and use the resulting polynomial sum as an approximation to the desired function.

This is the usual practice in physics, where we first use linear approximation, then—if needed—a quadratic one, etc. [4, 9]. This is how most functions are computed in a computer: e.g., to compute  $\exp(x)$ , we take into account this function's Taylor expansion

$$\exp(x) = 1 + x + \frac{x^2}{2!} + \dots + \frac{x^n}{n!} + \dots,$$

and use an approximating polynomial

$$\exp(x) \approx 1 + x + \frac{x^2}{2!} + \dots + \frac{x^n}{n!}.$$

**How many computational steps do we need to compute a polynomial?** For a generic linear polynomial

$$f(x_1, \dots, x_n) = a_0 + \sum_{i=1}^n a_i \cdot x_i,$$

we need to compute the sum of  $n$  terms, so we need  $O(n)$  computational steps.

To compute a generic quadratic polynomial

$$f(x_1, \dots, x_n) = a_0 + \sum_{i=1}^n a_i \cdot x_i + \sum_{i=1}^n \sum_{j=1}^n a_{ij} \cdot x_i \cdot x_j,$$

we need to compute the sum of  $O(n^2)$  terms, so we need  $O(n^2)$  computational steps.  
To compute a generic cubic polynomial

$$f(x_1, \dots, x_n) = a_0 + \sum_{i=1}^n a_i \cdot x_i + \sum_{i=1}^n \sum_{j=1}^n a_{ij} \cdot x_i \cdot x_j + \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n a_{ijk} \cdot x_i \cdot x_j \cdot x_k,$$

we need to compute the sum of  $O(n^3)$  terms, so we need  $O(n^3)$  computational steps.  
To compute a generic polynomial of degree  $d$

$$f(x_1, \dots, x_n) = a_0 + \sum_{i=1}^n a_i \cdot x_i + \dots + \sum_{i_1=1}^n \dots \sum_{i_d=1}^n a_{i_1 \dots i_d} \cdot x_{i_1} \cdot \dots \cdot x_{i_d},$$

we need to compute the sum of  $O(n^d)$  terms, so we need  $O(n^d)$  computational steps.

**How difficult is it to approximate majority rule by a polynomial?** It is known [2, 3, 8] that to approximate the majority-rule function  $f(x_1, \dots, x_n)$  by a polynomial, we need a polynomial of degree  $d = \Omega(n)$ —i.e.,  $d \geq c \cdot n$  for some  $c$ . This means that we need  $O(n^d) = O(n^{\Omega(n)})$ , i.e., exponentially many computational steps—which, for large  $n$ , is not practically feasible: for large  $n$ , we will need more steps than the lifetime of the Universe.

**What if we use non-smooth approximating functions?** If we allow non-smooth functions like min and max, then we can easily describe the majority rule in a very simple and easy-to-compute, as

$$f(x_1, \dots, x_n) = \max(-1, \min(x_1 + \dots + x_n, 1)).$$

Indeed:

- If most people voted “for”, this means that we have more positive terms  $x_i = 1$  than negative terms  $x_i = -1$ . Thus, the resulting sum  $x_1 + \dots + x_n$  is positive. Since all the values  $x_i$  are integers, their sum is also an integer, so it must be a positive integer. Every positive integer is greater than or equal to 1, so

$$\min(x_1 + \dots + x_n, 1) = 1.$$

Thus,

$$\max(-1, \min(x_1 + \dots + x_n, 1)) = \max(-1, 1) = 1,$$



which is exactly what we want.

- If most people voted “against”, this means that we have more negative terms  $x_i = -1$  than positive terms  $x_i = 1$ . Thus, the resulting sum  $x_1 + \dots + x_n$  is negative. Since all the values  $x_i$  are integers, their sum is also an integer, so it must be a negative integer. Every negative integer is smaller than or equal to  $-1$ , so  $\min(x_1 + \dots + x_n, 1) = x_1 + \dots + x_n$  and

$$\max(-1, \min(x_1 + \dots + x_n, 1)) = \max(-1, x_1 + \dots + x_n) = -1,$$

which is exactly what we want.

**Conclusion.** Already in the very simplest case of decision making, the use of non-smooth functions drastically decreases the computation time needed for approximating the desired dependence.

**Acknowledgements** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

## References

1. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)
2. Bun, M., Kothari, R., Thaler, J.: Quantum algorithms and approximating polynomials for composed functions with shared inputs. In: Proceedings of the ACM-SIAM Symposium on Discrete Algorithms SODA'2019, San Diego, California (2019)
3. Bun, M., Thaler, J.: Approximate degree in classical and quantum computing. ACM SIGACT News **51**(4), 48–72 (2020)
4. Feynman, R., Leighton, R., Sands, M.: The Feynman Lectures on Physics. Addison Wesley, Boston, Massachusetts (2005)
5. Fuentes, O., Parra, J., Anthony, E., Kreinovich, V.: Why rectified linear neurons are efficient: a possible theoretical explanations. In: Kosheleva, O., Shary, S., Xiang, G., Zapatrin, R. (eds.) Beyond Traditional Probabilistic Data Processing Techniques: Interval, Fuzzy, etc, pp. 603–613. Methods and Their Applications, Springer, Cham, Switzerland (2020)
6. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge, Massachusetts (2016)
7. Kreinovich, V., Kosheleva, O.: Optimization under uncertainty explains empirical success of deep learning heuristics. In: Pardalos, P., Rasskazova, V., Vrahatis, M.N. (eds.) Black Box Optimization, pp. 195–220. Machine Learning and No-Free Lunch Theorems, Springer, Cham, Switzerland (2021)
8. Tal, A.: Formula lower bounds via the quantum model. In: Proceedings of the 2017 ACM Symposium on Theory of Computing STOC'2017, Montreal, Quebec, Canada, June 19–23, 2017, pp. 1256–1268
9. Thorne, K.S., Blandford, R.D.: Modern Classical Physics: Optics, Fluids, Plasmas, Elasticity, Relativity, and Statistical Physics. Princeton University Press, Princeton, New Jersey (2017)

# Why Residual Neural Networks



Sofia Holguin and Vladik Kreinovich

**Abstract** In the traditional neural networks, the outputs of each layer serve as inputs to the next layer. It is known that in many cases, it is beneficial to also allow outputs from pre-previous etc. layers as inputs. Such networks are known as residual. In this paper, we provide a possible theoretical explanation for the empirical success of residual neural networks.

## 1 Formulation of the Problem

**What are neural networks: a brief reminder.** Lately, neural networks have shown to be the most efficient machine learning tools; see, e.g., [1]. The basic computations unit of a neural network is a *neuron*. It transforms inputs  $x_1, \dots, x_n$  into a value

$$s(a_0 + a_1 \cdot x_1 + \dots + a_n \cdot x_n) \tag{1}$$

for some constants  $a_i$ . Here  $s(x)$  is a nonlinear function known as an *activation function*. In a neural network:

- some neurons process the inputs,
- some neurons process the results of other neurons.

Usually, neurons form *layers*:

- neurons from layer 1 process inputs,
- neurons of layer 2 process the results of neurons of layer 1, etc.

---

S. Holguin · V. Kreinovich (✉)

Department of Computer Science, University of Texas at El Paso 500 W. University, El Paso, TX 79968, USA

e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

S. Holguin

e-mail: [seholguin2@miners.utep.edu](mailto:seholguin2@miners.utep.edu)

In the last layer, usually, we simply compute a linear combination of the signals from the previous layer.

**What are residual neural networks.** The main idea behind residual neural networks is that each neuron at layer  $i$  can use, as inputs:

- not only the outputs of the previous ( $i - 1$ )st layer,
- but also outputs from the layers before it: ( $i - 2$ )nd, etc.

**Residual neural networks are efficient, but why?** Empirically, residual neural networks are often more efficient than the traditional ones; see, e.g., [1]. In this paper, we provide a possible theoretical explanation for this efficiency.

## 2 Our Explanation

**Our model.** In real life applications, most dependencies are smooth. Functions describing many smooth dependencies can be expanded in Taylor series. In this case, the sum of the first few terms in these Taylor series provides a good approximation to the resulting dependence. This is how most special functions like exp, sin, etc. are usually computed. For example, the exponential function is usually computed as

$$\exp(x) \approx 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots + \frac{x^n}{n!}. \quad (2)$$

The simplest nonlinear approximation is when we take into account only constant, linear, and quadratic terms in the general Taylor expansion. Then, we consider expressions of the type

$$f(x_1, \dots, x_n) = a_0 + \sum_{i=1}^n a_i \cdot x_i + \sum_{i,j=1}^n a_{ij} \cdot x_i \cdot x_j. \quad (3)$$

This approximation is what we will consider in our model, both:

- in the description of the function that we want to approximate and
- in description of the activation function.

In both cases, we will ignore cubic and higher order terms, and assume that all these functions are quadratic.

**It is sufficient to consider neurons with activation function  $s(x) = x^2$ .** First, we show that in this approximation, we can replace each neuron by a neuron with  $s(x) = x^2$ . This can be done at the expense of changing the coefficients in the corresponding linear terms  $a_0 + a_1 \cdot x_1 + \dots$

Indeed, any nonlinear quadratic function of one variable  $s(x) = a \cdot x^2 + b \cdot x + c$ , with  $a \neq 0$ , can be represented as

$$s(x) = a \cdot \left(x + \frac{b}{2a}\right)^2 + \left(c - \frac{b^2}{4a}\right). \quad (4)$$

Thus, the output

$$y = s(a_0 + a_1 \cdot x_1 + \dots + a_n \cdot x_n) \quad (5)$$

of this neuron can be computed by the simple quadratic neuron  $s(x) = x^2$  as

$$y = a \cdot \left(\left(a_0 + \frac{b}{2a}\right) + a_1 \cdot x_1 + \dots + a_n \cdot x_n\right)^2 + \left(c - \frac{b^2}{4a}\right). \quad (6)$$

Vice versa, for each nonlinear quadratic expression  $s(x) = a \cdot x^2 + b \cdot x + c$ , from the formula (4), we conclude that

$$s\left(x - \frac{b}{2a}\right) = a \cdot x^2 + \left(c - \frac{b^2}{4a}\right), \quad (7)$$

thus

$$a \cdot x^2 = s\left(x - \frac{b}{2a}\right) - \left(c - \frac{b^2}{4a}\right), \quad (8)$$

and

$$x^2 = \frac{1}{a} \cdot s\left(x - \frac{b}{2a}\right) - \frac{1}{a} \cdot \left(c - \frac{b^2}{4a}\right). \quad (9)$$

Thus, the output

$$y = (a_0 + a_1 \cdot x_1 + \dots + a_n \cdot x_n)^2 \quad (10)$$

of the simple quadratic neuron can be computed by the neuron with activation function  $s(x)$  as

$$y = \frac{1}{a} \cdot s\left(\left(a_0 - \frac{b}{2a}\right) + a_1 \cdot x_1 + \dots + a_n \cdot x_n\right) - \frac{1}{a} \cdot \left(c - \frac{b^2}{4a}\right). \quad (11)$$

Because of this equivalence, in the following text, we will consider the simplest quadratic neuron, with activation function  $s(x) = x^2$ .

**In this approximation, one nonlinear layer is sufficient.** A general quadratic expression is a linear combination of terms  $x_i^2$ ,  $x_i \cdot x_j$ ,  $x_i$ , and 1. Each of these terms can be computed by a single layer; indeed:

- Each term  $x_i^2$  can be obtained by a single quadratic neuron.
- Each term  $x_i \cdot x_j$  can be obtained as

$$\frac{(x_i + x_j)^2 - (x_i - x_j)^2}{4}. \quad (12)$$

- Each term  $x_i$  can be obtained as

$$\frac{(x_i + 1)^2 - (x_i - 1)^2}{4}. \quad (13)$$

So one nonlinear layer is sufficient to represent any quadratic expression.

**How many neurons we need.** Let us denote by  $k$  the rank of the matrix  $a_{ij}$ . We can use new coordinates  $z_1, \dots, z_n$  in which coordinate axes are proportional to eigenvectors. Then, the given quadratic expression takes the form

$$c_0 + \sum_{i=1}^n c_i \cdot z_i + \sum_{i=1}^k c_{ii} \cdot z_i^2. \quad (14)$$

When  $k < n$ , then:

- traditional neural network needs at least  $k + 1$  neurons, since otherwise it cannot cover terms proportional to  $z_{k+1}, z_{k+2}$ , etc., but
- with residual neural network, the above formulas enables us to use only  $k$  nonlinear neurons.

This explains why residual neural networks are more efficient.

**Acknowledgements** This work was supported in part by the National Science Foundation grants:

- 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and
- HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported:

- by the AT&T Fellowship in Information Technology, and
- by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

The authors are thankful to all the participants of the 26th Annual UTEP/NMSU Workshop on Mathematics, Computer Science, and Computational Science (El Paso, Texas, November 5, 2021) for valuable discussions.

## Reference

1. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge, Massachusetts (2016)

# Why Semi-supervised Learning Makes Sense: A Pedagogical Note



Olga Kosheleva and Vladik Kreinovich

**Abstract** The main idea behind semi-supervised learning is that when we do not have enough human-generated labels, we train a machine learning system based on what we have, and we add the resulting labels (called *pseudo-labels*) to the training sample. Interesting, this idea works well, but why is somewhat a mystery: we did not add any new information so why is this working? There exist explanations for this empirical phenomenon, but most of these explanations are based on complicated math. In this paper, we provide a simple intuitive explanation.

## 1 Formulation of the Problem

**Usual (supervised) machine learning.** In the usual machine learning, we have several ( $K$ ) objects. Each object  $k = 1, \dots, K$  is characterized by parameters  $x^{(k)} = (x_1^{(k)}, \dots, x_n^{(k)})$ . For these objects, we also know the values  $y^{(k)}$  of some characteristics  $y$ . In the simplest case, the values  $y$  come from a small finite set  $Y$ —e.g., we know which object is a cat and which is a dog. In this discrete cases, the corresponding values  $y$  are called *labels*.

Based on this information, we want to come up with an algorithm that, given a new object  $x = (x_1, \dots, x_n)$ , will predict the value  $y$  corresponding to this object. This is exactly what many efficient machine learning algorithms do, including deep learning algorithms [1, 2].

Many such algorithms provide, for each object  $x$ , not only the corresponding label, they also provide a degree to which the system is confident in this label.

---

O. Kosheleva · V. Kreinovich (✉)

Department of Teacher Education, University of Texas at El Paso, 500 W. University El Paso, El Paso, TX 79968, USA

e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

O. Kosheleva

e-mail: [olgak@utep.edu](mailto:olgak@utep.edu)

**Limitations of supervised machine learning.** The more patterns  $(x^{(k)}, y^{(k)})$  we have, the more information we have about the desired dependence, the more accurate is the resulting algorithm  $y = f(x_1, \dots, x_n)$ .

When the values  $y$  come from experiments, we can get thousands and even millions of patterns, and thus reach a high accuracy of the resulting algorithm. The problem is that in many cases, the values  $y^{(k)}$  have to be provided by humans, and so it is not realistic to expect that many patterns. This is important, e.g., when we teach a computer to analyze videos or visual schemes. Because the number of patterns is limited, the resulting accuracy is not so good.

**Idea of semi-supervised learning.** Since we do have that many labeled patterns, when we train a machine learning algorithm on whatever labeled patterns we have, we get a not very accurate description. For some objects  $x$ , the system provides the estimate  $y$  with higher degrees of confidence, for some, lower.

The idea of semi-supervised learning is that, after setting some threshold  $p_0$ , we assume that all the labels assigned with confidence  $p_0$  or higher are correct, and repeat the training with the correspondingly enlarged set of patterns. To distinguish the new labels from the ones provided by human experts, the newly added labels are called *pseudo-labels*.

We can stop here or, alternatively, after this second training, we can again select labels assigned with confidence greater than or equal to  $p_0$ , add them, etc.

Interestingly, this idea works very well; see, e.g., [2].

**But why does this idea work?** At first glance, the fact that this idea works seems like magic. We did not add any new expert information, we did not add any new knowledge about the classified objects, so why does this improve the accuracy?

**What we do in this paper.** There are rather complicated mathematical explanations of why this idea works. In this paper, we provide a simpler more intuitive explanation.

## 2 An Explanation

**Simplified setting.** Our explanation is based on a simple but natural idea: if we have two classes, e.g., cats and dogs, and a new object is closer to some known cats than to all known dogs, then it is natural to classify this object as a cat.

*Comment.* Of course, this is a very simplified version of machine learning, but it is definitely one of the main ideas behind the discrete case of machine learning.

**From this viewpoint, when are we more confident.** From this viewpoint, the larger the ratio between the distance between this object and the nearest dog and its distance to the nearest cat, the more confident we are that the new object is a cat.

**So what new information do we add?** Suppose that originally, we have one sample cat and one sample dog. Suppose, for simplicity, that all cats are largely alike, while dogs differ a lot—by size, etc. Then, when we perform a crude first approximation,

most cats will be correctly classified as cats, but only dogs close in size to the original dog will be confidently classified as dogs. Let us call them *1st generation*.

When we add all cats and all 1st generation dogs as new patterns and apply training again, we will get new dogs confidently classified as dogs—namely, those that are close to dogs of 1st generation. Let us call them *2nd generation*. We then add 2nd generation dogs, etc.

At each point, we add dogs which are somewhat close to dogs from the previous generation. Any two dogs can be connected by such a sequence of not large transitions. So, at the end, we get a good classification of all the dogs.

In a nutshell: to the previous sample patterns, we added information about closeness: which objects are close to each other. Objects close to objects of known type are probable to belong to the same type.

**Illustrative example.** Let us have a simple 1-D example illustrating this explanation. Suppose that each object is characterize by only one parameter  $x_1$ , and that we have two groups of objects:

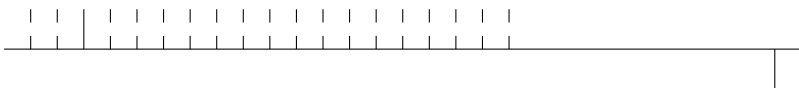
- the first group (the left one) is spread a lot, while
- in the second group (to the right) all the objects are so close together that they are practically indistinguishable.

Each object will be denoted by a dashed vertical segment:

- objects from the first group correspond to segments pointed up,
- objects from the second group are marked by segments pointing down.

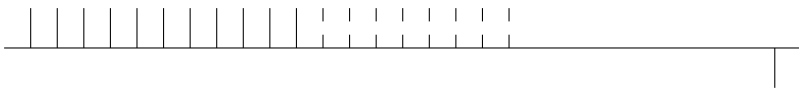
Objects for which we do not originally know the labels are marked by an interrupted segments.

Here is our original status.



Suppose that we confidently identify an object as belonging to a class if its distance to the nearest object from this class is at least twice smaller than its distance to the nearest object of another class.

In this case, after the first application of machine learning, some objects of the first class will be correctly classified as such:



However, many other objects from the first class remain unclassified. But now, we can perform the second iteration, with all newly classified objects as pseudo-labels. Now, more objects from the first class will be correctly classified:

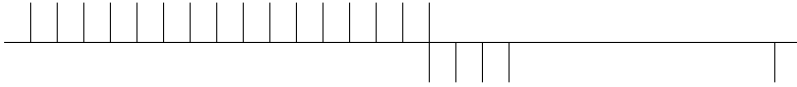




A few objects from the first class are still unclassified, but if we apply the same procedure for the third time, all objects will be correctly classified:



*Comment.* If we simply classified objects based on their closeness to the original labels, we would get several objects of the first class misclassified as belonging to the second class (with one object—as exactly the same distance from both original labels—left uncertain):



**Acknowledgements** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

## References

1. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)
2. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)

# How to Gauge the Quality of a Multi-class Classification When Ground Truth Is Known with Uncertainty



Ricardo Mendez, Osagumwenro Osaretin, and Vladik Kreinovich

**Abstract** The usual formulas for gauging the quality of a classification method assume that we know the ground truth, i.e., that for several objects, we know for sure to which class they belong. In practice, we often only know this with some degree of certainty. In this paper, we explain how to take this uncertainty into account when gauging the quality of a classification method.

## 1 Formulation of the Problem

Traditional methods of gauging the quality of a classification method assume that we know the ground truth. In other words, we assume that for some elements, we know, with certainty, to which class they belong. E.g., in medical diagnostics, we assume that for some patients, we know, with absolute certainty, what was the correct diagnosis.

In real life, however, we are rarely absolutely certain. Usually, there is some degree of uncertainty, some of the “known” classification may turn out to be wrong. Because of this, the values  $\tilde{v}$  of the quality measures that we get when we assume the known classifications to be absolutely true are, in general, different from the ideal values  $v$ —that we would have gotten if we knew the actual ground truth. How can we gauge the resulting uncertainty in  $v$ ?

---

R. Mendez · O. Osaretin · V. Kreinovich (✉)

Department of Computer Science, University of Texas at El Paso 500 W. University, El Paso, TX 79968, USA

e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

R. Mendez

e-mail: [rmendez13@miners.utep.edu](mailto:rmendez13@miners.utep.edu)

O. Osaretin

e-mail: [ooosaretin@miners.utep.edu](mailto:ooosaretin@miners.utep.edu)

In the previous papers, this problem was analyzed for the case of 2-class (“yes”–“no”) classification; see, e.g., [1]. In this paper, we start extending these ideas and results to the general multi-class case. Specifically, we analyze the uncertainty in accuracy.

## 2 Notations: Traditional Approach

Let us introduce the notations needed to describe the traditional methods—that assume that we know the ground truth.

- Let  $C$  denote the number of possible classes.
- Classes will be denoted by numbers  $c = 1, 2, \dots, C$ .
- Let  $N$  be the number of objects whose classification we know.
- Let  $P_c$  denote the set of all the objects in the  $c$ th class.
- Let  $S_c$  be the set of all objects that the tested method classifies as belonging to the  $c$ th class.
- By  $|S|$ , we denote the number of elements in the set  $S$ .

The *accuracy*  $A$  is defined as the proportion of correctly classified objects:

$$A = \frac{M}{N}, \text{ where } M \stackrel{\text{def}}{=} \sum_{c=1}^C |P_c \cap S_c|.$$

## 3 Realistic Approach: Formulation of the Problem

In practice, experts are not 100% sure about their classification.

- We have the number  $\tilde{N}$  of objects about which experts provided opinions.
- We know the sets  $\tilde{P}_c$  of all objects that experts classified to the  $i$ th class.
- For each object  $i$ , we know the expert’s probability  $p_i$  that his/her classification of this object is correct.

Based on the expert opinions, we compute the accuracy as

$$\tilde{A} = \frac{\sum_{c=1}^C |\tilde{P}_c \cap S_c|}{\tilde{N}}.$$

An important question is: how close is this estimate to the actual accuracy  $A$ ?

## 4 Our Solution

Let  $\xi(i)$  be 0 or 1 depending on whether the expert's classification of the  $i$ th object is correct. Then:

- with probability  $p_i$ , we have  $\xi(i) = 1$ , and
- with the remaining probability  $1 - p_i$ , we have  $\xi(i) = 0$ .

Thus, the mean value and the variance of these variables are

$$E[\xi(i)] = p_i \text{ and } V[\xi(i)] = p_i \cdot (1 - p_i).$$

In these terms,  $A = \frac{M}{N}$ , where:

$$N = \sum_{i=1}^{\tilde{N}} \xi(i) \text{ and } M = \sum_{c=1}^C |P_c \cap S_c| = \sum_{i \in \bigcup_{c=1}^C E_c \cap S_c} \xi(i).$$

For large  $\tilde{N}$ , a linear combination of a large number of relatively small independent random variables is, in effect, normally distributed. This follows from the Central Limit Theorem; see, e.g., [2]. Thus, both  $N$  and  $M$  are normally distributed. We can therefore find the distribution of  $A$  as the ratio of two random variables  $M/N$  with a joint normal distribution.

A joint normal distribution is uniquely determined by its means, variances, and covariance. Here:

$$E[N] = \sum_{i=1}^{\tilde{N}} p_i, \quad V[N] = \sum_{i=1}^{\tilde{N}} p_i \cdot (1 - p_i),$$

$$E[M] = \sum_{i \in \bigcup_{c=1}^C E_c \cap S_c} p_i, \quad V[M] = \sum_{i \in \bigcup_{c=1}^C E_c \cap S_c} p_i \cdot (1 - p_i).$$

Here,  $N - M$  and  $M$  contain different variables and are, thus, independent. Similarly,  $(N - E[N]) - (M - E[M])$  and  $M - E[M]$  are also independent, with mean 0. Thus:

$$E[((N - E[N]) - (M - E[M])) \cdot (M - E[M])] =$$

$$E[(N - E[N]) - (M - E[M])] \cdot E[(M - E[M])] = 0.$$

Hence, for the covariance, we get

$$C(N, M) \stackrel{\text{def}}{=} E[(N - E[N]) \cdot (M - E[M])] = \\ E[((N - E[N]) - (M - E[M])) \cdot (M - E[M])] + E[(M - E[M])^2] = V[M].$$

**Acknowledgements** This work was supported in part by the National Science Foundation grants:

- 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and
- HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported:

- by the AT&T Fellowship in Information Technology, and
- by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

The authors are thankful to all the participants of the 26th Annual UTEP/NMSU Workshop on Mathematics, Computer Science, and Computational Science (El Paso, Texas, November 5, 2021) for valuable discussions.

## References

1. Gray, N., Ferson, S., Kreinovich, V.: How to gauge the quality of a testing method when ground truth is known with uncertainty. In: Proceedings of the 9th International Workshop on Reliable Engineering Computing REC'2021, Taormina, Italy, May 16–20, 2021, pp. 265–278 (2021)
2. Sheskin, D.J.: Handbook of Parametric and Non-Parametric Statistical Procedures. Chapman & Hall/CRC, London, UK (2011)

# An AlphaZero-Inspired Approach to Solving Search Problems



Evgeny Dantsin, Vladik Kreinovich, and Alexander Wolpert

**Abstract** AlphaZero and its extension MuZero are computer programs that use machine-learning techniques to play at a superhuman level in chess, go, and a few other games. They achieved this level of play solely with reinforcement learning from self-play, without any domain knowledge except the game rules. It is a natural idea to adapt the methods and techniques used in AlphaZero for solving search problems such as the Boolean satisfiability problem (in its search version). Given a search problem, how to represent it for an AlphaZero-inspired solver? What are the “rules of solving” for this search problem? We describe possible representations in terms of *easy-instance solvers* and *self-reductions*, and we give examples of such representations for the satisfiability problem. We also describe a version of Monte Carlo tree search adapted for search problems.

## 1 Introduction

AlphaZero [10] and its extension MuZero [8] are computer programs developed by Google’s subsidiary DeepMind. They use machine-learning techniques to play at a superhuman level in chess, go, and a few other games. AlphaZero achieved this level of play solely with reinforcement learning from self-play, with no human data, no handcrafted evaluation functions, and no domain knowledge except the game rules. In comments on playing chess, the play style of AlphaZero is called “alien”: it sometimes wins by making moves that would seem unthinkable to a human chess player.

The purpose of this paper is to adapt the methods and techniques used in AlphaZero for solving search problems such as, for example, the Boolean satisfiability problem (in its search version). Reinforcement learning has been applied to combinatorial

---

E. Dantsin (✉) · A. Wolpert  
Roosevelt University, Chicago, US  
e-mail: [edantsin@roosevelt.edu](mailto:edantsin@roosevelt.edu)

V. Kreinovich  
The University of Texas at El Paso, El Paso, US

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023  
M. Ceberio and V. Kreinovich (eds.), *Decision Making Under Uncertainty and Constraints*, Studies in Systems, Decision and Control 217,  
[https://doi.org/10.1007/978-3-031-16415-6\\_20](https://doi.org/10.1007/978-3-031-16415-6_20)

optimization [6, 12], but with using expert knowledge and handcrafted heuristics, which differs these applications from AlphaZero’s approach.

To design an AlphaZero-inspired solver for a search problem  $\Pi$ , we first need to represent  $\Pi$  as a one-player combinatorial game, where the player starts from an initial position and tries to reach a winning position by making moves from one position to another. That is, we need to define what we mean by positions, winning positions, and possible moves.

We think of any instance of  $\Pi$  as a possible position in the game. For example, if  $\Pi$  is SAT, then any formula in CNF (a set of clauses) is viewed as a position. Certain instances of  $\Pi$  are thought as “easy” instances, assuming that we already have an efficient solver for such instances. An easy instance plays the role of a winning position. In the case of SAT, a set of easy instances could contain formulas with the empty clause, formulas where each clause contains a pure literal, formulas in 2-CNF, etc. By a possible move we mean a transition from an instance  $x$  to an instance  $x'$  such that the following holds:

- $x$  has a solution if and only if  $x'$  has a solution;
- a solution to  $x$  can be computed from a solution to  $x'$ .

The resolution rule gives examples of possible moves:  $x'$  is obtained from  $x$  by choosing two clauses and adding their resolvent to  $x$ . Another example is the pure literal elimination rule:  $x'$  is obtained from  $x$  by removing all clauses that contain pure literals.

Thus, we define the “game rules” for  $\Pi$  by specifying two components:

- a set of easy instances and a solver for this restriction of  $\Pi$ ;
- for each non-easy instance of  $\Pi$ , a set of possible moves from this instance.

We call such a specification a *setup* for solving  $\Pi$ . Section 2 gives a formal definition of setups in terms of *easy-instance solvers* and *self-reductions*. Section 3 gives examples of setups for SAT.

Suppose we have chosen a setup for solving  $\Pi$ . A solver for  $\Pi$  based on a given setup is described in Sect. 4. This solver uses adapted versions of two key algorithms of AlphaZero: a *reinforcement-learning algorithm* and a *parameter-adjustment algorithm*. The former one uses Monte Carlo tree search to find a sequence of moves from an input instance of  $\Pi$  to an easy instance. This algorithm has many parameters that are adjusted with help of the latter algorithm. The parameter-adjustment algorithm trains a deep neural network to find better values of the parameters; the choice of architecture of this network depends on how we represent instances of  $\Pi$ .

The solver described in Sect. 4 can also be applied to another task called *per-instance algorithm selection* [4, 5]. In this task, we wish to design a “meta-solver” that solves a search problem  $\Pi$  by automatically choosing (on a per-instance basis) a solver from a “portfolio” of solvers for  $\Pi$ , see Sect. 5 for details.

## 2 Setups for Solving Search Problems

**Search Problems** A *search problem* is one of the standard types of computational problems. It is common to represent a search problem by a binary relation  $R \subseteq X \times Y$  where  $X$  is a set of *instances* and  $Y$  is a set of *solutions*. If  $(x, y) \in R$  then  $y$  is called a *solution to  $x$* .

The *satisfiability problem* (in its search version) is an example of a search problem, see Sect. 3 for details. The corresponding set  $X$  of instances consists of Boolean formulas in CNF. The set  $Y$  of solutions consists of assignments of truth values to variables. An instance  $x \in X$  has a solution  $y \in Y$  if  $y$  is a satisfying assignment of  $x$ . There are many ways to encode instances and solutions of the satisfiability problem; no particular encoding is specified in our example.

**Solvers** Let  $\Pi$  be a search problem. A *solver* for  $\Pi$  is an algorithm  $S$  that either finds a solution, or reports that there is no solution, or may give up saying “don’t know”. That is, on every instance  $x \in X$ ,

- if  $x$  has a solution, then  $S$  returns some solution to  $x$  or says “don’t know”;
- if  $x$  has no solution, then  $S$  says “no solution” or says “don’t know”.

Solvers may have parameters, additional input, and additional output. For example, in Sect. 4, we describe a solver that takes as input not only an instance  $x$  but also additional data  $\theta$  with information about previous traces; the solver outputs an answer for  $x$  and updates  $\theta$ .

**Easy Instances** We assume that the set of instances of  $\Pi$  has a designated subset  $E \subseteq X$  whose elements are called *easy* instances. The assumption behind  $E$  is that it is “easy” to determine whether an instance  $x \in E$  has a solution and, moreover, if a solution exists, it is “easy” to find it. To formalize this assumption, we equip  $\Pi$  with an algorithm denoted by  $\mathcal{E}$  and called an *easy-instance solver*. On every instance  $x \in X$ , this algorithm determines whether  $x$  is an easy instance and, if so, the algorithm finds a solution to  $x$  or reports that  $x$  has no solution:

$$\mathcal{E}(x) = \begin{cases} \text{“not easy”} & \text{if } x \notin E \\ \text{“no solution”} & \text{if } x \in E \text{ and } x \text{ has no solution} \\ \text{some solution to } x & \text{if } x \in E \text{ and } x \text{ has a solution} \end{cases}$$

Section 3 gives examples of the set  $E$  for the satisfiability problem. For example,  $E$  can be the set of formulas  $\phi$  such that  $\phi$  is the empty set (this formula is satisfiable) or  $\phi$  contains the empty clause (this formula is unsatisfiable).

**Self-reductions and Moves** We define a *self-reduction* of  $\Pi$  to be a pair  $r = (f_r, g_r)$ , where  $f_r$  and  $g_r$  are computable functions such that for every instance  $x \in X$ ,

- $f_r(x)$  is a finite set of instances;
- if  $x$  has a solution, then each instance in  $f_r(x)$  has a solution;



- for every instance  $x' \in f_r(x)$  and for every solution  $y \in Y$ , if  $y$  is a solution to  $x'$ , then  $g_r(x, x', y)$  is a solution to  $x$ .

If  $x' \in f_r(x)$ , then we say that the self-reduction  $r$  offers a *move* from  $x$  to  $x'$ . Thus, for each instance,  $f_r$  defines the set of all moves from this instance. We call  $f_r$  the *move function* of the self-reduction  $r$ . For a move from  $x$  to  $x'$ , the function  $g_r$  computes a solution to  $x$  from a solution to  $x'$  (if any). We call  $g_r$  the *solution function* of  $r$ .

Examples of self-reductions of the satisfiability problem are given in Sect. 3. Here we just mention two of them. The first example is a self-reduction  $r = (f_r, g_r)$  where the move function  $f_r$  is in fact the *pure literal elimination* rule. This move function maps a CNF formula  $\phi$  to a one-element set  $\{\phi'\}$  where  $\phi'$  is a CNF formula obtained from  $\phi$  by successively removing all clauses containing pure literals. Another example is a self-reduction  $r = (f_r, g_r)$  that uses the *resolution rule*. For every CNF formula  $\phi$ , the set  $f_r(\phi)$  consists of all CNF formulas obtained from  $\phi$  by choosing two clauses and adding their resolvent to  $\phi$ . In both examples, the solution functions  $g_r$  are defined in the obvious way, see Sect. 3 for details.

**Paths** Let  $\mathcal{R}$  be a finite set of self-reductions of  $\Pi$ . Let  $x$  and  $x'$  be instances of  $\Pi$ . By a *path* from  $x$  to  $x'$  we mean a sequence

$$x_0, r_1, x_1, r_2, x_2, \dots, x_{n-1}, r_n, x_n$$

where  $x_0 = x$ ,  $x_n = x'$ , and  $r_i$  is a self-reduction from  $\mathcal{R}$  that offers a move from  $x_{i-1}$  to  $x_i$  for all  $i = 1, \dots, n$ . Clearly, given such a path, we have the following:

- if  $x$  has a solution, then  $x'$  also has a solution;
- if  $y$  is a solution to  $x'$ , then  $x$  has a solution that can be computed from  $y$  by successively computing solutions to  $x_{n-1}, \dots, x_1, x_0$ .

**Setups for Solving** An easy-instance solver  $\mathcal{E}$  and a finite set  $\mathcal{R}$  of self-reductions of  $\Pi$  suggest the following approach to solving  $\Pi$ :

1. Try to find a path from an input instance  $x$  to an easy instance  $x'$ .
2. If such a path is found, either return a solution to  $x$  (computed from a solution to  $x'$ ) or return “no solution” (in the event that  $x'$  has no solution). Otherwise, return “don’t know”.

The key step here is a search for a path and its success depends on the choice of  $\mathcal{E}$  and  $\mathcal{R}$ . We call the pair  $(\mathcal{E}, \mathcal{R})$  a *setup* for solving  $\Pi$ . Such a setup allows us to think of  $\Pi$  as a one-player combinatorial game, where the player tries to find a sequence of moves from an initial position to a winning position. From this point of view, a setup for solving  $\Pi$  defines the rules of this game.

Note that, in general, a setup  $(\mathcal{E}, \mathcal{R})$  is not required to be “complete” in the following sense: for every instance  $x$ , there must be a path from  $x$  to an easy instance. Section 3 shows examples of different setups for solving the satisfiability problem, including a setup where only satisfiable formulas have paths to easy instances. Solvers

based on “incomplete” setups output “don’t know” on instances that do not have paths to easy instances.

### 3 Examples for the Satisfiability Problem

There are many possible setups for solving the satisfiability problem that make sense in our context. In this section, we give three setups for the purpose of illustration.

**Satisfiability** Although the satisfiability problem is very well known and described in numerous books and articles [1], we give the basic definitions here to avoid ambiguity (notation and terminology slightly vary in the literature).

Let  $V = \{v_1, v_2, \dots\}$  be a set of *variables*. A *literal* is a variable from  $V$  or its negation; each of them is the *complement* of the other. The complement of a literal  $a$  is denoted by  $\neg a$ . A *clause* is a finite set of literals that contains no pair of complements (a clause is thought of as the disjunction of its literals). A *formula* is a finite set of clauses (a formula is thought of as the conjunction of its clauses). An *assignment* is usually defined as a function from a finite set of variables to  $\{0, 1\}$ , but it is convenient for us to use an equivalent definition: an *assignment* is a finite set of literals without any pair of complements (this set is thought of as the conjunction of its literals). An assignment  $\alpha$  *satisfies* a clause  $C$  if the intersection  $\alpha \cap C$  is not empty. An assignment  $\alpha$  *satisfies* a formula  $\phi$  if  $\alpha$  satisfies every clause of  $\phi$ ; we also call  $\alpha$  a *satisfying assignment* for  $\phi$ .

The definitions above allow the empty clause and the empty formula. No assignment satisfies the empty clause and, thus, every formula with the empty clause is unsatisfiable. The formula consisting of only the empty clause is denoted by  $\perp$ . The empty formula (“no constraints at all”) is denoted by  $\top$ . By definition,  $\top$  is satisfied by the empty assignment.

It is common to denote the following decision problem by SAT: given a formula  $\phi$ , does it have a satisfying assignment? Slightly abusing this notation, we write SAT to refer to the satisfiability problem in its search version: given a formula  $\phi$ , find a satisfying assignment or return “no solution”. In terms of Sect. 2, this search version is defined as follows. The set  $X$  of instances consists of all formulas over  $V$ . The set  $Y$  of solutions consists of all possible assignments, i.e., all finite subsets of literals over  $V$  without any pair of complements. An assignment  $\alpha \in Y$  is a solution to an instance  $\phi \in X$  if and only if  $\alpha$  satisfies  $\phi$ .

**Example 1: Setup Based on Resolutions** There are only two easy instances:  $\top$  and  $\perp$ . Thus, an empty-instance solver  $\mathcal{E}$  is trivial. A set  $\mathcal{R}$  of self-reductions consists of the following three self-reductions commonly used in SAT solving:

- *Resolution rule.* Let  $\phi$  be a formula with clauses  $C_1$  and  $C_2$  such that  $C_1$  contains a literal  $a$  and  $C_2$  contains its complement  $\neg a$ . If the set  $C_1 \cup C_2 - \{a, \neg a\}$  contains no pair of complements, then we call this set the *resolvent* of  $C_1$  and  $C_2$ . If  $\phi'$  is the formula obtained from  $\phi$  by adding this resolvent, we say that  $\phi'$  is obtained

from  $\phi$  by the *resolution rule*. The *resolution self-reduction* is a pair  $r = (f_r, g_r)$  where the move function is defined by

$$f_r(\phi) = \{\phi' \mid \phi' \text{ is obtained from } \phi \text{ by the resolution rule}\}$$

and the solution function  $g_r$  is defined as follows: for all formulas  $\phi$ , if  $\alpha$  is a solution to a formula  $\phi' \in f_r(\phi)$  then  $g_r(\phi, \phi', \alpha)$  is  $\alpha$ .

- *Subsumption rule*. If a clause  $C_1$  is a proper subset of a clause  $C_2$ , we say that  $C_1$  is *subsumed* by  $C_2$  and we call the clause  $C_2$  *unnecessary*. The *subsumption self-reduction* is the following self-reduction  $r = (f_r, g_r)$ . The move function  $f_r$  maps a formula  $\phi$  to a one-element set  $\{\phi'\}$  where the formula  $\phi'$  is obtained from  $\phi$  by removing all unnecessary clauses. The solution function  $g_r$  is obvious:  $g_r(\phi, \phi', \alpha) = \alpha$  for all  $\phi, \phi'$ , and  $\alpha$ .
- *Pure literal elimination*. A literal  $a$  in  $\phi$  is called a *pure literal* if no clause of  $\phi$  contains  $\neg a$ . The *pure literal self-reduction*  $r = (f_r, g_r)$  is defined as follows. The move function  $f_r$  maps a formula  $\phi$  to a one-element set  $\{\phi'\}$  where  $\phi'$  is obtained from  $\phi$  by successively removing all clauses containing pure literals until  $\phi'$  has no pure literals. If  $\alpha$  is a satisfying assignment for  $\phi'$ , then  $g_r(\phi, \phi', \alpha)$  is the extension of  $\alpha$  that assigns “true” to all pure literals in  $\phi$ .

This setup  $(\mathcal{E}, \mathcal{R})$  is “complete”: for every formula  $\phi$ , there is a path from  $\phi$  to either  $\top$  or  $\perp$ , see for example [2].

**Example 2: Setup Based on Resolutions and the Extension Rule** The setup described above can be extended by adding a self-reduction based on the *extension rule* [11]. Let  $\phi$  be a formula and let  $v$  be a variable not appearing in  $\phi$ : no clause of  $\phi$  contains  $v$  or  $\neg v$ . Let  $a$  and  $b$  be literals such that their underlying variables appear in  $\phi$ . The extension rule adds clauses

$$\{a, \neg v\}, \{b, \neg v\}, \{\neg a, \neg b, v\}$$

to  $\phi$ . In the corresponding self-reduction  $r = (f_r, g_r)$ , the move function  $f_r$  is defined by

$$f_r(\phi) = \{\phi' \mid \phi' \text{ is obtained from } \phi \text{ by the extension rule}\}$$

and the solution function  $g_r$  is the same as in the resolution self-reduction:  $g_r(\phi, \phi', \alpha)$  is  $\alpha$ .

The extension rule makes resolution proof systems much stronger, but there are no good heuristics for choosing extension literals  $a$  and  $b$ . This problem of using the extension rule in practical SAT solvers is discussed in [2, section 7.8], where the authors note that “if this could be done well, the gains would be enormous” and “the main bottleneck appears to be that we have no good heuristics for how to choose extension formulas”.

**Example 3: Setup Based on Flipping** A variable is called a *positive* literal; its negation is called a *negative* literal. We define an easy instance to be a formula in

which every clause has at least one positive literal. Obviously, such a formula is satisfied by the set of these positive literals. An algorithm  $\mathcal{E}$  determines whether an instance  $\phi$  is an easy instance and if so,  $\mathcal{E}(\phi)$  is the corresponding set of positive literals. The *flipping rule* transforms a formula  $\phi$  taking the following two steps:

1. Choose a clause  $C \in \phi$  in which all literals are negative (if  $\phi$  is not an easy instance, such a clause exists).
2. Choose a literal  $\neg v_i \in C$  and “flip” all of its occurrences in  $\phi$ , i.e., replace  $\neg v_i$  with  $v_i$  everywhere in  $\phi$ .

We can define  $\mathcal{R}$  to be a set of one or more self-reductions based on the *flipping rule*. The move function in a such a self-reduction maps  $\phi$  into a set of formulas obtained from  $\phi$  by applying the flipping rule. Note that the setup  $(\mathcal{E}, \mathcal{R})$  is not “complete”. If  $\phi$  is satisfiable, then there is a path from  $\phi$  to an easy instance. Otherwise,  $\phi$  has no path to any easy instance (all easy instances are satisfiable).

## 4 Solvers Based on Setups

Let  $(\mathcal{E}, \mathcal{R})$  be a setup for solving a search problem  $\Pi$ . We describe a solver for  $\Pi$  based on this setup. This solver, denoted by  $\mathcal{S}$ , tries to find a path from an input instance  $x$  to an easy instance and, if such a path is found,  $\mathcal{S}$  outputs an answer for  $x$ . The solver has parameters whose values change from run to run, and  $\mathcal{S}$  updates these values itself. The key point is that  $\mathcal{S}$  uses machine-learning techniques for both tasks, namely, for a path search and for updating values of the parameters. Roughly,  $\mathcal{S}$  uses a reinforcement-learning algorithm  $\mathcal{A}_1$  to search for a path and it uses a parameter-adjustment algorithm  $\mathcal{A}_2$  to search for “better” values of the parameters. We first describe a bird’s eye view of  $\mathcal{S}$  and then give more details.

**Input and Output** The input to  $\mathcal{S}$  has two parts: an instance  $x \in X$  and a binary string  $\theta \in \{0, 1\}^*$  called a *parameter string*. This string encodes values of the parameters of  $\mathcal{S}$  and information about the solver’s previous traces. We assume that  $\theta$  is stored in a data store outside  $\mathcal{S}$ . We also assume that  $\theta$  is initialized before the first run of  $\mathcal{S}$  and it is updated after each next run. Thus, the output of  $\mathcal{S}$  on  $x$  and  $\theta$  is an answer for  $x$  (either a solution to  $x$ , or “no solution”, or “don’t know”) and the updated parameter string  $\theta'$ .

**Solver  $\mathcal{S}$ .** On input  $x$  and  $\theta$ , the solver  $\mathcal{S}$  works as follows:

1. Run  $\mathcal{A}_1$ . This algorithm produces a path

$$x_0, r_1, x_1, r_2, x_2, \dots, x_{n-1}, r_n, x_n \tag{1}$$

where  $x_0 = x$ . Note that  $x_n$  is not necessarily an easy instance. In the course of producing this path,  $\mathcal{A}_1$  generates other paths and measures the “quality” of the moves occurring in these paths: one move is better than another if it is expected to have a better chance of leading to an easy instance. Information about the

quality of the moves is stored as *quality data*  $\delta$ . The output of  $\mathcal{A}_1$  is path (1) and  $\delta$ .

2. Return an answer for  $x$ :

- (a) If  $\mathcal{E}(x_n)$  is “not easy”, then return “don’t know”.
- (b) If  $\mathcal{E}(x_n)$  is “no solution”, then return “no solution”.
- (c) If  $\mathcal{E}(x_n)$  is a solution to  $x_n$ , work backwards from  $x_n$  and use the solution functions  $g_r$  from  $\mathcal{R}$  to find successively solutions to  $x_{n-1}, \dots, x_0$ . Finally, return the solution to  $x_0$ .

3. Run  $\mathcal{A}_2$ . This algorithm takes  $\delta$  and merges it with similar quality data collected in the previous runs of  $\mathcal{S}$ . The result of merge is used for training and updating parameters  $\theta$  to new parameters  $\theta'$ .

4. Return the updated parameter string  $\theta'$  for storing.

**Reinforcement-learning algorithm  $\mathcal{A}_1$ .** This algorithm is a Monte Carlo tree search algorithm adapted for search problems. More exactly,  $\mathcal{A}_1$  is a version of the Adaptive Multistage Sampling algorithm (AMS) described in [3]. The algorithm  $\mathcal{A}_1$  cannot apply AMS as a black-box algorithm because the input to AMS is not given explicitly. Instead,  $\mathcal{A}_1$  supplies the input data in a “just-in-time” manner as follows.

- *Initialization of rewards.* In each recursive call, AMS initializes rewards of moves. Given an instance  $x$ , the *reward* of a move is a measure for the belief that this move is on a bounded-length path from  $x$  to an easy instance. The reward is maximum if the move is on such a path to an easy instance. The algorithm  $\mathcal{A}_1$  needs a belief estimation algorithm that computes initial reward values for moves. This estimation is implemented by a deep neural network  $N$  that uses parameters given in  $\theta$ . The initial rewards are improved by training this network.
- *Sampling algorithm.* The algorithm  $\mathcal{A}_1$  provides a *sampling algorithm* for AMS. On an instance  $x$ , this algorithm uses the parameters in  $\theta$  to sample the moves from  $f_r(x)$  for each self-reduction  $r \in \mathcal{R}$ . The sampling algorithm can be implemented using the same deep neural network  $N$ , or it can be a different neural network that shares weights with  $N$ . The distributions for self-reductions are improved by training  $N$ , which means that the improved distributions assign higher probabilities to moves with higher accumulated rewards.
- *Output.* According to the description of AMS in [3], this algorithm returns a path that has the maximum accumulated reward. In addition to this optimal path,  $\mathcal{A}_1$  collect the following quality data  $\delta$  and returns it for training:
  - for every instance  $x$  and every self-reduction  $r$  explored in the run, the accumulated probability distribution on  $f_r(x)$ ;
  - for every instance  $x$  explored in the run, the accumulated quality of  $x$  (“value” of  $x$  in the AMS terminology).

**Parameter-adjustment algorithm  $\mathcal{A}_2$ .** After taking the quality data  $\delta$  and merging it with similar datasets,  $\mathcal{A}_2$  trains the deep neural network  $N$  to adjust the parameters  $\theta$ . Note that the choice of architecture of  $N$  is dictated by instance representation. For

example, convolutional neural networks can be used in the case of finite dimensional tensor representation. If instances are represented by binary strings of variable length, recurrent neural networks can be used. In the case of SAT, it is natural to represent instances by graphs and, therefore,  $N$  can be implemented as a graph neural network. In particular an extension of the network constructed in [9] could be used. Also note that the architecture of  $N$  determines what instance features can be discovered from training.

What datasets can be used for the initial training of  $N$ ? It is more or less common to train a neural network using randomly shuffled data. Certain sets of instances (for example, industrial instances of SAT) expose self-similarity: large instances have the same properties as smaller ones. In such cases, it makes sense to train  $N$  using *curriculum learning* [7] where the training starts from samples of small size and moves to larger ones.

## 5 Concluding Remark

In this paper, we described how to adapt AlphaZero's techniques for designing a solver for a search problem. This adaptation can also be used for another task called *per-instance algorithm selection* [4, 5]. In this task, we are given a search problem  $\Pi$  and a "portfolio" of solvers for  $\Pi$ . We wish to design a "meta-solver" that automatically chooses a solver from the portfolio on a per-instance basis and, thereby, it achieves better performance than any single solver from the portfolio.

Suppose all solvers in the portfolio are of the following type. Such a solver takes as input an instance  $x$  of  $\Pi$  and produces another instance  $x'$  such that (1)  $x'$  has a solution if and only if  $x$  has a solution and (2) a solution to  $x$  can be computed from a solution to  $x'$ . If  $x'$  is an easy instance then the solver returns an answer, otherwise the solver returns  $x'$  and says "don't know". Many SAT solvers are of this type, for example, iterative solvers like resolution-based solvers with a limited number of iterations. The portfolio with such solvers can be viewed as a self-reduction where the move function maps an input instance  $x$  to the set of all instances  $x'$  produced by the solvers. Thus, we can use the solver described in Sect. 4 as a meta-solver for  $\Pi$ .

## References

1. Biere, A., Heule, M., van Maaren, H., Walsh, T. (Eds.): Handbook of Satisfiability, 2nd edn. IOS Press (2021)
2. Buss, S., Nordström, J.: Proof complexity and SAT solving. In: Handbook of Satisfiability, 2nd edn., vol. 336, pp. 233–350. IOS Press (2021)
3. Chang, H.S., Fu, M.C., Hu, J., Marcus, S.I.: An adaptive sampling algorithm for solving Markov decision processes. *Oper. Res.* **53**(1), 126–139 (2005)

4. Hoos, H.H., Hutter, F., Leyton-Brown, K.: Automated configuration and selection of SAT solvers. In: *Handbook of Satisfiability*. Volume 336 of *Frontiers in Artificial Intelligence and Applications*, 2nd edn., pp. 481–507. IOS Press (2021)
5. Kerschke, P., Hoos, H.H., Neumann, F., Trautmann, H.: Automated algorithm selection: survey and perspectives. *Evol. Comput.* **27**(1), 3–45 (2019)
6. Mazyavkina, N., Sviridov, S., Ivanov, S., Burnaev, E.: Reinforcement learning for combinatorial optimization: a survey. *Comput. Oper. Res.* **134**, 105400 (2021)
7. Narvekar, S., Peng, B., Leonetti, M., Sinapov, J., Taylor, M.E., Stone, P.: Curriculum learning for reinforcement learning domains: a framework and survey. *J. Mach. Learn. Res.* **21**, 181:1–181:50 (2020)
8. Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T.P., Silver, D.: Mastering atari, go, chess and shogi by planning with a learned model. *Nature* **588**, 604–609 (2020)
9. Selsam, D., Lamm, M., Bünz, B., Liang, P., de Moura, L., Dill, D.L.: Learning a SAT solver from single-bit supervision. In: *Proceedings of the 7th International Conference on Learning Representations, ICLR 2019* (2019)
10. Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., Hassabis, D.: A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* **362**(6419), 1140–1144 (2018)
11. Tseitin, G.S.: On the complexity of derivation in propositional calculus. *Zapiski Nauchnykh Seminarov LOMI* **8**, 234–259 (1968). In Russian. Reprinted in: Siekmann, J., Wrightson, G. (eds.): *Automation of Reasoning 2: Classical Papers on Computational Logic 1967–1970*, pp. 466–483. Springer (1983)
12. Vesselinova, N., Steinert, R., Perez-Ramirez, D.F., Boman, M.: Learning combinatorial optimization on graphs: a survey with applications to networking. *IEEE Access* **8**, 120388–120416 (2020)

# **Applications to Physics**



# Fuzzy Techniques, Laplace Indeterminacy Principle, and Maximum Entropy Approach Explain Lindy Effect and Help Avoid Meaningless Infinities in Physics



Julio Urenda, Sean Aguilar, Olga Kosheleva, and Vladik Kreinovich

**Abstract** In many real-life situations, the only information that we have about some quantity  $S$  is a lower bound  $T < S$ . In such a situation, what is a reasonable estimate for  $S$ ? For example, we know that a company has survived for  $T$  years, and based on this information, we want to predict for how long it will continue surviving. At first glance, this is a type of a problem to which we can apply the usual fuzzy methodology—but unfortunately, a straightforward use of this methodology leads to a counter-intuitive infinite estimate for  $S$ . There is an empirical formula for such estimation—known as Lindy Effect and first proposed by Benoit Mandelbrot—according to which the appropriate estimate for  $S$  is proportional to  $T$ :  $S = C \cdot T$ , where, with some confidence, the constant  $C$  is equal to 2. In this paper, we show that a deeper analysis of the situation enables fuzzy methodology to lead to a finite estimate for  $S$ , moreover, to an estimate which is in perfect accordance with the empirical Lindy Effect. Interestingly, a similar idea can help in physics, where also, in some problems, straightforward computations lead to physically meaningless infinite values.

---

J. Urenda

Departments of Mathematical Science and Computer Science, University of Texas at El Paso, 500 W. University El Paso, El Paso, TX 79968, USA

e-mail: [jcurenda@utep.edu](mailto:jcurenda@utep.edu)

S. Aguilar · V. Kreinovich (✉)

Department of Computer Science, University of Texas at El Paso, 500 W. University El Paso, El Paso, TX 79968, USA

e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

S. Aguilar

e-mail: [raguilar4@miners.utep.edu](mailto:raguilar4@miners.utep.edu)

O. Kosheleva

Department of Teacher Education, University of Texas at El Paso, 500 W. University El Paso, El Paso, TX 79968, USA

e-mail: [olgak@utep.edu](mailto:olgak@utep.edu)

## 1 Formulation of the Problem

**What is Lindy Effect.** In this paper, we analyze a phenomenon known as *Lindy Effect*; see, e.g., [9, 14]. Its main idea—as described later—is intuitively clear, but its formal description is not that well known, so let us start by describing what is the Lindy Effect.

**Lindy Effect: intuitive idea.** If we have a company that has been in successful existence for many decades and another company which is a recent startup, what are the chances that each of these companies will survive for another decade? Intuitively, it is clear that the company that has been successful for many years, that have successfully survived many crises, will probably survive for another decade (and probably even longer), while a start-up has a high risk of not surviving—as most startups do.

This is an important issue if we plan a long-term investment: the stocks of which of the two companies shall we mostly buy?

If we have a building that has been standing since the 19 century, and another modernist experimental building built a few years ago, which of them has a better chance of survival? Clearly, the one that has been standing for more than 100 years will probably stand some more, while an experimental building, built by using not-yet-fully-tested technology, is at risk of needing repairs soon.

If we have a family that has recently celebrated its 50th anniversary and another family whose marriage has just been announced—who has a bigger chance of not divorcing?

In all these cases, it is quite possible that an old company will crumble while a startup will turn into a new Microsoft, that an old building will catch fire and collapse while the new one will persist, that the old couple will divorce after 50 years of marriage while the newlyweds will live happily even after—but in all these cases, the opposite is much more frequent.

**Why is this called Lindy Effect?** This name came from New York’s *Lindy’s Delicatessen*, which in the 1960s was a favorite gathering place for New York comedians—and in those days, this meant the majority of top US comedians. Once in a while, a new comedian would burst into the stage, so a natural question was: will he (it was usually a he) last for long? Young people may have believed in every single newcomer’s success, but more experience folks—who remembered that many new promising comedians did not last long—would cool down the younger folks’ optimism.

**Lindy Effect: towards formalization.** In all the above situations—and in many similar ones:

- we know that some object has already survived for  $T$  years, and
- we are trying to predict the amount of time  $t$  during which it will most probably survive in the future as well.

Alternatively, we can say that we want to predict the overall survival time  $S \stackrel{\text{def}}{=} T + t$ .

If the value  $T$  is all we know, then we need to estimate the future value  $t$  based only on this information. Let us denote the corresponding estimate by  $t = f(T)$ . Which function  $f(T)$  should we use for this estimation?

**Lindy Effect: qualitative idea.** The above informal discussion enables us to conclude that the larger survival-so-far time  $T$ , the largest should be our estimate  $t = f(T)$ . In other words, the desired estimation function  $f(T)$  should be increasing.

However, there are many increasing functions. Which one should we choose?

**Lindy Effect: precise formulation(s).** The first person who tried to come up with a precise formula for the Lindy Effect was Benoit Mandelbrot—the father of fractals. By considering several actual situations, he concluded that the desired dependence is linear: there exists a constant  $c > 0$  such that if a system survived for  $T$  years, it will, with high probability, survive for another  $t = c \cdot T$  years; see, e.g., [9].

Later, Nassim Nicholas Taleb analyzed even more cases and concluded that we can safely take  $c = 1$  and  $t = T$ ; see, e.g., [14]. In plain English, this means that if a company survived for 100 years, it is reasonable to expect that it will survive for another 100 years.

**Weak and Strong Lindy Effect.** We have two versions of Lindy Effect:

- The first version—that  $t = c \cdot T$  for some  $c > 0$ —is somewhat more accurate, since we have a parameter here that we can adjust to make a better fit.
- The second version—that  $t = T$ —is somewhat less accurate but stronger.

To distinguish between these two formulations, we will call the dependence  $t = c \cdot T$  a *weak* Lindy Effect and the dependence  $t = T$  the *strong* Lindy effect.

**Why?** Both formulations seem to be consistent with data, so they are real. The fact that they are ubiquitous, that they cover all kinds of phenomena, seems to indicate that there must be a general first-principles explanation for this effect.

**What we do in this paper.** In this paper, we will try to come up with this explanation.

All this is very imprecise (“fuzzy”), so a natural idea is to try to use fuzzy techniques; see, e.g., [2, 6, 10–12, 16]. On the complications side, we will see that in the process of these tries, we will encounter a need to somewhat modify the way such problems are usually described by fuzzy techniques.

The resulting complications will not be fully in vain: they will enable us to come up with a natural way to avoid meaningless infinities in computations related to physics.

## 2 Let Us Use Fuzzy Techniques: A Straightforward Approach and Why It Does Not Work in This Case

**Starightforward approach: idea.** At first glance, we have a typical problem of the type solved by fuzzy techniques—e.g., in fuzzy control. We have rules which are imprecise—in the sense that by themselves, they do not lead to an exact answer.

- In the control case, we may have rules like “if  $x$  is small, then control should be small”—which allow many different control values (as long as they are small).
- In our case, all we know is that the overall survival time  $S$  should be larger than the survival-so-far time  $T$ . This also allows many different values  $S$ —as long as they are larger than  $T$ .

In fuzzy control, the fuzzy methodology means that:

- we describe the knowledge in terms of fuzzy degrees,
- we come up with a fuzzy recommendation, and then
- we apply an appropriate defuzzification procedure to come up with the numerical recommendation.

Let us try to apply the same idea to our problem.

**Straightforward approach: let us try.** If all we know is the value  $T$ , and the only thing that we know about the desired value  $S$  is that  $S > T$ , then the corresponding membership function  $\mu(S)$  describing this knowledge is straightforward:

- it assigns  $\mu(S) = 1$  to all the values  $S > T$ , and
- it assigns  $\mu(S) = 0$  to all other values.

So far so good, but the problem starts when we try to apply defuzzification.

The most natural idea is to select the value in which we have most confidence, i.e., for which the corresponding value of the membership function is the largest. In our case, this does help at all: the largest value  $\mu(S) = 1$  is attained for *all* numbers  $S > T$ , so this idea does not allow us to select any specific value at all.

OK, this happens in fuzzy control as well. To avoid this non-uniqueness, fuzzy control applications usually use *centroid defuzzification*, i.e., transform a membership function  $\mu(x)$  into a value

$$\bar{x} = \frac{\int x \cdot \mu(x) dx}{\int \mu(x) dx}.$$

Of course, we cannot directly apply this formula to our membership function  $\mu(S)$ , since for this function, both integrals—in the numerator and in the denominator—are infinite. However, what we *can* do is to consider our function  $\mu(S)$  as the limit of functions  $\mu_n(S)$  which coincide with  $\mu(S)$  up to  $S = T + n$  and are equal to 0 after that. In the limit  $n \rightarrow \infty$ , the functions  $\mu_n(S)$  tend to the desired function  $\mu(S)$ . So, it makes sense:

- first, to apply defuzzification to each of these functions  $\mu_n(S)$ , resulting in values  $\bar{S}_n$ , and
- then use the limit  $\bar{S} = \lim_{n \rightarrow \infty} \bar{S}_n$  of the resulting values  $\bar{S}_n$  as the desired estimate for  $S$ .

Unfortunately, this does not work either: for each function  $\mu_n(S)$ , centroid defuzzification leads to  $\bar{S}_n = T + (n/2)$ , and thus, the limit  $\bar{S} = \lim_{n \rightarrow \infty} \bar{S}_n = \infty$ . Mathematically, it is correct, but it does not convey the meaning that we want: instead of

saying that a company will survive for  $c \cdot T$  more years, this conclusion says that the company will last forever.

We know that this is not true: many companies do survive for a long time, but most of them eventually stop functioning. There are not that many companies that have survived for many centuries: maybe Lloyd insurance is the only one.

### 3 Let Us Add Common Sense to Mathematics

**So what can we do?** At first glance, the above negative results may sound like a paradox that shows limitations of the fuzzy approach. But a deeper analysis shows that nothing is wrong with fuzzy approach, it is that we relied too much on mathematics and did not use enough common sense.

Specifically, we naively assumed that  $\mu(S) = 1$  for all  $S > T$ . Mathematically, it makes sense, but do we really believe—with confidence 1—that a company that survived for 100 years will survive for 1000 years more? If you believe this, how about 1 million years? 1 billion years? Clearly not.

From the viewpoint of common sense, the value of the membership function  $\mu(S)$  describing a seemingly crisp property  $S > T$  should not stay constant, but should instead decrease as  $S$  increases.

**What is an adequate membership function: analysis of the problem.** We are interested in designing, for each  $T$ , a membership function  $\mu_T(S)$  that describes our degree of belief that, once the system has survived for time  $T$ , it will survive for a longer time  $S \geq T$ .

What should be reasonable properties of these functions?

First, we know for sure that the system has survived for time  $T$ , so we should have  $\mu_T(T) = 1$ .

Second, the longer the time  $S$ , the smaller is our belief that the system will survive for this time. Thus, for each  $T$ , the function  $\mu_T(S)$  should be decreasing. We will call this property *monotonicity*.

Third, if we originally observed the system surviving for time  $T$ , and then later, it turns out that it has survived for time  $T' > T$ , this means that from the original function  $\mu_T(S)$ , we should only consider values  $S \geq T'$ . Of course, since the function  $\mu_T(S)$  is decreasing, the largest remaining value is the value  $\mu_T(T')$  which is smaller than  $\mu_T(T) = 1$ . In fuzzy techniques, we usually consider *normalized* membership functions, i.e., functions whose maximum is 1. So, to obtain the appropriate function  $\mu_{T'}(S)$ , we need to normalize the resulting restriction of the original function  $\mu_T(S)$  to values  $S \geq T'$ . Normalization is usually performed by dividing all the membership degrees by the largest one—which, is in this case, is equal to  $\mu_T(T')$ . Thus, we must have

$$\mu_{T'}(S) = \frac{\mu_T(S)}{\mu_T(T')}$$

for all  $S \geq T'$ . We will call this property *consistency*.

Finally, since we are trying to understand the phenomenon of Lindy Effect, which is reasonably universal, we want the expressions  $\mu_T(S)$  to be *universal*. In particular, it means that this effect should be the same whether we consider micro-objects or macro-objects or mega-objects (how long will the Sun continue to shine?). The corresponding membership degrees should thus not change if we simply change the units in which we measure time. If we replace the original unit of time with the one which is  $\lambda$  times smaller, then numerical values of both  $T$  and  $S$  are multiplied by  $\lambda$ : we get  $\lambda \cdot T$  instead of  $T$  and  $\lambda \cdot S$  instead of  $S$ . In these terms, universality means that  $\mu_{\lambda \cdot T}(\lambda \cdot S) = \mu_T(S)$ .

**Definitions and the main result.** Now, we are ready to formulate our first result.

**Definition 1** By a *family of membership functions corresponding to  $>$* , we mean a family of membership functions  $\mu_T(S)$  with parameter  $T > 0$  each of which is defined for all  $S \geq T$  and which satisfy the following properties:

- for each  $T$ , we have  $\mu_T(T) = 1$ ;
- for each  $T$ , the function  $\mu_T(S)$  is decreasing with  $S$  (*monotonicity*);
- for each  $T < T' \leq S$ , we have

$$\mu_{T'}(S) = \frac{\mu_T(S)}{\mu_T(T')}; \text{ (consistency), and}$$

- for each  $T \leq S$  and for each  $\lambda > 0$ , we have

$$\mu_{\lambda \cdot T}(\lambda \cdot S) = \mu_T(S) \text{ (universality).}$$

**Proposition 1** *Every family of membership functions corresponding to  $>$  has the form  $\mu_T(S) = \left(\frac{T}{S}\right)^\alpha$  for some  $\alpha > 0$ .*

**Proof** For  $T = 1$  and  $\lambda \geq 1$ , universality implies that

$$\mu_\lambda(\lambda \cdot S) = \mu_1(S).$$

On the other hand, due to consistency, with  $T = 1 < T' = \lambda$ , we have

$$\mu_\lambda(\lambda \cdot S) = \frac{\mu_1(\lambda \cdot S)}{\mu_1(\lambda)}.$$

Equating the resulting two expressions for the same value  $\mu_\lambda(\lambda \cdot S)$ , we conclude that

$$\mu_1(S) = \frac{\mu_1(\lambda \cdot S)}{\mu_1(\lambda)},$$

i.e., equivalently,

$$\mu_1(\lambda \cdot S) = \mu_1(\lambda) \cdot \mu_1(S) \tag{1}$$

In particular, for  $S = \lambda^{-1}$ , we get

$$1 = \mu_1(1) = \mu_1(\lambda) \cdot \mu_1(\lambda^{-1}),$$

hence

$$\mu_1(\lambda^{-1}) = \frac{1}{\mu_1(\lambda)}. \tag{2}$$

For each  $\lambda < 1$  and  $S$ , and for  $S' = \lambda \cdot S$  and  $\lambda' = 1/\lambda > 1$ , the formula (1) leads to  $\mu_1(\lambda' \cdot S') = \mu_1(\lambda') \cdot \mu_1(S')$ , i.e.,  $\mu_1(S) = \mu_1(1/\lambda) \cdot \mu_1(\lambda \cdot S)$ , and thus, due to (2), to the formula (1).

For  $\lambda = 1$ , the property (1) is trivially true. Thus, the property (1) is satisfied for all  $\lambda > 0$  and for all  $S$ .

Functions that satisfy this property are known as *multiplicative*, and it is known that every monotonic multiplicative function has the form  $\mu_1(x) = x^{-\alpha}$  for some real value  $\alpha$ ; see, e.g., [1]. Since all membership functions  $\mu_T(S)$  are decreasing, we must have  $\alpha > 0$ .

For each  $T \leq S$ , we can then use the universality property with  $\lambda = T^{-1}$  and get  $\mu_T(S) = \mu_1(S/T)$ , thus  $\mu_T(S) = (S/T)^{-\alpha}$ . The proposition is proven.

**This explains (weak) Lindy Effect.** To make sure that for the membership function  $\mu_T(S) = \left(\frac{T}{S}\right)^\alpha$ , both numerator and denominator integrals in the formula for centroid defuzzification are finite, we must have  $\alpha > 2$ . In this case,

$$\int_T^\infty S \cdot \left(\frac{T}{S}\right)^\alpha dS = T^\alpha \cdot \frac{1}{\alpha - 2} \cdot T^{2-\alpha} = \frac{1}{\alpha - 2} \cdot T^2$$

and

$$\int_T^\infty \left(\frac{T}{S}\right)^\alpha dS = T^\alpha \cdot \frac{1}{\alpha - 1} \cdot T^{1-\alpha} = \frac{1}{\alpha - 1} \cdot T,$$

thus

$$\bar{S} = \frac{\int S \cdot \mu_T(S) dS}{\int \mu_T(S) dS} = \frac{\alpha - 1}{\alpha - 2} \cdot T.$$

Thus, the remaining time  $t = S - T$  is indeed proportional to  $T$ , which is exactly what we called weak Lindy Effect.

## 4 What About Probabilistic Case

**Probabilistic case: (almost) the same result.** In the previous section, we considered the case when we use fuzzy logic to describe the corresponding uncertainty. What if instead we use probabilities?

In this case, for each  $T$ , we have the probability  $p_T(S)$  that the system will survive for time  $S$  once it has survived for time  $T$ . The same arguments as in the fuzzy case show that this function:

- should also satisfy the condition  $p_T(T) = 1$ ,
- should also be decreasing as  $S$  increases, and
- should also not depend on the choice of the measuring unit, i.e., we should have  $p_{\lambda \cdot T}(\lambda \cdot S) = p_T(S)$  for all  $T \leq S$  and  $\lambda > 0$ .

And if we have already observed the system for time  $T' > T$  and the system survived during this time, then the new probabilities  $p_{T'}(S)$  should be computed by using the formulas for conditional probability:  $p_{T'}(S) = \frac{p_T(S)}{p_T(T')}$ .

Thus, the new functions should satisfy the same conditions as described in Definition 1, and thus, by Proposition 1, it should have the same form  $p_T(S) = \left(\frac{T}{S}\right)^\alpha$  for some  $\alpha > 0$ .

In the probabilistic case, a natural numerical estimate is the mean value  $\bar{S} = \int S \cdot \rho_T(S) dS$ , where the probability density function  $\rho_T(S)$  can be obtained by differentiating the function  $p_T(S)$ —which is, in effect, equal to 1 minus the cumulative distribution function; see, e.g., [13]. In this case, we get  $\bar{S} = \frac{\alpha - 1}{\alpha}$ . So, in this case, we also get the weak Lindy Effect.

### Why do fuzzy and probabilistic approaches lead, in effect, to the same formula?

The fact that by using such different techniques as fuzzy and probabilistic, we get the exact same result—that the expected remaining survival time  $t$  is proportional to the survival-so-far time  $T$ —is a good indication that there is an even more fundamental reason behind this dependence, reason not depending on which technique we use to describe uncertainty.

And indeed, such a reason is easy to describe: the reason is what we called *universality*, that the result should not depend on the choice of the measuring unit. Our original problem was to find the estimate  $t = f(T)$ . In terms of the estimating function  $f(x)$ , universality means that if we have  $t = f(T)$  in the original units, then the same relation  $t' = f(T')$  should hold if we describe the times in the new units, i.e., if we take  $t' = \lambda \cdot t$  and  $T' = \lambda \cdot T$ .

**Formulating the problem in precise terms.** Let us describe this requirement in precise terms.



**Definition 2** We say that the function  $t = f(T)$  is *universal* if for all  $t, T$ , and  $\lambda > 0$ , the equality  $t = f(T)$  implies that  $t' = f(T')$ , where  $t' = \lambda \cdot t$  and  $T' = \lambda \cdot T$ .

**Proposition 2** Every universal function has the form  $f(T) = c \cdot T$  for some constant  $c$ .

**Proof** Let us denote  $f(1)$  by  $c$ , so that  $c = f(1)$ . Then, for each  $T$ , if we take  $\lambda = T$ , then universality enables us to imply that  $T \cdot c = f(T \cdot 1)$ , i.e., that indeed  $f(T) = c \cdot T$ . The proposition is proven.

**Discussion.** So, indeed, universality implies the weak Lindy Effect.

## 5 Why Strong Lindy Effect

**Reminder.** In the above text, we explained the *weak* Lindy Effect, according to which the remaining survival time  $t$  is related to the survival-so-far time  $T$  by the formula  $t = c \cdot T$ , for some constant  $t$ . However, as we have mentioned, there is strong evidence that this constant  $c$  is equal to 1, i.e., that we have what we called the *strong* Lindy Effect  $t = T$ .

How can we explain this?

**A simplified (somewhat naive) explanation.** A simplified explanation comes from Laplace Indeterminacy Principle (see, e.g., [5]), according to which if we have no reason to believe that two quantities are different, it makes sense to assume that they are equal.

From this viewpoint, since we do not have any reason to believe that the remaining survival time  $t$  is smaller or larger than the survival-so-far time  $T$ , so it makes sense to take  $t = T$ .

**A better explanation: fuzzy case.** In our problem, we know the value  $T$ , and know that  $T < S$ . In this case, as we have mentioned earlier, the straightforward fuzzy approach does not lead to any meaningful estimate for  $S$ .

But what if we *reverse* the problem: what is we assume that  $S$  is known, and the only information that we have about  $T$  is that  $0 \leq T \leq S$ . In this case, the corresponding (crisp) knowledge leads to the following membership function:  $\mu_S(T) = 1$  when  $0 \leq T \leq S$  and  $\mu_S(T) = 0$  otherwise. For this membership function, centroid defuzzification leads to  $\bar{T} = S/2$ .

So, if we know  $S$ , then we should take  $T = S/2$ . It is therefore natural to conclude that if we know  $T$ , then we should take  $S$  for which  $T = S/2$ . For this  $S$ , we have  $S = 2T$ , so the remaining survival time is  $t = S - T = T$ , which is exactly the strong Lindy Effect.

**Probabilistic case.** We can apply the same reversal idea to the case of probabilistic uncertainty.

Suppose that we know the value  $S$ , and the only information that we have about  $T$  is that  $T$  is between 0 and  $S$ . In this case, the maximum likelihood approach—a natural formalization of the Laplace Indeterminacy Principle—implies that the corresponding probability distribution on the interval  $[0, S]$  is uniform [5]. For this uniform distribution, the mean value is  $\bar{T} = S/2$ , which also prompts us to use the estimate  $S = 2T$  and thus,  $t = T$ .

*Comment.* In [4, 8], a similar idea was used to explain why in engineering, after we get an estimate of uncertainty based on known factors, practitioners usually double this estimate to take into account possible unknown factors as well.

This leads, e.g., to doubling the safety margins computed based only on the known factors.

## 6 Application to Physics: How to Avoid Physically Meaningless Infinite Values

**Problem: reminder.** It is known that in physics, some computations lead to meaningless infinite values. The simplest example of such a phenomenon is computing the overall energy of an electron's electric field; see, e.g., [3, 15] for detail.

An electron is an elementary particle, which means that it has no independent parts. According to special relativity, all velocities are bounded by the speed of light. Thus, if the electron was not point-wise, if it had at least two spatially separated points, then it would take some time for these points to influence each other—and therefore, during this time, these two points would act independently. So, an electron has to be a point-wise particle.

For a point-wise particle, the value of its electric field  $\mathbf{E}$  at any point  $x$  is determined by the Coulomb Law, as proportional to the  $r^{-2}$ , where  $r$  is the distance between this point and the location of the electron.

It is known that the energy density  $\rho(x)$  is proportional to the square of the electric field, i.e., to  $r^{-4}$ . The overall energy  $E$  can be computed by integrating this density over the whole 3-D space:  $E = \int \rho(x) dx$ . The problem is the resulting integral is infinite:

$$E = \int r^{-4} dx = \int_0^{\infty} r^{-4} \cdot 4\pi \cdot r^2 dr = 4\pi \cdot \int_0^{\infty} r^{-1} dr = 4\pi \cdot r^{-1} \Big|_0^{\infty} = \infty.$$

So, we get a physically meaningless value for a physically meaningful quantity—the overall energy of the electron's electric field.

How can we make the corresponding estimate physically meaningful—i.e., finite?

*Comment.* There are many such infinities in classical physics—the existence of such infinities was one of the main reasons why quantum physics was discovered in the first place. However, in contrast to many other cases when the answer become finite in the quantum case, for the overall energy of the electron’s electric field remains infinite in the quantum cases as well.

**Known idea.** A previously proposed possible way to solve this problem is to take into account that measurements are always imprecise, that at any given moment of time, there is a limit on how accurately we can measure, e.g., the distance—and probably there is a fundamental limit; see, e.g., [7].

So, instead of the actual distance  $r$ , we can only conclude that the actual distance is between  $r - \varepsilon$  (to be more precise,  $\max(0, r - \varepsilon)$ , since the distance cannot be negative) and  $r + \varepsilon$  for some  $\varepsilon$ . Thus, the value of the electric field at any point  $x$  is somewhere between  $(r + \varepsilon)^{-2}$  and  $(\max(0, r - \varepsilon))^{-2}$ , and, correspondingly, the overall energy is between

$$\underline{E} = \int (r + \varepsilon)^{-4} dx \text{ and } \overline{E} = \int (\max(0, r - \varepsilon))^{-4} dx.$$

One can check that the first integral  $\underline{E}$  is finite—for small  $r$ , the integrated function  $(r + \varepsilon)^{-4}$  is bounded from above by the value  $\varepsilon^{-4}$ . However, the second integral is clearly infinite—since for  $r \leq \varepsilon$ , we have  $\max(0, r - \varepsilon) = 0$  and thus,

$$(\max(0, r - \varepsilon))^{-4} = \infty.$$

So, instead of the infinite value for the total energy  $E$  of the electron’s electric field, we have a semi-infinite interval of possible values  $[E, \infty)$ . In other words, the only information that we have about the overall energy is that it is larger than or equal to  $\underline{E}$ .

**Lindy Effect helps.** The situation when the only information that have about an unknown quantity  $S$  is that it is larger than or equal to some known quantity  $T$  is exactly the situation described by the Lindy Effect.

According to the Lindy Effect—which we explained in this paper—in such a situation, the appropriate estimate for the unknown value  $E$  is a finite estimate  $E = c \cdot \underline{E}$  (where it is highly probable that  $c = 1$ ).

So, we have a finite estimate for the overall energy—thus avoiding the meaningless infinity.

**Acknowledgements** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478. The authors are greatly thankful to the anonymous referees for valuable suggestions.

## References

1. Aczél, J., Dhombres, J.: *Functional Equations in Several Variables*. Cambridge University Press, Cambridge (2008)
2. Belohlavek, R., Dauben, J.W., Klir, G.J.: *Fuzzy Logic and Mathematics: A Historical Perspective*. Oxford University Press, New York (2017)
3. Feynman, R., Leighton, R., Sands, M.: *The Feynman Lectures on Physics*. Addison Wesley, Boston (2005)
4. Gholamy, A., Kreinovich, V.: Safety factors in soil and pavement engineering: theoretical explanation of empirical data. In: *Abstracts of the 23rd Joint UTEP/NMSU Workshop on Mathematics, Computer Science, and Computational Sciences*, El Paso, Texas, November 3, 2018
5. Jaynes, E.T., Bretthorst, G.L.: *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, UK (2003)
6. Klir, G., Yuan, B.: *Fuzzy Sets and Fuzzy Logic*. Prentice Hall, Upper Saddle River (1995)
7. Kosheleva, O., Kreinovich, V.: Interval (set) uncertainty as a possible way to avoid infinities in physical theories. In: *Abstracts of the 18th International Symposium on Scientific Computing, Computer Arithmetic, and Verified Numerical Computation SCAN'2018*, Tokyo, Japan, September 10–15, 2018
8. Lorkowski, J., Kreinovich, V.: How to gauge unknown unknowns: a possible theoretical explanation of the usual safety factor of 2. *Math. Struct. Model.* **32**, 49–52 (2014)
9. Mandelbrot, B.B.: *The Fractal Geometry of Nature*. Henry Holt & Co., New York (1983)
10. Mendel, J.M.: *Uncertain Rule-Based Fuzzy Systems: Introduction and New Directions*. Springer, Cham, Switzerland (2017)
11. Nguyen, H.T., Walker, C.L., Walker, E.A.: *A First Course in Fuzzy Logic*. Chapman and Hall/CRC, Boca Raton (2019)
12. Novák, V., Perfilieva, I., Močkoř, J.: *Mathematical Principles of Fuzzy Logic*. Kluwer, Boston, Dordrecht (1999)
13. Sheskin, D.J.: *Handbook of Parametric and Non-parametric Statistical Procedures*. Chapman & Hall/CRC, London, UK (2011)
14. Taleb, N.N.: *Antifragile: Things That Gain from Disorder*. Random House, New York (2012)
15. Thorne, K.S., Blandford, R.D.: *Modern Classical Physics: Optics, Fluids, Plasmas, Elasticity, Relativity, and Statistical Physics*. Princeton University Press, Princeton (2017)
16. Zadeh, L.A.: Fuzzy sets. *Inf. Control.* **8**, 338–353 (1965)

# Dimension Compactification Naturally Follows from First Principles



Julio C. Urenda, Olga Kosheleva, and Vladik Kreinovich

**Abstract** According to modern physics, space-time originally was of dimension 11 or higher, but then additional dimensions became compactified, i.e., size in these directions remains small and thus, not observable. As a result, at present, we only observed 4 dimensions of space-time. There are mechanisms that explain *how* compactification may have occurred, but the remaining question is *why* it occurred. In this paper, we provide two first-principles-based explanations for space-time compactification: based on Second Law of Thermodynamics and based on geometry and symmetries.

## 1 Formulation of the Problem

**What is dimension compactification.** According to modern physics (see, e.g., [5, 9, 11]), the requirement that the quantum field theory be consistent implies that the dimension of space-time should be at least 11. How can we combine this conclusion with the fact that the observed space-time is only 4-dimensional?

A usual explanation is that while in the beginning, space-time may have had 11 or more equally prominent dimensions, with time, most of these dimensions has been *compactified*: i.e., the size in the direction of these additional dimension remains as

---

J. C. Urenda

Department of Mathematical Sciences and Department of Computer Science, University of Texas at El Paso, 500 W. University El Paso, El Paso, TX 79968, USA

e-mail: [jcurenda@utep.edu](mailto:jcurenda@utep.edu)

O. Kosheleva

Department of Teacher Education, University of Texas at El Paso, 500 W. University El Paso, El Paso, TX 79968, USA

e-mail: [olgak@utep.edu](mailto:olgak@utep.edu)

V. Kreinovich (✉)

Department of Computer Science, University of Texas at El Paso, 500 W. University El Paso, El Paso, TX 79968, USA

e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

153

M. Ceberio and V. Kreinovich (eds.), *Decision Making Under Uncertainty*

*and Constraints*, Studies in Systems, Decision and Control 217,

[https://doi.org/10.1007/978-3-031-16415-6\\_22](https://doi.org/10.1007/978-3-031-16415-6_22)

small as the Universe was in its first moments, while other dimensions expanded to the current astronomical sizes.

**Compactification: how and why.** There are several mechanisms that explain *how* compactification could have happened. However, these mechanisms do not explain *why* it happened.

In this paper, we provide arguments that compactification naturally follows from first principle. We actually provide *two* first-principles explanations for space-time compactification:

- an explanation based on the Second Law of Thermodynamics and
- an explanation based on geometry and symmetries.

## 2 Explanation Based on the Second Law of Thermodynamics

**Second Law of Thermodynamics: a brief reminder.** According to the Second Law of Thermodynamics (see, e.g., [2, 11]), the entropy of the Universe (and of any closed system) increases with time (or, in some cases, stays the same)—and there is no limit to such increase, eventually we get closer and closer to the state with the largest possible entropy.

**What is entropy: a brief reminder.** In general, the entropy is defined as [6, 10]:

$$S = - \int \rho(x) \cdot \ln(\rho(x)) dx,$$

where  $\rho(x)$  is the probability distribution of the set of all possible micro-states.

**How is entropy depending on dimension.** In general:

- close points or close particles are strongly correlated, while
- distant particles are independent.

A simplified description of this phenomenon can be obtained if we assume that all the points are divided into groups of nearby ones, so that:

- within each group there is a correlation, but
- between the groups there is no correlation.

It is known (see, e.g., [6]) that if we have several independent random processes, then the overall entropy is equal to the sum of the entropies of these processes. Thus, to find the overall entropy of the Universe in this approximation, it is sufficient to compute the entropy corresponding to each group, and then add up the resulting entropies.

How many points  $n$  are in each such group? Let us consider first the case when we only consider immediate neighbors—i.e., points whose all coordinates different from this one by no more than 1 appropriate unit of distance. In a coordinate system in which a central particle is at the point  $(0, \dots, 0)$ , each of  $d$  coordinates of an immediate neighbor is equal to  $-1$ ,  $0$ , or  $1$ —three options. So overall, we have  $n = 3^d$  points. If we consider neighbors of neighbors, we can have  $5^d$  points—and, in general,  $n = a^d$  for some  $a > 1$ .

This number clearly grows with the dimension  $d$ . So, when we go from a higher dimension  $d$  to a lower dimension  $d' < d$ , the number of neighbors decreases. This means that:

- instead of the original group of size  $n$  in which all particles were correlated,
- we have several subgroups of smaller size, and there is no longer correlation between different subgroups.

It is known—see, e.g., [6]—that if we know distributions corresponding to all the subgroups, then the entropy of the overall distribution for the whole group is the largest if and only if these subgroups are independent. Thus, when we divide a group in which all elements were correlated into smaller independent subgroups, we increase entropy.

Since, according to the usual interpretation of the Second Law of Thermodynamics, there are no limitations to the increase in entropy, eventually, we should also encounter a decrease in spatial dimension as a way to increase entropy—and this is exactly what compactification is about.

*Comment.* The above argument does not imply that compactification will stop at our 3 dimensions: it can go further, to having a 2- and even 1-dimensional space. Maybe this is what is already happening in the Universe, with 1D superclusters of Galaxies; see, e.g., [1, 7].

### 3 Explanation Based on Geometry and Symmetries

**Our second explanation is based on a natural physical process.** The original distribution of matter was uniform. However, the uniform distribution is not stable:

- if at some point, due to fluctuations, the density becomes larger than at the neighboring points,
- then this point start attracting matter from its neighbors—thus further increasing its density.

As a result, you get a large disturbance.

**Symmetries and statistical physics: general idea.** The original distribution in a  $d$ -dimensional space was invariant under shifts, rotations, and scaling (i.e., transformation  $x_i \rightarrow \lambda \cdot x_i$ ).

According to statistical physics (see, e.g., [2, 11]):

- It is not very probably that from a highly symmetric state, we go straight into a completely asymmetric one.
- Usually, the most probably transition is to a state that preserves as many symmetries as possible.

So, we expect the shapes of the disturbances to have some symmetries.

**Analysis of the problem.** What is the shape that has the largest number of symmetries—i.e., for which the dimension of the corresponding symmetry group is the largest?

If the shape is invariant with respect to all rotations in the  $d$ -dimensional space, then it must consist of spheres, and a sphere has only rotations—so the dimension of the corresponding symmetry group is  $\frac{d \cdot (d - 1)}{2}$ . Indeed, infinitesimal rotations are described by asymmetric matrices which have exactly as many parameters. So, in this case, the dimension of the symmetry group is

$$\frac{d^2 - d}{2}.$$

If the shape includes a  $(d - 1)$ -dimensional space, then we have  $d - 1$  independent shifts,  $\frac{(d - 1) \cdot (d - 2)}{2}$  independent rotations, and 1 scaling, to the total of

$$d - 1 + \frac{(d - 1) \cdot (d - 2)}{2} + 1 = \frac{d^2 - d + 2}{2},$$

which is larger than for the sphere.

If we have all  $(d - 1)$ -dimensional rotations but not all shifts or scaling, then we have fewer symmetries.

What if we only have rotations in a  $(d - 2)$ -dimensional space, to the total of  $\frac{(d - 2) \cdot (d - 3)}{2}$ ? We cannot have  $d - 1$  shifts, because this would lead to a  $(d - 1)$ -dimensional space. Thus, we can have no more than  $d - 2$  independent shifts. Even if we have  $d - 2$  shifts and rotations, we will have

$$d - 2 + \frac{(d - 2) \cdot (d - 3)}{2} + 1 < d - 1 + \frac{(d - 1) \cdot (d - 2)}{2} + 1$$

independent symmetries.

**Conclusion.** The most probable result of a natural spontaneous symmetry violation of a  $d$ -dimensional space is a  $(d - 1)$ -dimensional space. Since fluctuations continue, we will then get space of dimension  $d - 2$ , etc.

This provides another explanation of why the original space has lost many of its dimensions.



*Comments.*

- We have two explanations of the same phenomenon, but these explanations are not contradicting each other—both are based on statistical physics, we just took into account different aspects of it.
- The above idea of shapes motivated by symmetries has been used in physics—e.g., it explains the existing shapes of celestial bodies; see, e.g., [3, 4, 8].

**Acknowledgements** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes). It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

## References

1. Fairall, A.: Large-Scale Structures in the Universe. Wiley, New York (1998)
2. Feynman, R., Leighton, R., Sands, M.: The Feynman Lectures on Physics. Addison Wesley, Boston (2005)
3. Finkelstein, A., Kosheleva, O., Kreinovich, V.: Astrogeometry: towards mathematical foundations. *Int. J. Theor. Phys.* **36**(4), 1009–1020 (1997)
4. Finkelstein, A., Kosheleva, O., Kreinovich, V.: Astrogeometry: geometry explains shapes of celestial bodies. *Geoinformatics* **VI**(4), 125–139 (1997)
5. Green, M.B., Schwarz, J.H., Witten, E.: Superstring Theory, vols. 1, 2. Cambridge University Press, Cambridge (1988)
6. Jaynes, E.T., Bretthorst, G.L.: Probability Theory: The Logic of Science. Cambridge University Press, Cambridge, UK (2003)
7. Kreinovich, V.: Dimension compactification – a possible explanation for superclusters and for empirical evidence usually interpreted as dark matter. In: Ceberio, M., Kreinovich, V. (eds.), *How Uncertainty-Related Ideas Can Provide Theoretical Explanation for Empirical Dependencies*. Springer, Cham, Switzerland, to appear
8. Li, S., Ogura, Y., Kreinovich, V.: Limit Theorems and Applications of Set Valued and Fuzzy Valued Random Variables. Kluwer Academic Publishers, Dordrecht (2002)
9. Polchinski, J.: String Theory, vols. 1, 2. Cambridge University Press, Cambridge (1998)
10. Sheskin, D.J.: Handbook of Parametric and Nonparametric Statistical Procedures. Chapman & Hall/CRC, Boca Raton (2011)
11. Thorne, K.S., Blandford, R.D.: Modern Classical Physics: Optics, Fluids, Plasmas, Elasticity, Relativity, and Statistical Physics. Princeton University Press, Princeton (2017)

# Is Our World Becoming Less Quantum?



Lidice Castro and Vladik Kreinovich

**Abstract** According to the general idea of quantization, all physical dependencies are only approximately deterministic, and all physical “constants” are actually varying. A natural conclusion—that some physicists made—is that Planck’s constant (that determines the magnitude of quantum effects) can also vary. In this paper, we use another general physics idea—the second law of thermodynamics—to conclude that with time, this constant can only decrease. Thus, with time (we are talking cosmological scales, of course), our world is becoming less quantum.

## 1 Formulation of the Problem

**Our world is a quantum world.** According to modern physics, our world is a quantum world, a world described by quantum physics.

In order to formulate the problem that we will be solving in this paper, let us recall the main physical idea behind quantization. To convincingly describe this idea, let us briefly recall how physics came up with the quantum ideas in the first place; see, e.g., [1, 8].

**Classical mechanics.** Before quantum physics appeared, physics was described by deterministic equations, namely, by Newton’s equations. According to Newton’s equations

$$m \cdot \ddot{x}_i = F_i, \quad (1)$$

the trajectory  $x_i(t)$  of a particle with mass  $m$  is uniquely determined by this particle’s original location  $x_i(t_0)$ , original velocity  $\dot{x}_i(t_0)$ , and the forces  $F_i(x_j, t)$ .

---

L. Castro · V. Kreinovich (✉)

Department of Computer Science, University of Texas at El Paso, 500 W. University El Paso, El Paso, TX 79968, USA

e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

L. Castro

e-mail: [lcastrojim@miners.utep.edu](mailto:lcastrojim@miners.utep.edu)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

M. Ceberio and V. Kreinovich (eds.), *Decision Making Under Uncertainty*

and *Constraints*, Studies in Systems, Decision and Control 217,

[https://doi.org/10.1007/978-3-031-16415-6\\_23](https://doi.org/10.1007/978-3-031-16415-6_23)

The forces acting on a particle are, in their turn, uniquely determined by the locations and velocities of other particles. For example, the gravitational force  $F_i^{(a)}(t)$  acting on the particle  $a$  of mass  $m^{(a)}$  located at the point  $x_i^{(a)}$  is equal to

$$F_i^{(a)}(t) = G \cdot \sum_b \frac{m^{(a)} \cdot m^{(b)} \cdot (x_i^{(b)} - x_i^{(a)})}{\|x^{(b)} - x^{(a)}\|^3}, \quad (2)$$

where the sum is taken over all other bodies  $b$ , and  $\|x\| \stackrel{\text{def}}{=} \sqrt{x_1^2 + x_2^2 + x_3^2}$ .

Usually, the forces are described in terms of the corresponding field—which is, in turn, uniquely determined by the locations and velocities of all the particles.

**Need to go beyond classical mechanics.** In the traditional (non-quantum) mechanics, all the processes are deterministic.

It turns out that many processes in the world—radioactivity was probably the first example—are probabilistic. We cannot predict when an atom will experience a radioactive decay—we can only predict the probability of this happening.

**Original (first) quantization.** The need to take into account that many processes in the real world are probabilistic led to the development of the original quantum mechanics. In this formalism, the particle's trajectory is *not* determined uniquely by its original state and all the forces. When we know the original location and velocity of a particle, and we know the fields (hence the forces), we can only predict the probability distribution on the set of all possible trajectories—or, to be more precise, the probabilities of different possible results of measuring coordinates and velocities.

To what extent predictions are probabilistic is determined by a constant  $\hbar$  introduced by Max Planck, one of the founders of quantum physics. The smaller the Planck's constant, the closer all the trajectories to the Newton's ones.

**How this probabilistic idea is related to a more traditional understanding of quantization.** Planck did not start with the probabilistic nature of physics.

His original idea was different—that while in Newtonian physics, the values of all physical quantities change continuously, in reality, some quantities can only take values from some discrete set. In this case, transitions have to be abrupt. So, whether the object will change to a new state cannot be determined only by the state itself: for some time, the object stays in the same state, and then jumps to another state. This cannot be deterministic—thus we need a probabilistic description.

**Towards second quantization.** The original quantum mechanics worked very well—until it turned out that its predictions are not always in full accordance with the experiment. The solution—known as second quantization—came from the observation that while in the original quantum mechanics, the dependence of the trajectory on the fields is probabilistic, this theory still assumed that the fields themselves are uniquely determined by the positions and velocities of all the particles.

A natural idea was therefore to take into account that the fields are also not uniquely determined by the positions and velocities of all the particles, that all this information about the particles only enables us to predict the probabilistic distribution

on the set of all possible fields. This idea enabled researchers to match the theory with experimental data. The resulting quantum field theories are, at present, the main way how the world's processes are described in modern physics.

**Beyond second quantization.** In the first quantization, the probability distribution of the set of all the trajectories is uniquely determined by the particle's initial locations and velocities. The dependence of this probability distribution on the particle's locations and velocities is determined by the corresponding fields which are, in turn, uniquely determined by the particles' locations and velocities.

In the second quantization, the dependence of the field on the particles' initial locations and velocities also becomes probabilistic. Thus, the probability distribution on the set of all trajectories is no longer uniquely determined by the particle's initial locations and velocities.

A natural next idea is to assume that the probability distribution on the set of all possible fields is also not uniquely determined by the particles' initial locations and velocities, that all we can predict is the probability distribution on the set of all probability distributions, etc. The effect of this "third" quantization is too small to be noticeable at present, but this led many physicists—most famous of them John Wheeler—to formulate the general idea of quantization as saying that every dependence is probabilistic, and to analyze interesting consequences of this idea with respect to space-time; see, e.g., [5].

**Mathematical interruption: but is not probability distribution of the set of probability distributions the same as just a probability distribution?** Not really. Suppose that we have a probability 0.5 that a coin falls head. This means that for each coin out of the large set of minted coins, if we flip this coin many times, in half of the cases this coin will fall head, and in half of the cases, it will fall tail.

Suppose now that instead of the fixed probability  $p = 0.5$ , we have a probability distribution on the set of all possible values  $p$ . This would mean that for some coin, if we flip it many times, we will consistently get head 0.6 of the time, while for some other coin, we may consistently get head 0.4 of the time. Yes, if we combine all the experiment results together, we still get 0.5, but overall, the experiment results are different from what we would have observed if we had probability  $p = 0.5$ .

**Back to physics: according to the general quantization idea, Planck's constant is no longer a constant.** The same general logic leads to a conclusion that the local value of any physical constant is no longer a constant, that it can fluctuate from one moment to another, from one spatial location to another. For the speed of light—the parameter that, according to relativity theory, describing the space-time—these variations are well known: this is exactly what General Relativity teaches, that the space-time differs from one point to another.

But an interesting—and not as well accepted—conclusion is that the Planck's constant—that determines how deterministic is the dependence—is also not constant, it fluctuates from one point of space-time to another. Theories in which Planck's constant is actually a new physical field have indeed been proposed.

Now, we can formulate the problem that we analyze in this paper.

**What are the possible consequences of taking into account that Planck's constant is not a constant?** This is the question that we study in this paper.

## 2 Analysis of the Problem and the Resulting Conclusion

**Idea.** In our analysis, we cannot rely on specific equations—since the whole idea is that all equations are approximate. Instead, we have to rely on general principles.

One of these principles is the second law of thermodynamics—that the entropy  $S$  of any closed system, including the world as a whole, can only increase.

**What is entropy?** For a probabilistic distribution, entropy is the average number of “yes”-“no” questions that one needs to ask to uniquely determine the state; see, e.g., [4, 7].

If we have  $N$  original states, then we can divide these states into two equal parts and by a single question determine whether the current state belongs to the first half or to the second half. So, each question divides the number of states by 2. Thus,  $k$  questions divide the number of possible states by  $2^k$ , to  $N/2^k$ . Hence, to be left with a single possible state, the needed number of questions  $k$  is determined by the condition that  $N/2^k = 1$ , i.e., that  $2^k = N$  and  $k = \log_2(N)$ .

**We need to take uncertainty principle into account.** The world consists of particles. At each moment of time, the state of each particle is characterized by its location and its velocity—or, what is equivalent, its momentum. In quantum physics, we cannot determine both coordinate  $x_i$  and the corresponding momentum  $p_i$  exactly: there is the Uncertainty Principle, according to which the accuracies  $\Delta x_i$  and  $\Delta p_i$  with which we can determine these two quantities satisfy the inequality  $\Delta x_i \cdot \Delta p_i \geq \hbar$ . In other words, the state of each particle is characterized not by a single point  $(x, p) = (x_1, x_2, x_3, p_1, p_2, p_3)$  in the 6-D space (known as *phase space*), but by an area of 6-D volume  $\hbar^3$ .

Thus, the number  $N$  of distinguishable states can be obtained if we divide the 6-D volume  $V$  of the set of all possible points  $(x, p)$  by  $\hbar^3$ :  $N = V/\hbar^3$ .

**Conclusion: our world is becoming less quantum.** The entropy  $k = \log_2(N)$  can only increase, thus the number of states  $N$  can also only increase. For a fixed  $V$ , the only way for the number of states  $N$  to increase is when the Planck's constant  $\hbar$  decreases.

Thus, once we accept the general conclusion that Planck's constant can change, the only direction is which it can globally change is by decreasing. Since the value of this constant determines the intensity of quantum effects, this means that our world is becoming less and less quantum.

**Not to worry.** Of course, we are talking changes in cosmological time: so far, no macro-time experiments have found any change in the Planck's constant.

**How this affects our ability to compute—and thus, to predict.** On the one hand, if the world is becoming more deterministic, it will become easier to predict its future

state: all we need to do is predict one state, not the whole probability distribution on the set of all possible states.

On the other hand, our general ability to compute will decrease—since it will no longer be possible to use quantum computing, which is known to drastically decrease the computation time of many important computations; see, e.g., [2, 3, 6].

**Acknowledgements** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD2034030 (CAHSI Includes). It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

## References

1. Feynman, R., Leighton, R., Sands, M.: *The Feynman Lectures on Physics*. Addison Wesley, Boston, Massachusetts (2005)
2. Grover, L.K.: A fast quantum mechanical algorithm for database search. In: *Proceedings of the 28th ACM Symposium on Theory of Computing*, pp. 212–219 (1996)
3. Grover, L.K.: Quantum mechanics helps in searching for a needle in a haystack. *Phys. Rev. Lett.* **79**(2), 325–328 (1997)
4. Jaynes, E.T., Bretthorst, G.L.: *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, UK (2003)
5. Misner, W., Wheeler, J.A., Thorne, K.S.: *Gravitation*. Freeman & Co., San Francisco (1973)
6. Nielsen, M.A., Chuang, I.L.: *Quantum Computation and Quantum Information*. Cambridge University Press, Cambridge, U.K. (2000)
7. Sheskin, D.J.: *Handbook of Parametric and Non-Parametric Statistical Procedures*. Chapman & Hall/CRC, London, UK (2011)
8. Thorne, K.S., Blandford, R.D.: *Modern Classical Physics: Optics, Fluids, Plasmas, Elasticity, Relativity, and Statistical Physics*. Princeton University Press, Princeton, New Jersey (2017)

# **Applications to Psychology and Decision Making**

# As Complexity Rises, Meaningful Statements Lose Precision—But Why?



Miroslav Svítek, Olga Kosheleva, and Vladik Kreinovich

**Abstract** One of the motivations for Zadeh’s development of fuzzy logic—and one of the explanations for the success of fuzzy techniques—is the empirical observation that as complexity rises, meaningful statements lose precision. In this paper, we provide a possible explanation for this empirical phenomenon.

## 1 Formulation of the Problem

**Empirical fact.** Many researchers are familiar with Lotfi Zadeh’s observation that “As complexity rises, precise statements lose meaning and meaningful statements lose precision”; see, e.g., [3], p. 43. This is one of the most cited phrases by Zadeh. This empirical fact served as one of the main motivations for developing fuzzy techniques. This empirical fact also serves as a good explanation for why these techniques have been successful in many applications; see, e.g., [1, 2, 4–7].

**But why?** But how can we explain this empirical fact? In this paper, we provide a possible explanation.

---

M. Svítek

Faculty of Transportation Sciences, Czech Technical University in Prague, Konviktska 20, 110 00  
Prague 1, Czech Republic  
e-mail: [svitek@fd.cvut.cz](mailto:svitek@fd.cvut.cz)

O. Kosheleva · V. Kreinovich (✉)

University of Texas at El Paso, 500 W. University El Paso, El Paso, TX 79968, USA  
e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

O. Kosheleva

e-mail: [olgak@utep.edu](mailto:olgak@utep.edu)



## 2 Analysis of the Problem

**Towards reformulating the question in precise terms.** In general, we can have both precise and imprecise (“fuzzy”) statements about a system. The empirical fact—as observed by Zadeh—is that:

- when a system is simple, this system is adequately described by precise statements, while
- as the system becomes more complex, its adequate description requires more and more fuzzy statements.

How can we explain this empirical fact?

**Towards a corresponding model.** Let us consider possible statements  $S_1, \dots, S_n$  that we can make about a system.

In general, for each system, for each statement  $S_i$ , we can—following the general fuzzy methodology—describe our degree of confidence in this statement by a number  $x_i$  from the interval  $[0, 1]$ . So, our description boils down to a tuple  $x = (x_1, \dots, x_n)$  of numbers from the interval  $[0, 1]$ —i.e., to a point in an  $n$ -dimensional cube  $[0, 1]^n$ .

**What we want from a model.** We want our model to be consistent with all the different observation patterns characterising the system’s behavior. Let us denote the number of such patterns by  $p$ , and let us denote the requirement that the tuple  $x$  is consistent with the  $j$ th pattern by  $f_j(x) = 0$ .

Among all the models that are consistent with all the patterns, we should select a model which is the best: this could be the simplest to describe, the simplest to use, the least deviating from the current model, etc. “The best” means that some objective function  $a(x)$  take the largest possible value. We will call the value of this objective function for a given model  $x$  the “degree of quality” of this model.

In this term, selecting, among all the descriptions for which  $f_j(x) = 0$  for all  $j$ , the description  $x$  which is the best, means selecting the description for which the degree  $a(x)$  is the largest possible.

**Our descriptions are not ideal.** In general, every description is approximate. To get an ideal “most adequate” description, we need to consider more than  $n$  statements. In geometric terms, the ideal description is outside our  $n$ -dimensional cube  $[0, 1]^n$ .

It is reasonable to assume that the closer we are to this ideal description, the more adequate our model. At the point  $x$  that corresponds to the ideal model, the objective function  $a(x)$  attains its largest possible value, i.e., its global maximum. At every other point, if we get slightly closer to the ideal model, then the model becomes more adequate—i.e., the value of the corresponding objective function  $a(x)$  increases. Thus, the objective function cannot have any local maxima—because in the vicinity of a local maximum, the value of the function does not increase no matter in what direction we go. So, we expect the quality function  $a(x)$  to have no local maxima—its only maximum is the global maximum.

**Using known facts from calculus.** It is known that if a function has no local maxima inside an area, then its maximum in this area is attained on the border of this area.

**Let us start with the case when we have no observation patterns at all.** Let us first consider the trivial case when we have no observation patterns at all, i.e., in mathematical terms, when we have no constraints. As we have argued, the global maximum of this objective function is attained outside the cube, and there are no local maxima inside the cube. Thus, in line with the above fact from calculus, in this case, the desired maximum of the quality function  $a(x)$  is attained on the border of the  $n$ -dimensional cube.

This border consists of faces, which are described by the equations  $x_i = 0$  or  $x_i = 1$ . On each of these faces, we also do not expect to have a local maximum, so the optimal description should correspond to the border of each face, i.e., to the set of all points where two of the values  $x_i$  are equal to 0 or 1.

Following the same line of reasoning, we conclude that the maximum of the objective function  $a(x)$  on the  $n$ -dimensional cube is attained at an extreme point of the cube, i.e., at a point where each of the values  $x_i$  is equal to 0 or to 1.

So, in the absence of any observation patterns, the best description is a crisp description.

**What if we take observation patterns into account.** In general, the same argument as in the previous subsection leads us to the conclusion that the maximum of the quality function  $a(x)$  is attained at one of the extreme points of the corresponding area.

If we take observation patterns into account, this means that the corresponding area consists of all the tuples  $x$  for which  $f_j(x) = 0$  for all  $j$  from 1 to  $p$ , i.e., this area is equal to the following set:

$$S \stackrel{\text{def}}{=} \{x = (x_1, \dots, x_n) : 0 \leq x_i \leq 1 \text{ for all } i \text{ and } f_j(x) = 0 \text{ for all } j\}.$$

This set  $S$  is a particular case of a set defined by equalities  $f_j(x) = 0$  and inequalities  $g_k(x) \geq 0$ . For our sets, the inequalities are:

- $x_i \geq 0$ , i.e.,  $g_i(x) \geq 0$ , where  $g_i(x) \stackrel{\text{def}}{=} x_i$ , and
- $x_i \leq 1$ , i.e.,  $g_{n+i}(x) \geq 0$ , where  $g_{n+i}(x) \stackrel{\text{def}}{=} 1 - x_i$ .

In general, for a set defined by equalities and inequalities, an extreme point is when as many inequalities  $g_k(x) \geq 0$  as possible become equalities, i.e., satisfy the condition  $g_k(x) = 0$ . In general:

- if the number of equations is smaller than the number of unknowns, then we have many solutions;
- if the number of equations is equal to the number of unknowns, then we have a unique (or at least locally unique) solution; and
- if the number of equations is larger than the number of unknowns, then the system, in general, does not have a solution.

Thus, for a tuple  $x$  consisting of  $n$  real values, the largest number of equalities that this tuple can satisfy is  $n$ . So, extreme points correspond to the case when  $n$  equalities are satisfied.

We already have  $p$  equalities  $f_j(x) = 0$  that are satisfied. Thus, for an extreme point for which  $n$  equalities are satisfied,  $n - p$  remaining inequalities become equalities. These remaining inequalities have the form  $0 \leq x_i \leq 1$ , i.e., the form  $x_i \geq 0$  and  $1 - x_i \geq 0$ . Thus, the fact that these inequalities become equalities means that for the corresponding values  $i$ , we have:

- either  $x_i = 0$
- or  $1 - x_i = 0$ , i.e.,  $x_i = 1$ .

The fact that  $x_i = 0$  or  $x_i = 1$  means that in this description, the  $i$ th statement is crisp. We therefore conclude that in the best model, out of  $n$  statements  $S_i$ ,  $n - p$  of them are crisp.

The remaining truth values are determined by  $p$  equations  $f_j(x) = 0$ . In the general case, all components of a solution of a system of  $p$  equations with  $p$  unknowns are different from 0 and 1. Thus, in the general case, for the remaining  $p$  statements  $k$ , we have  $0 < x_k < 1$ —i.e., these statements are, in general, not crisp.

**Mathematical conclusion.** So, in the general case, if we have  $p$  observation patterns, then in the best description, we have:

- $p$  fuzzy statements, and
- $n - p$  crisp statements.

**How this is related to system complexity.** The more complex a system, the more different behavioral patterns it exhibits. This is, in a nutshell, is what we mean by a complex system. For example:

- a pendulum shows the same behavior all the time; in this sense, it is a simple system;
- on the other hand, a human being has many different patterns of behavior and is, thus, a complex system.

In the previous subsection of this section, we presented the conclusion of our analysis: that the more different patterns of behavior a system exhibits, the larger the number of fuzzy statements in this system's best description. So, indeed, as complexity rises, more meaningful statements become fuzzy—i.e., lose precision.

This is exactly Zadeh's observation. Thus, our analysis indeed explains this observation.

**Acknowledgements** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes), and by the AT&T Fellowship in Information Technology.

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478, and by a grant from the Hungarian National Research, Development and Innovation Office (NRDI).

## References

1. Belohlavek, R., Dauben, J.W., Klir, G.J.: *Fuzzy Logic and Mathematics: A Historical Perspective*. Oxford University Press, New York (2017)
2. Klir, G., Yuan, B.: *Fuzzy Sets and Fuzzy Logic*. Prentice Hall, Upper Saddle River (1995)
3. McNeill, D., Freiberger, P.: *Fuzzy Logic*. Simon & Schuster, New York (1993)
4. Mendel, J.M.: *Uncertain Rule-Based Fuzzy Systems: Introduction and New Directions*. Springer, Cham, Switzerland (2017)
5. Nguyen, H.T., Walker, C.L., Walker, E.A.: *A First Course in Fuzzy Logic*. Chapman and Hall/CRC, Boca Raton (2019)
6. Novák, V., Perfilieva, I., Močkoř, J.: *Mathematical Principles of Fuzzy Logic*. Kluwer, Boston, Dordrecht (1999)
7. Zadeh, L.A.: Fuzzy sets. *Information and Control* **8**, 338–353 (1965)

# Why People Overestimate Small Probabilities?



David Amparan and Vladik Kreinovich

**Abstract** It is a known empirical fact that people overestimate small probabilities. This fact seems to be inconsistent with the fact that we humans are the product of billions years of improving evolution—and that we therefore perceive the world as accurately as possible. In this paper, we provide a possible explanation for this seeming contradiction.

## 1 Formulation of the Problem

**People overestimate small probabilities.** It is known that people routinely overestimate small probabilities when making decisions. They overestimate the probability of rare side effects—and thus, refuse to take important vaccinations.

Experiments performed by the Nobelist Daniel Kahneman and his team show that indeed, most people overestimate small probabilities; see, e.g., [1] (see also [2, 3]).

**But why?** This is a fact, but how can we explain this fact from the biological viewpoint?

At first glance, the more adequately we understand the situation, the more adequate decision we can make. So why did evolution preserve this clearly biased perception of small probabilities?

**What we do in this paper.** In this paper, we provide a possible answer to this “why”-question.

---

D. Amparan · V. Kreinovich (✉)  
Department of Computer Science, University of Texas at El Paso, El Paso, TX 79968, USA  
e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

D. Amparan  
e-mail: [daamparan@miners.utep.edu](mailto:daamparan@miners.utep.edu)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023  
M. Ceberio and V. Kreinovich (eds.), *Decision Making Under Uncertainty and Constraints*, Studies in Systems, Decision and Control 217,  
[https://doi.org/10.1007/978-3-031-16415-6\\_25](https://doi.org/10.1007/978-3-031-16415-6_25)

173

## 2 How Do We Know Probabilities?

To provide an explanation, let us recall how we learn the probabilities.

Probabilities are estimates based on our experience. If we saw some event  $n$  times out of  $N$ , then we estimate the probability as the ratio  $n/N$ .

Of course, this is only an approximate estimate. If we flip a perfectly symmetric coin 10 times, we may get  $n = 5$  heads, but we may also get 6 or 4 or 7.

## 3 Which Outcomes Are Possible?

If an event has probability  $p$ , how many times out of  $N$  can it occur? If the actual probability is  $p$ , then out of  $N$  tries:

- the event happens on average in  $\mu \stackrel{\text{def}}{=} p \cdot N$  times, and
- the variance of number of events is equal to  $\sigma^2 = N \cdot p \cdot (1 - p)$ .

For small  $p$ , we have  $1 - p \approx 1$ , so  $\sigma^2 \approx \mu$  and thus,  $\mu \approx \sigma^2$ .

Usually:

- if we have a distribution with a known mean and standard deviation,
- we conclude—with high confidence—that the actual value is somewhere between  $\mu - k \cdot \sigma$  and  $\mu + k \cdot \sigma$ ; see, e.g., [4].

Here,  $k = 2, 3, 6, \dots$  depending on the desired level of confidence.

## 4 So What Can We Conclude About the Probability?

Suppose that some event occurred  $n$  time out of  $N$ . So, the only information that we can conclude about its probability  $p$  is that  $\mu - k \cdot \sigma \leq n \leq \mu + k \cdot \sigma$ .

Since  $\mu = \sigma^2$ , equivalently,

$$\sigma^2 - k \cdot \sigma \leq n \leq \sigma^2 + k \cdot \sigma, \quad (1)$$

where  $p = \sigma^2/N$ .

- The first of the two inequalities (1) is the inequality  $\sigma^2 - k \cdot \sigma \leq n$ , i.e., equivalently,  $\sigma^2 - k \cdot \sigma - n \leq 0$ . Due to the known properties of quadratic functions, this inequality means that the non-negative value  $\sigma$  is between the roots of the corresponding quadratic equation  $\sigma^2 - k \cdot \sigma - n = 0$ , i.e., that

$$\frac{k - \sqrt{k^2 + 4n}}{2} \leq \sigma \leq \frac{k + \sqrt{k^2 + 4n}}{2}.$$

The left-hand side expression is always non-positive, so the left inequality is always satisfied. Thus, to satisfy the first of the two inequalities (1), it is sufficient to have

$$\sigma \leq \frac{k + \sqrt{k^2 + 4n}}{2}. \tag{2}$$

- The second of the two inequalities (1) is the inequality  $n \leq \sigma^2 + k \cdot \sigma$ , i.e., equivalently,  $\sigma^2 + k \cdot \sigma - n \geq 0$ . Due to the known properties of quadratic functions, this inequality means that the non-negative value  $\sigma$  is either smaller than the smaller root of the corresponding quadratic equation  $\sigma^2 + k \cdot \sigma - n = 0$  or larger than the larger of the roots, i.e., that

$$\sigma \leq \frac{-k - \sqrt{k^2 + 4n}}{2} \text{ or } \frac{-k + \sqrt{k^2 + 4n}}{2} \leq \sigma.$$

The expression  $\frac{-k - \sqrt{k^2 + 4n}}{2}$  is always non-positive, so the first inequality is never satisfied. Thus, to satisfy the second of the two inequalities (1), it is sufficient to have

$$\frac{-k + \sqrt{k^2 + 4n}}{2} \leq \sigma. \tag{3}$$

By combining the inequalities (2) and (3), we conclude that

$$\frac{\sqrt{k^2 + 4n} - k}{2} \leq \sigma \leq \frac{\sqrt{k^2 + 4n} + k}{2}, \text{ so}$$

$$\underline{p} \stackrel{\text{def}}{=} \frac{2n + k^2 - k \cdot \sqrt{k^2 + 4n}}{2N} \leq p \leq \bar{p} \stackrel{\text{def}}{=} \frac{2n + k^2 + k \cdot \sqrt{k^2 + 4n}}{2N}.$$

## 5 Which Probability Value Should We Select?

We know that

$$\underline{p} \stackrel{\text{def}}{=} \frac{2n + k^2 - k \cdot \sqrt{k^2 + 4n}}{2N} \leq p \leq \bar{p} \stackrel{\text{def}}{=} \frac{2n + k^2 + k \cdot \sqrt{k^2 + 4n}}{2N}.$$

We have no reason to consider one of the values from the interval  $[p, \bar{p}]$  as more probable. So, it makes sense to consider all these values equally possible.

In this case, a natural idea is to select the average of these values, i.e., the midpoint

$$\frac{\underline{p} + \bar{p}}{2} = \frac{n}{N} + \frac{k^2}{2N}.$$

This value is always larger than the frequency  $n/N$ —which is the usual (and unbiased) estimate of the actual probability.

This provides a possible explanation of why we, in general, overestimate the values of small probabilities.

**Acknowledgements** This work was supported in part by the National Science Foundation grants:

- 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and
- HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported:

- by the AT&T Fellowship in Information Technology, and
- by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

The authors are thankful to all the participants of the 26th Annual UTEP/NMSU Workshop on Mathematics, Computer Science, and Computational Science (El Paso, Texas, November 5, 2021) for valuable discussions.

## References

1. Kahneman, D.: Thinking, Fast and Slow. Farrar, Straus, and Giroux, New York (2011)
2. Lorkowski, J., Kreinovich, V.: Fuzzy logic ideas can help in explaining Kahneman and Tversky's empirical decision weights. In: Zadeh, L., et al. (eds.) Recent Developments and New Direction in Soft-Computing Foundations and Applications, pp. 89–98. Springer, Berlin (2016)
3. Lorkowski, J., Kreinovich, V.: Bounded Rationality in Decision Making Under Uncertainty: Towards Optimal Granularity. Springer, Cham, Switzerland (2018)
4. Sheskin, D.J.: Handbook of Parametric and Non-Parametric Statistical Procedures. Chapman & Hall/CRC, London, UK (2011)



# Why Ovals in Eliciting Intervals?



Joshua Zamora and Vladik Kreinovich

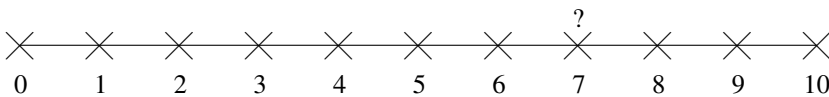
**Abstract** To elicit people’s opinions, we usually ask them to mark their degree of satisfaction on a scale—e.g., from 0 to 5 or from 0 to 10. Often, people are unsure about the exact degree: 7 or 8? To cover such situations, it is desirable to elicit not a single value but an interval of possible values. However, it turns out that most people are not comfortable with marking an interval. Empirically, it turned out that the best way to elicit an interval is to ask them to draw an oval whose intersection with the 0-to-10 line is the desired interval. Surprisingly, this seemingly more complex 2-D task is easier for most people than a seemingly simpler 1-D task of drawing an interval. In this paper, we provide a possible explanation of why eliciting an interval-related oval is more efficient than eliciting the interval itself.

## 1 Need to Elicit Intervals

People’s opinion is usually elicited by asking people to mark a point on a scale. This is how, e.g., students evaluate their instructors.

- In some cases, people are absolutely certain about their marks.
- However, in many other cases, they are not so sure. For example, a person may hesitate where to mark a good but not excellent service by 7 or 8 on a 0 to 10 scale.

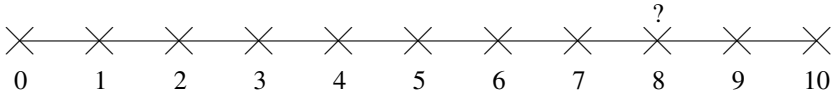
Since the usual scale only allows one mark, the person will put either 7 or 8.



---

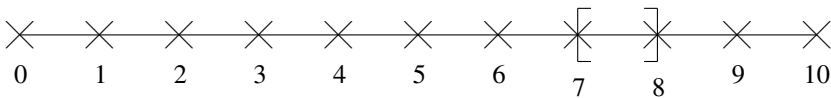
J. Zamora · V. Kreinovich (✉)  
Department of Computer Science, University of Texas at El Paso, El Paso, TX 79968, USA  
e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

J. Zamora  
e-mail: [jazamora6@miners.utep.edu](mailto:jazamora6@miners.utep.edu)



We could get a more adequate understanding of the people's opinions if we allow the user, in such situations, to explicitly explain that both 7 and 8—and thus, all the values in between—could be this person's marks.

In other words, we would get a more adequate description of people's opinions if we allow them to describe their opinion by intervals, and not just by the numerical values.

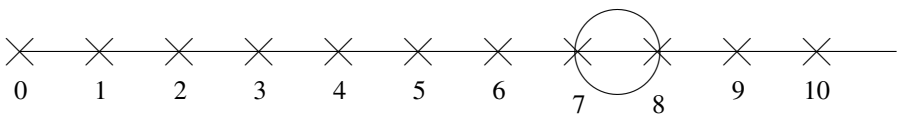


## 2 Eliciting Intervals Is not Easy

Eliciting intervals would be beneficial for processing people's opinions. However, people are not accustomed to marking intervals. Therefore, they are reluctant to do it.

To make this task easier for users, researchers tried different approaches. Interestingly, a successful approach came when researchers decided to elicit a 2-D figure.

Namely, they elicit an oval whose intersection with the straight line provides the desired interval; see [1].



## 3 Why: The Question and Our Explanation

**Why?** A 2-D oval contains more information than the resulting interval. So why is it easier for the users to provide ovals than to directly provide intervals?

**Our explanation.** Psychologists have found that the perceived complexity of a curve increases with the number of vertices; see, e.g., [2].

- Smooth curves like ovals are the simplest.



- On the other hand, an interval—with 2 vertices—is much more complex.



This explains why it is easier for people to draw an oval than to directly draw an interval.

**Acknowledgements** This work was supported in part by the National Science Foundation grants:

- 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and
- HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported:

- by the AT&T Fellowship in Information Technology, and
- by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

The authors are thankful to all the participants of the 26th Annual UTEP/NMSU Workshop on Mathematics, Computer Science, and Computational Science (El Paso, Texas, November 5, 2021) for valuable discussions.

## References

1. Ellerby, Z., Wagner, C.: Do people prefer to give interval-valued or point estimates and why? In: Proceedings of the 2021 IEEE International Conference on Fuzzy Systems FUZZ-IEEE'2021, Luxembourg, July 11–14, 2021
2. Wilder, J., Feldman, J., Singh, M.: Contour complexity and contour detection. *J. Vis.* **15**(5, 6), 1–16 (2015)

# Why Moments (and Generalized Moments) Are Used in Statistics and Why Expected Utility Is Used in Decision Making: A Possible Explanation



R. Noah Padilla and Vladik Kreinovich

**Abstract** Among the most efficient characteristics of a probability distribution are its moments and, more generally, generalized moments. One of the most adequate numerical characteristics describing human behavior is expected utility. In both cases, the corresponding characteristic is the sum of results of applying appropriate non-linear functions applied to individual inputs. In this paper, we provide a possible theoretical explanation of why such functions are efficient.

## 1 Formulation of the Problem

In this paper, we provide a new explanation of two seemingly unrelated phenomena:

- that moments (and, more generally, generalized moments) are effectively used in statistics; see, e.g., [8], and
- that expected utility is effectively used in decision making; see, e.g., [1–7].

Before we provide the corresponding explanations, let us first briefly describe these two phenomena.

**Moments and generalized moments: a brief reminder.** One of the most frequent ways to characterize a random variable  $x$  is to use moments—i.e., expected values  $E[x^k]$  of some integer power of this variable—and, more generally, generalized moments, i.e., expected values  $E[f(x)]$  of some function of the random variable.

For each random quantity  $q$ , its expected value is equal to the limit of its average observations  $q_1, \dots, q_n, \dots$ :

---

R. N. Padilla · V. Kreinovich (✉)

Department of Computer Science, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA

e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

R. N. Padilla

e-mail: [rnpadilla2@miners.utep.edu](mailto:rnpadilla2@miners.utep.edu)

$$E[q] = \lim_{n \rightarrow \infty} \frac{q_1 + \cdots + q_n}{n}.$$

By definition of the limit, this means that when  $n$  becomes larger and larger, the average

$$\frac{q_1 + \cdots + q_n}{n}$$

gets closer and closer to the expected value. Thus, a reasonable way to estimate the mean based on the observations  $q_i$  is to take the arithmetic average of all the observed values:

$$E[q] \approx \frac{q_1 + \cdots + q_n}{n}.$$

In particular, to estimate the value  $E[f(x)]$  of the generalized moment (or, in particular, of a usual moment corresponding to  $f(x) = x^k$ ) based on the observations  $x_1, \cdots, x_n$ , it is reasonable to use the corresponding arithmetic average

$$E[f(x)] \approx \frac{f(x_1) + \cdots + f(x_n)}{n}. \quad (1)$$

**Alternative formulas for moments and generalized moments.** In some cases, we have limited number of values  $v_1, \cdots, v_k$  ( $k \ll n$ ) that the variables  $x_i$  can take. In this case, each term  $f(x_i)$  in the sum

$$s \stackrel{\text{def}}{=} f(x_1) + \cdots + f(x_n) \quad (2)$$

is equal to one of the  $k$  values  $f(v_j)$ ,  $1 \leq j \leq k$ . In such cases, we can simplify the formula (2) by grouping together terms equal to  $f(v_1)$ , terms equal to  $f(v_2)$ , etc. Then, we get

$$s = f(v_1) + \cdots + f(v_1) (n_1 \text{ times}) + \cdots + f(v_k) + \cdots + f(v_k) (n_k \text{ times}),$$

where  $n_j$  denotes the number of terms  $f(x_i)$  which are equal to  $f(v_k)$ , or, equivalently,

$$s = f(x_1) + \cdots + f(x_n) = n_1 \cdot f(v_1) + \cdots + n_k \cdot f(v_k).$$

Substituting this expression into the formula (1), we conclude that

$$E[f(x)] \approx \frac{n_1}{n} \cdot f(v_1) + \cdots + \frac{n_k}{n} \cdot f(v_k). \quad (3)$$

Here, the ratio  $\frac{n_j}{n}$  is the frequency with which the value  $v_j$  appears in the observations, i.e., in effect, the probability  $p_j$  of this value—to be more precise, the probability is defined as the limit of such a frequency, but since we are considering large  $n$ ,

probability and frequency are approximately the same. Thus, the formula (3) takes the form

$$E[f(x)] \approx p_1 \cdot f(v_1) + \cdots + p_k \cdot f(v_k). \quad (4)$$

**Expected utility: a brief reminder.** It is known—see above references—that a rational person, when making a decision, should maximize the value of a special expression known as *expected utility*

$$u \stackrel{\text{def}}{=} p_1 \cdot u(v_1) + \cdots + p_k \cdot u(v_k), \quad (5)$$

where:

- $v_1, \dots, v_k$  are possible consequences of the selected action,
- $p_j$  is the (subjective) probability of getting an alternative  $v_j$ , and
- $u(v_j)$  is a number—called *utility*—that characterizes the value of the alternative  $v_j$  to the decision maker.

*Comment.* The main use of expected utility is to decide which alternative is better, i.e., which decision we should make. From this viewpoint, what is important are not the numerical values (5) themselves, but which values are larger and which are smaller. From this viewpoint, instead of the values  $u$ , we could use the values  $g(u)$  for any increasing function  $g(u)$ —since for an increasing function  $u < u'$  if and only if  $g(u) < g(u')$ .

**Is there a common explanation for these two formulas?** There exist explanations for both formulas (4) and (5), explanations based on different ideas; see, e.g., the above references. However, the fact that the expressions (4) and (5) are very similar—in both cases, we have a linear combination of the values of some function ( $f(v)$  in the first case,  $u(v)$  in the second case) applied to different values  $v_1, \dots, v_k$ —made us think that there also be a joint explanation for these two seemingly unrelated formulas. In this paper, we provide a possible common explanation.

## 2 Main Ideas Behind Our Explanation

**In many practical problem, computation time is a big issue.** Nowadays, we get a lot of data, and we have a lot of computational ability. However, still, computation time remains a big issue. For example, with numerous weather sensors almost everywhere, we get a lot of data that enables us to predict tomorrow's weather reasonably well—but because of the huge amount of data and, as a result, a huge amount of computations, the only way to predict weather is to use high-performance computers, where a large number of processors are working in parallel, and even on such

computers, weather prediction takes hours (and became possible only after special time-saving algorithms were implemented).

In many other problems we still cannot perform computations in desired time. For example, in principle, it is possible to predict somewhat accurately in what direction a potentially deadly tornado will go in the next 15 min—but the resulting computations so far require much longer than 15 min and are, therefore, practically useless. From this viewpoint, it is desirable to come up with computations that can be performed as fast as possible.

**Which computations are the fastest?** Of course, to make computations faster, we need to parallelize computations as much as possible. On a parallel computer, first, all the processors perform one computation step, then they all perform another step, etc. To minimize the overall computation time:

- we need to minimize the number of steps, and
- we need to minimize the time needed for each step—i.e., in other words, perform, at each step, computations which are as fast as possible.

**Which computational steps are the fastest?** When we process numbers, computation on a deterministic computer means, in effect, computing the value of some function of an input. Overall, the function we compute is a composition of functions computed on consequent steps.

Among different functions of several variables, linear functions, i.e., functions of the type

$$f(x_1, \dots, x_n) = a_0 + a_1 \cdot x_1 + \dots + a_n \cdot x_n \quad (6)$$

are the easiest (thus fastest) to compute.

However, if we only use linear computational steps, then, due to the fact that a composition of linear functions is linear, we will only be able to compute linear functions, while in real life, many processes are nonlinear. Thus, in addition to linear computational steps, we also need some nonlinear ones.

In general, the more inputs a function has, the longer it takes to process all these inputs and to compute the value of this function. From this viewpoint, among all nonlinear functions, the fastest to compute are nonlinear functions of one variable  $y = s(x)$ . Thus, fastest computations should consist of two types of computational steps:

- linear steps, on which we compute a linear combination (6) of the inputs, and
- nonlinear steps, on which we compute the value of a function of one variable

$$y = s(x).$$

**To make computations fast, consequent computational steps must be of different types.** Indeed, if we have a linear step followed by a linear step, then all these two steps compute is a composition of two linear functions—which, as we have mentioned, is also a linear function. Thus, instead of these two steps, we can have a single linear step, in which we directly compute this composition.

Similarly, if we have a nonlinear step  $y = s(x)$  followed by another nonlinear step  $z = s'(y)$ , then all these two steps compute is a composition  $z = s'(s(x))$  of these two functions—i.e., also a nonlinear function of one variable. Thus, instead of these two steps, we can have a single nonlinear step, in which we directly compute this composition.

So, in general, to make computations faster, we need to make sure that consequent computational steps are of different types, i.e., that:

- a linear computational step is followed by a nonlinear one, and
- a nonlinear computational step is followed by a linear one.

**What can we compute with the smallest possible number of computational steps.**

Now that we know which are the fastest computational steps, let us analyze which functions can be computed by using the smallest possible number of computational steps.

The smallest possible number of computational steps is 1. In one step, we can compute either a linear function or a function of one variable. In both statistics and decision making applications, we need to process several numbers:

- in the statistics cases, we need to take into account (and thus, to process) several observations  $x_1, \dots, x_n$ , and
- in the decision making cases, we need to take into account (and thus, to process) several different possible consequences  $v_1, \dots, v_k$  of the analyzed decision.

Thus, if we limit ourselves to a single computational step, we cannot use a function of one variable. Therefore, we have to use a linear function. In case of the statistical analysis, this corresponds to using the first moment

$$E[x] \approx \frac{x_1 + \dots + x_n}{n} = p_1 \cdot v_1 + \dots + p_k \cdot v_k,$$

for some values  $p_j$ . In case of decision making, this corresponds to having utility proportional to the numerical value  $v_j$  of each alternative:

$$u = p_1 \cdot v_1 + \dots + p_k \cdot v_k.$$

In line with the general fact that some real-life dependencies are nonlinear, both in statistical analysis and in decision making, we may need to use nonlinear functions to get a more adequate description. In this case, we need to use at least two computational steps.

**Two stages: possible options.** Due to the above, these stage must be different. So, we have two options:

- the first option is to have a linear stage followed by a nonlinear stage, and
- the second option is to have a nonlinear stage followed by a linear stage.



**Two stages: first option.** If the first stage is linear and the following one nonlinear, then, in general, we compute a function

$$f \left( a_0 + \sum_{j=1}^k a_j \cdot v_j \right).$$

Comparing such values is equivalent to comparing the corresponding linear combinations  $a_0 + \sum_{j=1}^k a_j \cdot v_j$ , and we know that such a linearized approach does not work for many real-life phenomena.

**Two stages: second option.** If the first stage is nonlinear and the second one linear, then we compute expressions  $a_0 + \sum_{j=1}^k a_j \cdot f_j(v_j)$ . This provides a more general opportunities for comparison.

In particular, if a priori, we have no reason to prefer some  $j$ 's, then it makes sense to use the same nonlinear function  $f_j(v) = f(v)$  to process all the inputs. Thus, we get the expression

$$a_0 + \sum_{j=1}^k a_j \cdot f(x_j). \quad (7)$$

**This expression is exactly what we wanted to explain.** The formula (7) is exactly what is used when we use generalized moments or expected utility. Thus, we have indeed explained the desired expressions.

**Acknowledgements** This work was supported in part by the National Science Foundation grants:

- 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and
- HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported:

- by the AT&T Fellowship in Information Technology,
- by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478, and
- by a grant from the Hungarian National Research, Development and Innovation Office (NRDI).

## References

1. Fishburn, P.C.: Utility Theory for Decision Making. Wiley, New York (1969)
2. Fishburn, P.C.: Nonlinear Preference and Utility Theory. The John Hopkins Press, Baltimore (1988)

3. Kreinovich, V.: Decision making under interval uncertainty (and beyond). In: Guo, P., Pedrycz, W. (eds.) *Human-Centric Decision-Making Models for Social Sciences*, pp. 163–193. Springer, Berlin (2014)
4. Luce, R.D., Raiffa, R.: *Games and Decisions: Introduction and Critical Survey*. Dover, New York (1989)
5. Nguyen, H.T., Kosheleva, O., Kreinovich, V.: Decision making beyond Arrow's 'impossibility theorem', with the analysis of effects of collusion and mutual attraction. *International Journal of Intelligent Systems* **24**(1), 27–47 (2009)
6. Nguyen, H.T., Kreinovich, V., Wu, B., Xiang, G.: *Computing Statistics under Interval and Fuzzy Uncertainty*. Springer, Berlin (2012)
7. Raiffa, H.: *Decision Analysis*. McGraw-Hill, Columbus (1997)
8. Sheskin, D.J.: *Handbook of Parametric and Non-Parametric Statistical Procedures*. Chapman & Hall/CRC, London, UK (2011)

# Decision Making Under Uncertainty: Cases When We Only Know an Upper Bound or a Lower Bound



Toshiki Kamio, Gavin Baechle, and Vladik Kreinovich

**Abstract** In situations when we have a perfect knowledge about the outcomes of several situations, a natural idea is to select the best of these situations. For example, among different investments, we should select the one with the largest gain. In practice, however, we rarely know the exact consequences of each action. In some cases, we know the lower and upper bounds on the corresponding gain. It has been proven that in such cases, an appropriate decision is to use Hurwicz optimism-pessimism criterion. In this paper, we extend the corresponding results to the cases when we only know an upper bound or a lower bound.

## 1 Formulation of the Problem

In investment, when a person knows the exact monetary consequence of each action, he/she naturally selects an action with the largest possible gain.

In practice, we usually know the consequences only with some uncertainty. For example, instead of the exact gain value, the whole set  $S$  of different possible gain values are consistent with our knowledge. How should we then make a decision? What is the equivalent price  $v(S)$  that we are willing to pay to participate in the corresponding action?

For example, we may know the lower bound  $a$  and the upper bound on the gain. In this case, the set  $S$  is the interval  $[a, b]$ .

---

T. Kamio · G. Baechle · V. Kreinovich (✉)  
Department of Computer Science, University of Texas at El Paso, El Paso, TX 79968, USA  
e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

T. Kamio  
e-mail: [tkamio@miners.utep.edu](mailto:tkamio@miners.utep.edu)

G. Baechle  
e-mail: [gpbachle@miners.utep.edu](mailto:gpbachle@miners.utep.edu)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023  
M. Ceberio and V. Kreinovich (eds.), *Decision Making Under Uncertainty  
and Constraints*, Studies in Systems, Decision and Control 217,  
[https://doi.org/10.1007/978-3-031-16415-6\\_28](https://doi.org/10.1007/978-3-031-16415-6_28)

189

Alternatively, we may know:

- only the lower bound, in which case  $S = [a, \infty)$  or
- only the upper bound, in which case  $S = (-\infty, b]$ .

## 2 How This Problem Is Solved If We Know Both Bounds

**Shift-invariance.** Suppose that we are willing to pay  $v(S)$  for the set  $S$ . Then, for the set  $S$  and a fixed amount  $c$ , we are willing to pay  $v(S) + c$ .

In this joint offer, the set of possible outcomes is

$$S + c \stackrel{\text{def}}{=} \{s + c : s \in S\}.$$

So, a reasonable price to pay for this joint offer is  $v(S + c)$ .

These are two different descriptions of the same situation. The price that are willing to pay to participate in this situation should not depend on how we describe this situation. So, we should have  $v(S + c) = v(S) + c$ . This property is called *shift-invariance*.

**Scale-invariance.** Another idea is that the transformation  $S \mapsto v(S)$  should not depend on the choice of the monetary unit. For example, if we select pesos instead of dollars, we should get the same equivalent value.

In precise terms, this means  $v(\lambda \cdot S) = \lambda \cdot v(S)$ , where

$$\lambda \cdot S \stackrel{\text{def}}{=} \{\lambda \cdot s : s \in S\}.$$

This property is known as *scale-invariance*.

**Additivity.** The third idea is that participation in two independence actions, with sets  $S_1$  and  $S_2$ , is equivalent to participation in a single action with the result

$$S_1 + S_2 = \{s_1 + s_2 : s_1 \in S_1 \ \& \ s_2 \in S_2\}.$$

These are two ways of representing the same situation. So we should have

$$v(S_1 + S_2) = v(S_1) + v(S_2).$$

This property is known as *additivity*.

**Known results** (see, e.g., [2]). For interval uncertainty, additivity implies Hurwicz formula  $v([a, b]) = \alpha \cdot b + (1 - \alpha) \cdot a$  for some  $\alpha \in [0, 1]$ . The same formula emerges if we assume shift- and scale-invariance.

### 3 What if We only Know the Lower Bound

**Description of the case.** Suppose that we only know the lower bound  $a$ . In this case, the set of possible gains is the infinite interval  $[a, \infty)$ . What is the price

$$f(a) \stackrel{\text{def}}{=} v([a, \infty))$$

that we should pay for this situation?

**What if we assume additivity.** For infinite intervals,

$$[a, \infty) + [b, \infty) = [a + b, \infty).$$

Thus, additivity implies that  $f(a + b) = f(a) + f(b)$ , for  $f(a) \geq a$ .

It is known that this functional equation implies that  $f(a) = k \cdot a$ ; see, e.g., [1]. The condition  $a \leq f(a)$  implies that  $k \geq 1$ .

**What if we assume scale-invariance.** Here,

$$\lambda \cdot [a, \infty) = [\lambda \cdot a, \infty).$$

Thus, scale-invariance means  $f(\lambda \cdot a) = \lambda \cdot f(a)$  for all  $\lambda > 0$  and  $a$ . In particular:

- for  $a = 1$ , we get  $f(\lambda) = k_+ \cdot \lambda$ , where  $k_+ \stackrel{\text{def}}{=} f(1)$ ; and
- for  $a = -1$ , we similarly get  $f(-\lambda) = k_- \cdot \lambda$ , i.e.,  $f(x) = (-k_-) \cdot x$ .

**What if we assume shift-invariance.** Here,

$$[a, \infty) + c = [a + c, \infty).$$

Thus, shift-invariance means that  $f(a + c) = f(a) + c$ . In particular, for  $a = 0$ , we get  $f(c) = a_0 + c$ , where we denoted  $a_0 \stackrel{\text{def}}{=} f(0)$ . Since  $f(0) \geq 0$ , we have  $a_0 \geq 0$ .

### 4 What if We only Know the Upper Bound

**Description of the case.** Suppose that we only know the upper bound  $a$ . In this case, the set of possible gains is the infinite interval  $(-\infty, a]$ . What is the price

$$g(a) \stackrel{\text{def}}{=} v((-\infty, a])$$

that we should pay for this situation?

**What if we assume additivity.** For infinite intervals,

$$(-\infty, a] + (-\infty, b] = (-\infty, a + b].$$

Thus, additivity implies that  $g(a + b) = g(a) + g(b)$ , for  $g(a) \leq a$ .

It is known that this functional equation implies that  $g(a) = k \cdot a$ ; see, e.g., [1]. The condition  $g(a) \leq a$  implies that  $k \leq 1$ .

**What if we assume scale-invariance.** Here,

$$\lambda \cdot (-\infty, a] = (-\infty, \lambda \cdot a].$$

Thus, scale-invariance means  $g(\lambda \cdot a) = \lambda \cdot g(a)$  for all  $\lambda > 0$  and  $a$ . In particular:

- for  $a = 1$ , we get  $g(\lambda) = k_+ \cdot \lambda$ , where  $k_+ \stackrel{\text{def}}{=} g(1)$ ; and
- for  $a = -1$ , we similarly get  $g(-\lambda) = k_- \cdot \lambda$ , i.e.,  $g(x) = (-k_-) \cdot x$ .

**What if we assume shift-invariance.** Here,

$$(-\infty, a] + c = (-\infty, a + c].$$

Thus, shift-invariance means that  $g(a + c) = g(a) + c$ . In particular, for  $a = 0$ , we get  $g(c) = a_0 + c$ , where we denoted  $a_0 \stackrel{\text{def}}{=} g(0)$ . Since  $g(0) \leq 0$ , we have  $a_0 \leq 0$ .

**Acknowledgements** This work was supported in part by the National Science Foundation grants:

- 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and
- HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported:

- by the AT&T Fellowship in Information Technology, and
- by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

The authors are thankful to all the participants of the 26th Annual UTEP/NMSU Workshop on Mathematics, Computer Science, and Computational Science (El Paso, Texas, November 5, 2021) for valuable discussions.

## References

1. Aczél, J., Dhombres, J.: Functional Equations in Several Variables. Cambridge University Press, Cambridge (2008)
2. Lorkowski, J., Aliev, R., Kreinovich, V., Towards decision making under interval, setvalued, fuzzy, and z-number uncertainty: a fair price approach. In: Proceedings of the IEEE World Congress on Computational Intelligence WCCI'2014, Beijing, China, July 6–11, 2014

# Why Do People Become Addicted: Towards a Theoretical Explanation for Eyal’s Experiment-Based Hook Model



Christopher Reyes and Vladik Kreinovich

**Abstract** Why do people become addicted, e.g., to gambling? Experiments have shown that simple lotteries, in which we can win a small prize with a certain probability, and not addictive. However, if we add a second possibility—of having a large prize with a small probability—the lottery becomes highly addictive to many participants. In this paper, we provide a possible theoretical explanation for this empirical phenomenon.

## 1 Formulation of the Problem

**Addiction: bad and not so bad.** The word “addiction” has a negative connotation: people get addicted to gambling, to drugs, to alcohol, to smoking; they try it first, and then they feel the urge to continue the corresponding habit. However, from the psychological viewpoint, the same habit-forming can have (and often has) positive effects as well: people get addicted to healthy lifestyle, like eating healthy food and exercising regularly, people get addicted to their creative activities ranging from art and music to scientific research, people fall in love with each other—which is usually a good type of addiction.

For bad addiction, we need to understand where it comes from so we can prevent it and—if it already happened—cure it. For good addition, we also need to understand where it comes from, so that we can have more people living healthy lives, we can have more people exploring their creativity, etc. In both cases, it is important to understand where addiction comes from, i.e., how we form the resulting habits.

**Eyal’s experiments and the resulting Hook Model.** Understanding can mean different things. We can discuss what physiological processes occur in the brain when

---

C. Reyes · V. Kreinovich (✉)

Department of Computer Science, University of Texas at El Paso, 500 W. University El Paso, El Paso, TX 79968, USA

e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

C. Reyes

e-mail: [creyes24@miners.utep.edu](mailto:creyes24@miners.utep.edu)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

M. Ceberio and V. Kreinovich (eds.), *Decision Making Under Uncertainty*

*and Constraints*, Studies in Systems, Decision and Control 217,

[https://doi.org/10.1007/978-3-031-16415-6\\_29](https://doi.org/10.1007/978-3-031-16415-6_29)

a person becomes addicted. In the future, this may help us prevent the formation of bad habits and promote formation of good ones, but as of now, the results of such an analysis are somewhat far away from practical applications. In general, we are not yet able to use this knowledge to prevent or promote habit forming.

More practical results have reasonably recently come from a different study: an analysis of which situations cause addictions and which do not—without the physiological analysis of how exactly addiction is formed in the brain. Such studies have indeed been performed, they are describe in Nir Eyal’s book; see [1] and references therein. Eyal’s results can be best explained on the example of gambling addiction—since in gambling (as opposed to other bad addictions), rewards and risks can be clearly stated in objective numerical form.

Eyal started with a seemingly natural simple gambling model, in which a person gets:

- a reward  $r$  with some probability  $p$ , and
- no reward at all with the remaining probability  $1 - p$ .

This can be a simplified model of playing a lottery, this can be a simplified version of playing the slot machine at a casino, etc. Somewhat surprisingly, this seemingly natural arrangement did not lead to any serious habit forming—participants played a little bit, but did not form a habit of playing.

The situation changed drastically when he introduced a somewhat more realistic description of a gambling situation, in which there are two levels of rewards:

- a very large reward  $R$  that happens with a very low probability  $p_\ell$ , and
- a medium-size (actually, small) reward  $r$  that happens with a medium-size probability  $p_m$ .

For example, in a lottery where a lottery ticket costs 1 dollar, many people get a \$5 prize and very few get a very big, multi-million dollar prize. In simulated situations, a significant proportion of participants became addicted to playing this lottery: they eagerly participated in it again and again.

**What we do in this paper.** In this paper, we provide a natural explanation for this phenomenon: namely, we explain why lotteries with two levels of rewards are more addictive.

## 2 Analysis of the Problem and the Resulting Explanation

**Naive picture of the situation.** In order to understand the situation, let us start with the first—as it turns out, naive—description of the situation. In this (naive) picture, when people engage in some repeated financial arrangements, they expect to earn some money. This is why people invest money in stocks or place them in a savings account—this way they expect to gain more than they invested. This is true for



investments, but can this explain why people play lotteries in the first place? As it turns out, not really.

Indeed, if a person plays the above-described simple lottery—in which we get a reward  $r$  with probability  $p$ —a large number of times  $N$ , then we get this reward in approximately  $p \cdot N$  cases, so the overall reward is equal to  $p \cdot N \cdot r$ . To get this reward, the person needs to buy  $N$  lottery tickets. So, if we denote the price of a ticket by  $t$ , the person spends the amount  $t \cdot N$ .

In this picture, a person should play the lottery only if his/her expected gain is larger than his/her investment, i.e., if  $p \cdot N \cdot r > t \cdot N$ . But where can this extra money come from? The only possibility is for this money to come from the lottery organizers, but this does not make sense: why would the lottery organizers give away money? Lotteries usually earn money for the state, not lose them. So, this naive picture not only does not explain why people get addicted, it does not even explain why people play lotteries in the first place.

A similar conclusion can be made for any lottery  $i$ , in which we get:

- money reward  $r_{i1}$  with probability  $p_{i1}$ ,
- money reward  $r_{i2}$  with probability  $p_{i2}$ , etc.,

In this case, after  $N$  plays, we get:

- money reward  $r_{i1}$  approximately  $N \cdot p_{i1}$  times,
- money reward  $r_{i2}$  approximately  $N \cdot p_{i2}$  times, etc.

So, the overall reward is equal to

$$N \cdot p_{i1} \cdot r_{i1} + N \cdot p_{i2} \cdot r_{i2} + \dots ,$$

and the average reward per play is equal to

$$p_{i1} \cdot r_{i1} + p_{i2} \cdot r_{i2} + \dots \tag{1}$$

Unless the lottery organizers give out money for free, this expected amount cannot be larger than the price of a lottery ticket. Thus, from this naive viewpoint, people should not play lotteries at all—but they do. Why?

**A more adequate picture of the situation.** Researchers have been analyzing human decision making for many decades. In particular, they analyzed a question of how a rational person should make decisions. Their conclusion (see, e.g., [2, 3, 5, 8–11]) is that a rational person, when presented several situations  $i$  in which he/she will get:

- money reward  $r_{i1}$  with probability  $p_{i1}$ ,
- money reward  $r_{i2}$  with probability  $p_{i2}$ , etc.,

should select an alternative  $i$  for which the following expression is the largest possible:

$$u_i = p_{i1} \cdot u(r_{i1}) + p_{i2} \cdot u(r_{i2}) + \dots , \tag{2}$$

for some function  $u(r)$  (called *utility function*) describing this person's preferences.

This formula is similar to the above "naive" formula (1), the main difference is that instead of computing the expected value (1) of the monetary gain  $r_{ij}$ , we compute the expected value of the *utility*  $u(r_{ij})$  of this gain—which, crudely speaking, describes the people "degree of happiness" upon receiving such gain.

That the degree of happiness is not directly proportion to the monetary amount makes sense. If it was, then every time you get an extra \$1, you would experience the same increase in happiness. In reality, however:

- if you have no money and someone gives you \$1, then you become very happy;
- on the other hand, if you already have \$100 and someone gives you \$1, then your degree of happiness does not change that much.

In other words, the perceived difference between having \$0 and \$1 is much higher than the perceived difference between having \$100 and \$101.

Empirical studies found that this aspect of human behavior can be reasonably well described if we use the square root utility function  $u(r) = \sqrt{r}$ ; see, e.g., [4, 7]. In this case, indeed, the difference  $u(101) - u(100) = \sqrt{101} - \sqrt{100} \approx 0.05$  is much smaller than the difference  $u(1) - u(0) = \sqrt{1} - \sqrt{0} = 1$ .

In this approach, the person is willing to play a lottery in which he/she gains  $r_j$  with probability  $p_j$ ,  $j = 1, 2, \dots$  if his/her expected utility

$$p_1 \cdot \sqrt{r_1} + p_2 \cdot \sqrt{r_2} + \dots$$

is larger than the utility  $\sqrt{t}$  corresponding to the ticket price  $t$ :

$$p_1 \cdot \sqrt{r_1} + p_2 \cdot \sqrt{r_2} + \dots \geq \sqrt{t}. \quad (1)$$

**This more adequate model still does not explain why people play lotteries.** Indeed, as one can easily check, the function  $f(x) = \sqrt{x}$  is strictly concave—since its second derivative is negative. Concaveness means that for all possible convex combinations  $r = \sum_{i=1}^n p_i \cdot r_i$ , where  $p_i \geq 0$  and  $\sum_{i=1}^n p_i = 1$ , we have

$$\sum_{i=1}^n p_i \cdot f(r_i) \leq f\left(\sum_{i=1}^n p_i \cdot r_i\right).$$

Strict concaveness means that unless one of the values  $p_i$  is equal to 1 and other to 0, we have a strict inequality:

$$\sum_{i=1}^n p_i \cdot f(r_i) < f\left(\sum_{i=1}^n p_i \cdot r_i\right).$$

In particular, for our case, when  $f(r) = \sqrt{r}$  and when some probabilities are different from 0 and 1, we get

$$\sum_{i=1}^n p_i \cdot \sqrt{r_i} < \sqrt{\sum_{i=1}^n p_i \cdot r_i}. \tag{2}$$

As we have mentioned, the folks organizing the lottery are not willing to lose money, so the average gain must be smaller than or equal to the price  $t$  of the lottery ticket:

$$\sum_{i=1}^n p_i \cdot r_i \leq t. \tag{3}$$

By taking square root of both sides of this inequality, we conclude that:

$$\sqrt{\sum_{i=1}^n p_i \cdot r_i} \leq \sqrt{t}. \tag{4}$$

Combining (2) and (4), we conclude that

$$\sum_{i=1}^n p_i \cdot \sqrt{r_i} < \sqrt{t},$$

i.e., that the condition (1) is never satisfied—and thus, that rational people should not play lotteries.

How can we explain that they not only play lotteries once in a while, but that many folks even become addicted to playing them?

**Let us use an even more adequate model.** The fact that the above model does not always explain human behavior means that we need to consider an even more adequate model of human behavior—a model that would take into account some additional features of human behavior.

One possibility for providing such more adequate model comes from the fact that the above model (implicitly) assumes that people adequately estimate the probabilities of different events. In reality, people tend to overestimate small probabilities. This phenomenon is described, e.g., in [4]. In [6, 7], provide a possible theoretical explanation for this phenomenon. Based on this explanation, provide a formula relating a subjective probability  $ps$  of an event—i.e., the values that people use to make decisions—and the actual probability  $p$ :  $ps = \frac{2}{\pi} \cdot \arcsin(\sqrt{p})$ . For small value  $p$ , this means  $ps = c_p \cdot \sqrt{p}$  for some constant  $c_p$ .

Thus, when making decisions, people maximize the expression

$$\sum_{i=1}^n p s_i \cdot u(r_i) = c_p \cdot \sum_{i=1}^n \sqrt{p_i} \cdot \sqrt{r_i} = c_p \cdot \sum_{i=1}^n \sqrt{p_i \cdot r_i}.$$

So, they play the lottery if

$$c_p \cdot \sum_{i=1}^n \sqrt{p_i \cdot u_i} > c_p \cdot \sqrt{t},$$

i.e., equivalently, if

$$\sum_{i=1}^n \sqrt{p_i \cdot u_i} > \sqrt{t}. \quad (5)$$

For a simple lottery, this means that  $\sqrt{p \cdot r} > \sqrt{t}$ . Since for a simple lottery, we must have  $p \cdot r \leq t$ —otherwise the lottery organizers will be losing money—the inequality  $\sqrt{p \cdot r} > \sqrt{t}$  is not possible.

This explains why simple lotteries are not addictive.

**What about more complex lotteries.** For an above-described more complex lottery, with two levels of rewards, when  $p_\ell \cdot R + p_m \cdot r \approx t$ , we have

$$(\sqrt{p_\ell \cdot R} + \sqrt{p_m \cdot r})^2 = p_\ell \cdot R + p_m \cdot r + 2\sqrt{(p_\ell \cdot R) \cdot (p_m \cdot r)} > t.$$

So, in this case,  $\sqrt{p_\ell \cdot R} + \sqrt{p_m \cdot r} > \sqrt{t}$ .

A similar inequality holds if we consider three or more different reward levels. This explains why more complex lotteries are addictive.

**Acknowledgements** This work was supported in part by the National Science Foundation grants:

- 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and
- HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported by the AT&T Fellowship in Information Technology, and by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

The authors are thankful to all the participants of the 26th UTEP/NMSU Workshop on Mathematics, Computer Science, and Computational Science (El Paso, Texas, November 6, 2021) for valuable discussions.

## References

1. Eyal, N., Hoover, R.: *Hooked: How to Build Habit-Forming Products*. Penguin, New York (2014)
2. Fishburn, P.C.: *Utility Theory for Decision Making*. Wiley, New York (1969)
3. Fishburn, P.C.: *Nonlinear Preference and Utility Theory*. The John Hopkins Press, Baltimore, Maryland (1988)
4. Kahneman, D.: *Thinking, Fast and Slow*. Farrar, Straus, and Giroux, New York (2011)
5. Kreinovich, V.: Decision making under interval uncertainty (and beyond). In: Guo, P., Pedrycz, W. (eds.) *Human-Centric Decision-Making Models for Social Sciences*, pp. 163–193. Springer, Berlin (2014)
6. Lorkowski, J., Kreinovich, V.: Fuzzy logic ideas can help in explaining Kahneman and Tversky's empirical decision weights. In: Zadeh, L., et al. (eds.) *Recent Developments and New Direction in Soft-Computing Foundations and Applications*, pp. 89–98. Springer, Berlin (2016)
7. Lorkowski, J., Kreinovich, V.: *Bounded Rationality in Decision Making Under Uncertainty: Towards Optimal Granularity*. Springer, Cham, Switzerland (2018)
8. Luce, R.D., Raiffa, R.: *Games and Decisions: Introduction and Critical Survey*. Dover, New York (1989)
9. Nguyen, H.T., Kosheleva, O., Kreinovich, V.: Decision making beyond Arrow's 'impossibility theorem', with the analysis of effects of collusion and mutual attraction. *Int. J. Intell. Syst.* **24**(1), 27–47 (2009)
10. Nguyen, H.T., Kreinovich, V., Wu, B., Xiang, G.: *Computing Statistics under Interval and Fuzzy Uncertainty*. Springer, Berlin (2012)
11. Raiffa, H.: *Decision Analysis*. McGraw-Hill, Columbus, Ohio (1997)

# Why Decimal System? Why Communities with More Than 150 Folks Tend to Split? New Consequences of the Seven Plus Minus Two Law



Leonardo Orea Amador and Vladik Kreinovich

**Abstract** Why are we using the decimal system to describe numbers? Why all over the world, communities with more than 150 folks tend to split? In this paper, we show that both phenomena—as well as some other phenomena—can be explained if we take into account the seven plus minus two law, according to which a person can keep in immediate memory from 5 to 9 items.

## 1 Formulation of the Problem

In this paper, we consider two seemingly unrelated questions, and we show that, somewhat surprisingly, they seem to have a common explanation.

**Why decimal system?** We currently use decimal system, but why? To us now, this may seem natural, but in the past, many different systems were used:

- Babylonians used 60-based system; see, e.g. [1, 3, 10, 11];
- Mayans used 20-based system; see, e.g., [1, 3, 8–11], etc.

For some reason, only the decimal system survived—why? A usual argument is that it is related to the fact that we have ten fingers on two hands, but the same logic explained 5-based system—when counting on one hand, or 12-based system—when we use knuckles instead of fingers.

**Why communities with more than 150 folks tend to split.** There is a sociological phenomenon that communities that have more than 150 folks tend to split into sub-communities; see, e.g., [5–7]. This magic number 150 was observed in many different cultures:

---

L. Orea Amador · V. Kreinovich (✉)

Department of Computer Science, University of Texas at El Paso, 500 W. University El Paso, El Paso, TX 79968, USA

e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

L. Orea Amador

e-mail: [lorea@miners.utep.edu](mailto:lorea@miners.utep.edu)

- Researchers looked at dozens of hunter-gatherer societies for which we have historical evidence, including Walbiri from Australia, Tauade from New Guinea, Ammassalik from Greenland, and Ona from Tierra del Fuego, Argentina. In all these societies, the average number people in a village is about 150.
- Throughout centuries, it was confirmed again and again that the largest size of a functional military fighting unit is about 150. In the beginning, this might have been a limitation on communication, but nowadays that communication is easy, this limitation remains—it is the limit on the ability of a large group of people to coordinate their actions and to successfully act together.
- This magic number is confirmed by the experience of Hutterites, a religious group similar to Amish and Mennonites, that when a community becomes larger than 150, it loses coherence and needs to be split into several smaller groups.
- Several manufacturing companies found out that 150 is the largest size for a successful coherent unit.

Why?

A possible explanation is that this is how our brain is set up—to be able to only handle communications with groups not exceeding 150 folks. But this leaves another question: why is our brain set up this way? How can we explain it by using some other well-studied phenomena?

## 2 Seven Plus Minus Two Law: A Brief Reminder

To explain both phenomena, we will use the *seven plus minus two law*, according to which a person can keep in immediate memory from 5 to 9 items. Similarly, when we classify things, we divide them into 5 to 9 groups. For some folks, it is 5, for some, it is 9; see, e.g., [13, 15] (see also [2, 17]).

Let us show how this law can explain both phenomena.

## 3 Why Decimal System: A Possible Explanation

Why do we use 10-based arithmetic? Up to nine objects some of us can keep in mind. Ten is the smallest number of objects which cannot be immediately remembered. We thus need to keep track of it, no matter how many smart people we use—we need to write it down.

This explains why 10 is, at present, the usual base for representing numbers.

## 4 Why Communities with More Than 150 Folks Tend to Split: A Possible Explanation

**Zipf law.** To explain this phenomenon, we need to also use another known law: Zipf law. This law was first discovered in linguistics: if we sort words by frequency and denote the frequency of the most frequent word by  $f$ , then:

- the second frequent word has frequency  $f/2$ ,
- the third frequent word has frequency  $f/3$ , etc.

Similar dependence has been observed for many phenomena—such as distribution of wealth, etc.; see, e.g., [4, 12].

**Let us apply Zipf law to our situation.** In a big group, we cannot pay equal attention to all the members of the group. So, we pay more attention to some folks, less attention to others. It is natural to expect that the same Zipf law will be applicable here: that if for each person  $P$ , we sort members of the group in the decreasing order of  $P$ 's attention, then:

- the person most involved with him/her receives full attention,
- the next involved requires  $1/2$  of the attention,
- then  $1/3$ , etc.

Overall, in a group of  $n$  folks, we get  $1 + 1/2 + \dots + 1/n$  full attentions. It is known that this sum is approximately equal to  $\ln(n)$ .

When  $n$  gets larger than approximately 150, this sum exceeds 5—so for some people, involvement with everyone in the community becomes impossible, and the community naturally splits.

## 5 Two Additional Related Phenomena

The same seven plus minus two law can explain other phenomena as well.

**First phenomenon: compensation recommended by the Bible.** According to Chap. 5 of the Book of Numbers, if a person is found guilty:

- this person should not only pay the damages,
- the guilty party should also pay an additional  $1/5$  of the damage amount as a compensation.

**How this can be explained.** To prevent people from wrongdoing, a natural idea is to institute an additional penalty rather than to require simply to return back what was wrongly taken. So, this additional penalty should not be insignificant, it should be felt by the guilty party.



On the other hand, the general trend in the Bible—although there are exceptions—is that a penalty should fit the crime, and thus, that it should not be too harsh. Thus, the additional penalty it should not be too excessive. In other words, this additional penalty should be equal to the smallest amount which will be felt by the guilty party.

What is this minimal-felt amount? In line with the general seven plus minus two law, depending on a person, this amount is between  $1/5$  and  $1/9$  of the whole. If we make it  $1/9$ —or any amount smaller than  $1/5$ —this penalty will not be felt by those for whom this number is 5. So, the smallest amount which is felt by all possible wrongdoers is  $1/5$ —and this is exactly what the Bible recommends.

**Second phenomenon: rating the dates.** Another interesting phenomenon is that women on dating sites rate 85% of men as below average in attractiveness; see, e.g., [14, 16].

**How this can be explained.** According to [14, 16], women want the best partners. Since, on average, they divide possible partners into 7 groups, this group of “best” partners is indeed, on average,  $1/7$  of all potential dates.

Thus, indeed, on average  $1 - 1/7 = 6/7 \approx 85\%$  of possible partners are dismissed as not good.

**Acknowledgements** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

## References

1. Bokati, L., Kosheleva, O., Kreinovich, V.: How can we explain different number systems? In: Ceberio, M., Kreinovich, V. (eds.) *How Uncertainty-Related Ideas Can Provide Theoretical Explanation for Empirical Dependencies*, pp. 21–26. Springer, Cham, Switzerland (2021)
2. Bokati, L., Kreinovich, V., Katz, J.: Why 7 plus minus 2? A possible geometric explanation. *Geoinformatics* **30**(1), 109–112 (2021)
3. Boyer, C.B., Merzbach, U.C.: *A History of Mathematics*. Wiley, New York (1991)
4. Cervantes, D., Kosheleva, O., Kreinovich, V.: Why Zipf’s law: a symmetry-based explanation. *International Mathematical Forum* **13**(6), 255–258 (2018)
5. Dunbar, R.I.M.: Neocortex size as a constraint on group size in primates. *J. Hum. Evol.* **20**, 469–493 (1992)
6. Dunbar, R.I.M.: *Grooming, Gossip, and the Evolution of Language*. Harvard University Press, Harvard (1996)
7. Gladwell, M.: *The Tipping Point: How Little Things Can Make a Big Difference*. Black Bay Books/Little, Brown, and Company, New York (2002)
8. Heath, T.L.: *A Manual of Greek Mathematics*. Dover, New York (2003)
9. Ifrah, G.: *The Universal History of Numbers: From Prehistory to the Invention of the Computer*. Wiley, Hoboken (2000)
10. Kosheleva, O.: Mayan and Babylonian arithmetics can be explained by the need to minimize computations. *Appl. Math. Sci.* **6**(15), 697–705 (2012)

11. Kosheleva, O., Villaverde, K.: *How Interval and Fuzzy Techniques Can Improve Teaching*. Springer, Cham, Switzerland (2018)
12. Mandelbrot, B.: *The Fractal Geometry of Nature*. Freeman, San Francisco (1983)
13. Miller, G.A.: The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* **63**(2), 81–97 (1956)
14. Peterson, J.B.: *12 Rules for Life: An Antidote for Chaos*. Random House Canada, Toronto, Ontario (2018)
15. Reed, S.K.: *Cognition: Theories and Application*. Wadsworth Cengage Learning, Belmont (2010)
16. Rudder, C.: *Dataclysm: Love, Sex, Race, and Identity*. Broadway Books, New York (2015)
17. Trejo, R., Kreinovich, V., Goodman, I.R., Martinez, J., Gonzalez, R.: A realistic (non-associative) logic and a possible explanations of  $7 \pm 2$  law. *Int. J. Approx. Reason.* **29**, 235–266 (2002)

# Lev Landau's Marital Advice Explained



Olga Kosheleva and Vladik Kreinovich

**Abstract** Nobelist physicist Lev Landau was known for applying mathematical and physical reasoning to human relations. His advices may have been somewhat controversial, but they were usually well motivated. However, there was one advice for which no explanation remains—that a person should not marry his/her first and second true loves, and only start thinking about marriage starting with the third true love. In this paper, we provide a possible Landau-style motivation for this advice.

## 1 Formulation of the Problem

**Who was Lev Landau.** Lev Landau was a Nobelist physicist.

In addition to his physics discoveries—and to a popular physics textbook he co-authored [3]—he was also well known for applying reasoning from mathematics and physics to human relations.

**Sometimes, his advice made perfect sense.** In some cases, Landau's advice about human relations made perfect sense—and if the audience did not understand the reason for this advice, he was always ready to provide reasonably convincing explanations.

For example, he claimed that there is an optimal distance at which a beautiful woman's face is the most beautiful. A similar statement about enjoying paintings is a known fact—for each painting, there is usually an optimal distance at which this painting looks the best. Because of this phenomenon, in an art museum, true connoisseurs follow a strange-looking trajectory: e.g., staying closer to smaller paint-

---

O. Kosheleva

Department of Teacher Education, University of Texas at El Paso, 500 W.University El Paso,  
Austin, TX 79968, USA  
e-mail: [olgak@utep.edu](mailto:olgak@utep.edu)

V. Kreinovich (✉)

Department of Computer Science, University of Texas at El Paso, 500 W.University El Paso,  
Austin, TX 79968, USA  
e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

ings while moving further away when a painting is larger, staying further away from impressionist painting but closer to realistic ones, etc.

However, for women's faces, the same conclusion sounded unusual. Landau's explanation for this advice was very simple:

- when you too far away, you cannot see anything but a blur, and
- when you are too close, you only see one feature and not the whole face, and thus, you cannot appreciate the full beauty.

Thus, there must be a distance at which the beauty is the most visible.

**Strange marital advice.** While most of Landau's non-physics advices were explained—usually, by Landau himself—one of Landau's advices remains unexplained. The advice was very straightforward—although shocking at the time when the usual advice was marrying your true love and live happily ever after. The advice was *not* to marry your first true love, and *not* to marry your second true love, and only start considering marriage starting with the third true love; see, e.g., [7].

A possible reason why this advice remains unexplained is that he gave this advice to his teenage niece. She was so shocked by this advice that she did not even ask for the reason. As she writes in her memoirs, she even pretended to experience her first true love with some imaginary person, so that her uncle would be happy that she followed his advice.

**What is the reason for this advice?** Knowing Landau, he was a very rational person: whatever he said was usually well justified. So this unusual marital advice puzzled us for some time.

Now we finally came up with a reasonable explanation, an explanation that we are describing in this paper.

## 2 Our Explanation

**Main idea.** Every person has some criteria—formal or informal—for selecting a spouse. The person wants to select someone who is the best according to this criterion. This sounds straightforward, but the problem is that to really understand the person, to check compatibility, one needs to get close to this person, spend some time with him/her. It is usually not possible to try it with several people at the same time, but if you spend some time with one person, and then decide to try someone else—that first person whom you ditched will be, in general, reluctant to re-start the relation.

So, we encounter a known problem known as a secretary problem, or as a fussy princess problem; see, e.g., [2, 4, 8]. Let us describe this problem in its princess form.

**Fuzzy princess problem: formulation.** Suppose that  $n$  princes seek the hand of a beautiful princess. They come to her and propose one by one. She needs to select the best prince to marry.

In the ideal-for-the-princess world, she would consider all of them, and then select the one that she likes the best (or, if she is a patriotic princess, the one that will bring the most beneficial alliance to her country). But the problem is that once a prince makes a proposal, he expects an immediate (or almost immediate) answer. If this answer is No, the prince's pride does not allow him to come back if the princess changes her mind.

Provided that the princes arrive in random order, what is the best strategy for the princess that will, on average, leads to the best possible choice?

**Fuzzy princess problem: solution.** The solution to this problem is known, and it is somewhat non-intuitive: for reasonably large values  $n$  (and actually already for moderate values  $n$ ), the best strategy is:

- to say No to the first  $n/e$  suitors, and then
- to select the first one who is better than the first  $n/e$  candidates (and if none is better, and if there is a need for a princess to marry, marry the last one).

**Conclusion for Landau's advice.** From this viewpoint, if we know how many true loves a person will encounter in her (or his) life, then a reasonable idea is indeed:

- to just enjoy the first  $n/e$  true loves (without thinking of marriage) and
- to only start considering marriage after that.

But how can we know this number  $n$ ?

**Where do we get the number  $n$ .** A princess may select one of the hundred princes whom she meets for the first time, this is how princesses do it in fairy tales.

For example, she may be motivated by the desire to bring a good alliance to her country. After all, at some point, the notorious Russian tsar Ivan the Terrible made a marriage proposal to none else but the great British Queen Elizabeth—not because he was in love with her (they never met, and I am not sure if even ever saw a picture of her), but because he believed—and there was some reason for that—that by combining their empires, they could easily defeat their enemies; see, e.g., [1].

With us common folks, the situation is different. We want a spouse that will be an important part of our lives, we do not usually want to marry an unknown stranger. So, in contrast to the princess, we only want to marry someone whom we know very well. And how many people can we know well?

In psychology, there is a known “several plus minus two” law, according to which a person can only simultaneously be seriously thinking about several plus minus two objects, i.e., between five and nine, on average seven (how many is different for different individuals); see, e.g., [5, 6].

Realistically, when you have been close to a person for some time, when he or she was your true love, the memories of that person stays in your heart forever. So, during the lifetime, we can only have seven plus minus two true loves—on average, seven. In other words,  $n = 7 \pm 2$ .

**This explains Landau's marital advice.** For  $n = 7$ , the value  $n/e$  is between 2 and 3. Moreover, if we dismiss the rarer extreme cases  $n = 5$  and  $n = 9$ , for all three intermediate values  $n = 6$ ,  $n = 7$ , and  $n = 8$ , the ratio  $n/e$  is between 2 and 3.

For all these values  $n$ , the advice to start thinking of marriage only after  $n/e$  true loves means indeed to start thinking about marriage only starting with the third true love. (And actually, the value  $n = 9$  for which  $n/e$  is between 3 and 4 also kinds of fits into the same advice, since here also we skip the first two true loves—the only difference is that for  $n = 9$ , we skip the third true love as well.)

So, we have indeed explained Landau's advice.

**Acknowledgements** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

## References

1. Bertolet, A.R.: The Tsar and the Queen: 'You speak a language that I understand not'. In: Beem, C. (ed.) *The Foreign Relations of Elizabeth I*, pp. 101–123. *Queenship and Power*, Palgrave Macmillan, New York (2011)
2. Bruss, F.T., Cam, L.L. (eds.): *Game Theory, Optimal Stopping, Probability & Statistics: Paper in Honor of Thomas S. Ferguson*, Institute of Mathematical Statistics, Cleveland, Ohio (2000)
3. Landau, L.D., Lifshitz, E.M.: *Course of Theoretical Physics, Vols. 1–9*, Butterworth-Heinemann, Oxford, UK (1976)
4. Leonardz, B.: *To Stop Or Not to Stop: Some Elementary Optimal Stopping Problems with Economic Interpretation*. Almqvist & Wiksell, Stockholm (1973)
5. Miller, G.A.: The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* **63**(2), 81–97 (1956)
6. Reed, S.K.: *Cognition: Theories and application*. Wadsworth Cengage Learning, Belmont, California (2010)
7. Ryndina, E.: Lev Landau: shtrikhi k portretu. *Vestnik*, Vol. 16, No. 6, pp. 16–20 (in Russian) (2004). <http://www.vestnik.com/issues/2004/0317/win/ryndina.htm>
8. Wong, D.: *Generalized Optimal Stopping Problems and Financial Markets*. CRC Press, Boca Raton, Florida (1997)

# Why Too Much Interaction Between Different Parts of the Brain Leads To Unhappiness



Ricardo Alvarez, Yamel Hernandez, and Vladik Kreinovich

**Abstract** Reasonably recent experiments show that unhappiness is strongly correlated with the excessive interaction between two parts of the brain—amygdala and hippocampus. At first glance, in situations when outside signals are positive, additional interaction between two parts of the brain that get signals from different sensors should only reinforce the positive feeling. In this paper, we provide a simple explanation of why, instead of the expected reinforcement, we observe unhappiness.

## 1 Formulation of the Problem

**General problem.** Sometimes, we are in a good mood, and sometimes, we are in a bad mood. In some cases, our mood is determined by the external circumstances, but sometimes, a person who has everything is still unhappy. How can we make people happier?

To be able to do this, it is important to understand what brain processes cause different moods. If we learn why people become unhappy, we may be able to help them become happier.

**This problem is very complex.** The brain is a very complex structure, with many processes happening at the same time. Because of this complexity, until recently, it was not clear which brain processes are correlated with mood.

This complexity is also affected by the fact that the usual ways of studying brain activities—via Magnetic Resonance or other remote methods—provide information about the average activity of reasonably large groups of neurons, and it looks like

---

R. Alvarez · Y. Hernandez · V. Kreinovich (✉)  
Department of Computer Science, University of Texas at El Paso, 500 W. University El Paso,  
Austin, TX 79968, USA  
e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

R. Alvarez  
e-mail: [ralvarezlo@miners.utep.edu](mailto:ralvarezlo@miners.utep.edu)

Y. Hernandez  
e-mail: [yeherandez2@miners.utep.edu](mailto:yeherandez2@miners.utep.edu)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023  
M. Ceberio and V. Kreinovich (eds.), *Decision Making Under Uncertainty  
and Constraints*, Studies in Systems, Decision and Control 217,  
[https://doi.org/10.1007/978-3-031-16415-6\\_32](https://doi.org/10.1007/978-3-031-16415-6_32)

this averaging filters out possible correlations—whose discovery probably requires more localized techniques.

**More localized measurements are available.** The possibility to use more localized techniques of brain study comes from the study of epileptic patients.

From the engineering viewpoint, epilepsy means excessive positive feedback, excessive amplification. Electric signals in the brain gets amplified—as they are amplified in all systems, to compensate for the natural signal decrease. The problem is that if this amplification is too high, the signal passing to and fro gets constantly amplified more and more—until it exceeds the safety limits and starts damaging the brain. This amplification is usually happening in one specific small part of the brain. To help the patient, it is therefore important to find the location of this area.

The brain is very important, we do not want to affect its functions, so we must pinpoint the defective area as accurately as possible. Such accuracy is often not possible if we only use non-invasive techniques. So, to help with this location, electrodes are implanted in several places in the suspected defect area, so that by measuring the corresponding brain activity, we will be able to very accurately pinpoint the defect area.

As a side effect, we also have a very localized description of brain activity.

**A recent breakthrough.** A recent (2018) study [3] analyzed this activity and found—for the first time—brain processes that clearly correlated with the person’s mood. Namely, it turned out that the person’s mood is determined by the interaction between two specific parts of the brain: amygdala and hippocampus.

**Unexpected feature.** The researchers expected to see how—directly or indirectly—signals related to external factors affect the person’s mood. This was indeed found. However, there was also an unexpected discovery—that for the same level of external signals, the mood was strongly affected by the degree of interaction between the above two parts of the brain: too much interaction between these two different parts of the brain leads to unhappiness.

**Why?** A natural question is: why? In this paper, we provide a possible explanation for this unexpected empirical phenomenon.

## 2 Analysis of the Problem and Resulting Explanation

**Let us describe a simple mathematical model.** Each part of the brain receives signals from our sensors, from different parts of the body, etc. Some of these signals are good—so they should lead to more happiness. Let us denote the overall level of positivity of signals coming to amygdala by  $p_1$ , and of the signals coming to hippocampus by  $p_2$ .

In general, different parts of the brain process different signals. The overall mood should depend on all these signals. So, it makes sense that there are interactions



between different parts of the brain—that enable us to combine these signals and thus, get the signal reflecting the overall mood.

Interaction means that the signal coming from one part of the brain affects the signal in the other part. Thus, the overall positivity level  $s_1$  at the amygdala is determined not only by the signals  $p_1$  coming to it from the corresponding sensors, but also by the signals coming to it from the hippocampus. The higher the activity level  $s_2$  at the hippocampus, the more signals come to amygdala, so we can say that

$$s_1 = p_1 + k_{12} \cdot s_2, \quad (1)$$

where the coefficient  $k_{12}$  describes the degree of interaction between these two parts of the brain.

Similarly, the resulting activity level  $s_2$  of the hippocampus is determined not only by the signals  $p_2$  coming to it from the corresponding sensors, but also by the signals coming to it from the amygdala. The higher the activity level  $s_1$  at the amygdala, the more signals come to hippocampus, so we can say that

$$s_2 = p_2 + k_{21} \cdot s_1, \quad (2)$$

where the coefficient  $k_{21}$  describes the degree of interaction between these two parts of the brain.

**At first glance, this cannot explain the empirical fact.** At first glance, it looks like our model (1)–(2) cannot explain the observed effect: based on the equations (1) and (2), the larger the degree of interaction between the two corresponding parts of the brain, the more positive will be the overall sense of happiness.

**A deeper analysis leads to the desired explanation.** Let us show, however, the deeper analysis of our model leads to the desired explanation.

Indeed, if we plug in the right-hand side of the formula (1) instead of  $s_1$  in the formula (2), we conclude that

$$s_2 = p_2 + k_{21} \cdot (p_1 + k_{12} \cdot s_2) = p_2 + k_{21} \cdot p_1 + k_{21} \cdot k_{12} \cdot s_2. \quad (3)$$

Moving all the terms containing the unknown  $s_2$  into the left-hand side, we conclude that

$$(1 - k_{21} \cdot k_{12}) \cdot s_2 = p_2 + k_{21} \cdot p_1, \quad (4)$$

hence

$$s_2 = \frac{p_2 + k_{21} \cdot p_1}{1 - k_{21} \cdot k_{12}}. \quad (5)$$

Similarly, if we plug in the right-hand side of the formula (2) instead of  $s_2$  in the formula (1), we conclude that

$$s_1 = p_1 + k_{12} \cdot (p_2 + k_{21} \cdot s_1) = p_1 + k_{12} \cdot p_2 + k_{12} \cdot k_{21} \cdot s_1. \quad (6)$$

Moving all the terms containing the unknown  $s_1$  into the left-hand side, we conclude that

$$(1 - k_{12} \cdot k_{21}) \cdot s_1 = p_1 + k_{12} \cdot p_2, \quad (7)$$

hence

$$s_1 = \frac{p_1 + k_{12} \cdot p_2}{1 - k_{12} \cdot k_{21}}. \quad (8)$$

When the interaction becomes too intensive, namely, when  $k_{12} \cdot k_{21} > 1$ , then even when the signals  $p_1$  and  $p_2$  are positive, the resulting states  $s_1$  and  $s_2$ —as described by the formulas (5) and (8)—become negative, i.e., indeed corresponding to unhappiness.

This explains the above-described empirical phenomenon.

**Comment.** From the mathematical viewpoint, this explanation is similar to a known explanation of another phenomenon—that an excess of empathy may lead to unhappiness—see, e.g., [1, 2, 4–6].

**Acknowledgements** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

## References

1. T. Bergstrom, “Love and spaghetti, the opportunity cost of virtue”, *Journal of Economic Perspectives*, 1989, Vol. 3, No., pp. 165–173
2. Bergstrom, T.: Systems of benevolent utility interdependence. University of Michigan, Technical Report (1991)
3. Kirkby, L.A., Luongo, F.J., Lee, M.B., Nahum, M., Van Veleet, T.M., Dawes, H.E., Chang, E.F., Sohal, V.S.: An amygdala-hippocampus subnetwork that encodes variation in human mood. *Cell* **175**, 1688–1700 (2018)
4. V. Kreinovich, *Paradoxes of Love: Game-Theoretic Explanation*, University of Texas at El Paso, Department of Computer Science, Technical Report UTEP-CS-90-16, July 1990
5. Nguyen, H.T., Kosheleva, O., Kreinovich, V.: “Decision making beyond Arrow’s ‘impossibility theorem’, with the analysis of effects of collusion and mutual attraction”. *International Journal of Intelligent Systems* **24**(1), 27–47 (2009)
6. Nguyen, H.T., Kreinovich, V., Wu, B., Xiang, G.: *Computing Statistics under Interval and Fuzzy Uncertainty*. Springer Verlag, Berlin, Heidelberg (2012)

# **Applications to Religion**

# Gödel's Proof of Existence of God Revisited



Olga Kosheleva and Vladik Kreinovich

**Abstract** In his unpublished paper, the famous logician Kurt Gödel provided arguments in favor of the existence of God. These arguments are presented in a very formal way, which makes them difficult to understand to many interested readers. In this paper, we describe a simplifying modification of Gödel's proof which will hopefully make it easier to understand. We also describe, in clear terms, why Gödel's arguments are just that—arguments—and not a convincing proof.

## 1 Formulation of the Problem

**What Gödel did.** In his originally unpublished paper, the famous logician Kurt Gödel provides arguments in favor of the existence of an object that can be interpreted as God; see [5], see also [1–4, 6–9].

**Problems with the original Gödel's proof.** Gödel's proof is somewhat over-complicated and, as a result, somewhat difficult to understand. It is therefore desirable to come up with a simplified version of this proof.

The fact that this proof is presented in a complicated way also makes it difficult to understand whether Gödel's arguments are simply arguments or a convincing proof.

**What we do.** In this paper, we provide a modified (namely, simplified) version of Gödel's proof. This simplification, hopefully, makes it easier to understand the proof itself—and also to understand why this is not a fully convincing proof.

---

O. Kosheleva

Department of Teacher Education, University of Texas at El Paso, 500 W. University El Paso, Austin, TX 79968, USA

e-mail: [olgak@utep.edu](mailto:olgak@utep.edu)

V. Kreinovich (✉)

Department of Computer Science, University of Texas at El Paso, 500 W. University El Paso, Austin, TX 79968, USA

e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

217

M. Ceberio and V. Kreinovich (eds.), *Decision Making Under Uncertainty*

*and Constraints*, Studies in Systems, Decision and Control 217,

[https://doi.org/10.1007/978-3-031-16415-6\\_33](https://doi.org/10.1007/978-3-031-16415-6_33)

## 2 Intuitive Idea

**Idea.** Intuitively, God is an object that has all good properties and no bad properties.

**We need to formalize this idea.** Let us show how to formalize this idea.

## 3 Towards Formalizing This Idea

**Possible worlds.** Our knowledge about the world is incomplete. This means that we usually do not have full information about the world. Even if we have a reasonably full information about the current state of the world, we may not be sure about its future state. Thus, there are several possible descriptions of the world which are all consistent with our knowledge. Such descriptions are usually called *possible worlds*.

If a statement  $S$  is true in all possible worlds, we say that it is *necessarily true* and denote it by  $\Box S$ . If a statement holds in at least one of the possible worlds, then we say that this statement is *possibly true* and denote it by  $\Diamond S$ .

**Good and bad properties.** In each possible world, there are objects  $x, x'$ , etc. that may have different properties  $\varphi, \psi$ , etc. Some properties are good; we will denote this by  $g(\varphi)$ . Other properties are bad; we will denote this by  $b(\varphi)$ .

It is reasonable to assume that goodness and badness are absolute—if a property is good in one world, it is good in every world—same for bad properties.

Intuitively, if a property is good, then this property cannot be bad, and its negation cannot be good:

$$g(\varphi) \rightarrow \neg b(\varphi) \text{ and } g(\varphi) \rightarrow \neg g(\neg\varphi). \quad (1)$$

Similarly, if a property is bad, then this property cannot be good, and its negation cannot be bad:

$$b(\varphi) \rightarrow \neg g(\varphi) \text{ and } b(\varphi) \rightarrow \neg b(\neg\varphi) \quad (2)$$

**Formal implication versus meaningful implication.** An important part of our knowledge are if-then statements—known as *implications*.

In mathematics, a statement “if  $A$  then  $B$ ” is denoted by  $A \rightarrow B$ . The general meaning of such a statement in mathematics is that if  $A$  is true, then  $B$  is true too. If  $A$  is false, then the implication has no limitation on  $B$ , so the statement  $A \rightarrow B$  is true. If  $A$  is true, then  $B$  should be true. Thus,  $A \rightarrow B$  means that either  $A$  is false or  $B$  is true.

This sounds reasonable at first glance, but it leads to meaningless implications. For example, if it will not rain tomorrow in El Paso and a volcano Erebus in Antarctica will be active, then the implication “if it will rain tomorrow in El Paso, then Erebus will be inactive” is, in mathematical sense, true.

While this implication is mathematically true, from the commonsense viewpoint, it is meaningless. Indeed, intuitively, “if  $A$  then  $B$ ” means that if we make  $A$  true,

then  $B$  also becomes true. However, if we force rain to fall in El Paso—e.g., by seeding the clouds—if will not affect the Erebus volcano.

An intuitive meaning of a natural-language if-then statement is that once we make  $A$  true,  $B$  will always be true, i.e., that the implication  $A \rightarrow B$  should be true in *all* possible worlds—and not just in our world as in the usual mathematical definition. This leads to the following formula:

$$\varphi \Rightarrow \psi \stackrel{\text{def}}{=} \Box(\forall x(\varphi(x) \rightarrow \psi(x))) \quad (3)$$

If this formula holds, we will say that  $\varphi$  *necessarily implies*  $\psi$ .

**Relation between good, bad, and meaningful implication.** Intuitively, if  $\varphi$  is a good property, and  $\varphi$  necessarily implies  $\psi$ , then the property  $\psi$  should also be good:

$$(g(\varphi) \ \& \ (\varphi \Rightarrow \psi)) \rightarrow g(\psi) \quad (4)$$

Similarly, if  $\varphi$  is a bad property, and  $\varphi$  necessarily implies  $\psi$ , then the property  $\psi$  should also be bad:

$$(b(\varphi) \ \& \ (\varphi \Rightarrow \psi)) \rightarrow b(\psi). \quad (5)$$

## 4 First (Preliminary) Result

**Formulation of the result.** The first result—proven by Gödel—is that for every good property  $\varphi$ , in some possible world, there is an object that satisfies this property.

**Proof.** Indeed, let us assume that this statement is not true. This means that in each world, for each object  $x$ , the statement  $\varphi(x)$  is false. By definition of the usual implication, this means, in particular, that in every world, for every object  $x$ , we have  $\varphi(x) \rightarrow \neg\varphi(x)$ . By definition of necessary implication, this means that  $\varphi \Rightarrow \neg\varphi$ . Since the property  $\varphi$  is good, by formula (4), it implies that its negation  $\neg\varphi$  is also good—but, according to formula (1), if a property is good, its negation cannot be good.

This contradiction shows that our assumption cannot be true. Thus, there must exist a possible world in which some object  $x$  satisfies the property  $\varphi$ .

## 5 It Is Good to Have Good Objects in All Possible Worlds

**Here we are modifying (and simplifying) Gödel's proof.** Up to now, we were following Gödel, but now, we will modify and simplify his arguments.

**Idea.** It would be nice to have objects satisfying a good property in all possible worlds.

**How to formalize this idea.** The above condition can be described as follows:

$$c(x) \stackrel{\text{def}}{=} \forall \varphi ((g(\varphi) \& \diamond \exists y \varphi(y)) \rightarrow (\Box \exists z \varphi(z))). \quad (6)$$

*Comment.* The property  $c(x)$  actually does not depend on  $x$ , the variable  $x$  is added solely for mathematical convenience.

**This property is clearly good.** The condition (6) is clearly good:

$$g(c). \quad (7)$$

## 6 Second Result

**Formulation of the result.** The second result is that for every good property  $\varphi$ , in every possible world, there is an object that satisfies this property.

**Proof.** Indeed, since the property  $c$  is good, according to the first result, it is true in some possible world. Thus, in some world, we have an implication

$$\forall \varphi ((g(\varphi) \& \diamond \exists y \varphi(y)) \rightarrow \Box \exists z \varphi(z)). \quad (8)$$

This implication does not depend on the world, thus it is just true.

According to the same first result, for every good property  $\varphi$ , there exists a world in which this property is true for some object, thus  $\diamond \exists y \varphi(y)$ . Thus, due to the implication (8), we conclude that  $\Box \exists z \varphi(z)$ , i.e., that the object satisfying this property indeed exists in all possible worlds.

## 7 What Is God?

**Now we can formalize the informal definition of God.** We want to say that an object  $x$  is God (we will denote it by  $G(x)$ ) is  $x$  has all good properties and no bad properties:

$$G(x) \stackrel{\text{def}}{=} \forall \varphi ((g(\varphi) \rightarrow \varphi(x)) \& (b(\varphi) \rightarrow \neg \varphi(x))). \quad (9)$$

**God is good.** Intuitively, being God is a good property:

$$g(G). \quad (10)$$

**Conclusion.** From the second result, we conclude that God exists in every possible world.

## 8 Word of Caution: Shall We All Run to a Place of Worship?

Does this result convincingly prove that God exist? Not necessarily.

The problem is in the definition of necessary implication. The way this notion is defined still enables us to make counterintuitive conclusions. Namely, if  $\psi(x)$  is always true, then the implication  $\forall x (\varphi(x) \rightarrow \psi(x))$  holds in all possible worlds, so, according to the above definition, we have  $\varphi \Rightarrow \psi$ .

For example, if  $\psi(x)$  is “the Sun will rise tomorrow”, then we get conclusion like “animal sacrifices necessarily imply that the Sun will rise tomorrow”. This is *not* what we intuitively mean by if-then rules, since whether the Sun rises or not clearly does not depend on whether we make an animal sacrifice or not.

A more adequate description is to only conclude that  $\varphi$  necessarily imply  $\psi$  when, in addition, it is possible that  $\psi$  will be false:

$$\varphi \Rightarrow \psi \stackrel{\text{def}}{=} \Box(\forall x (\varphi(x) \rightarrow \psi(x))) \ \& \ \Diamond \exists x \neg \psi(x). \quad (11)$$

However, if we use this more adequate definition in formula (4), we can now longer prove the very first Gödel's result—and thus, we are no longer able to conclude that God exists.

**Acknowledgements** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes). It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478. The authors are greatly thankful to Anatol Slissenko and to all the participants of the Fifth St. Petersburg Days of Logic and Computability Conference (St. Petersburg, Russia, May 12–15, 2021) for valuable discussions.

## References

1. Bjordal, F.A.: Understanding Gödel's ontological argumen. In: Childers, T. (Ed.), *The Logica Yearbook 1998*, Prague 1999, pp. 214–217
2. Bjordal, F.A.: All properties are divine, or God exists. *Logic and Logical Philosophy* **27**(3), 329–350 (2018)
3. Dawson, J.W., Jr.: *Logical Dilemmas: The Life and Work of Kurt Godel*. AK Peters, Wellesley, Massachusetts (1987)
4. Fitting, M.: *Types, Tableaus, and Godel's God*. Kluwer Academic, Dordrecht (2002)
5. Gödel, K.: In: S. Feferman, J.W. Dawson Jr., W. Goldfarb, C. Parsons, Solovay, R.M. (Eds.), *Unpublished Essays and Lectures, Collected Works III* (1st ed.), Oxford University Press, Oxford, UK, 1995, pp. 403–404 and 429–437
6. Hazen, A.P.: On Gödel's ontological proof. *Australasian J. Philos.* **76**(3), 361–377 (1998)
7. Oppy, G.: Ontological argument, In: Zalta, E.N. (Ed.), *Stanford Encyclopedia of Philosophy*, downloaded on May 14, 2021
8. Wang, H.: *Reflections on Kurt Gödel*. MIT Press, Cambridge, Massachusetts (1987)
9. Wang, H.: *A Logical Journey: from Gödel to Philosophy*. MIT Press, Cambridge, Massachusetts (1996)



# Blessings, God, Sacrifices: Possible Rational Explanations of Biblical Ideas



Olga Kosheleva and Vladik Kreinovich

**Abstract** In this paper, we show that many seemingly irrational Biblical ideas can actually be rationally interpreted: that God is everywhere, that we can only say what God is not, that God's name is holy, why cannot you bless as many people as you want, etc. We do not insist on our interpretations, there probably are many others, our sole objective was to show that many Biblical ideas can be rationally explained.

## 1 Formulation of the Problem

**Many Biblical ideas look irrational.** Many Biblical ideas sound irrational—at least at first glance.

**This can be expected: a religion cannot be fully rational.** Of course, religion, by definition, cannot be a fully rational enterprise, so some irrationality is natural.

**What we do in this paper.** However, what we plan to show in this paper is that many seemingly irrational ideas can have rational explanations.

**How we do it.** Some of our explanations come from common sense, some come from modern science—in which many seemingly counterintuitive ideas have been experimentally confirmed and have become a solid foundation of relativity and quantum physics; see, e.g., [3, 9].

*Comment.* We realize that from the theological viewpoint, our interpretations of Biblical ideas may be naive and oversimplified. This may be, but these interpretations

---

O. Kosheleva

Department of Teacher Education, University of Texas at El Paso, 500 W. University El Paso, Austin, TX 79968, USA

e-mail: [olgak@utep.edu](mailto:olgak@utep.edu)

V. Kreinovich (✉)

Department of Computer Science, University of Texas at El Paso, 500 W. University El Paso, Austin, TX 79968, USA

e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

do show that the Biblical ideas mentioned in this paper can be rationally explained—and this was exactly our objective.

## 2 Where Is God?

**Biblical idea.** While this may be not explicitly stated in the Bible, but the feeling one gets—and the feeling theologians get—is that, according to the Bible, God is (or at least can be) simultaneously at some specific place and at the same time in other places, probably even everywhere.

In the Reform Judaism prayerbook, this idea is described in the following poetic form: “Thou art as close to us as breathing and yet art farther than the farthestmost star” [1].

**At first glance, this sounds counterintuitive.** From the viewpoint of common sense, this is not rationally possible: if an object (or a person) is located in one place, the same object or person cannot be at the same time located at some other place.

**However, this is-everywhere idea is in perfect agreement with modern science.** In Newtonian physics, indeed, every particle was supposed to be limited to a specific location. Not so in quantum physics, where each particle is described by a so-called *wave function*  $\psi(x)$  that describes the probability with which this particle can be found at different locations. Specifically, for each spatial region  $S$ , the probability to find the particle in this region is equal to the integral  $\int_S |\psi(x)|^2 dx$  [3, 9].

The impossibility to exactly locate a particle is a part of Heisenberg’s uncertainty principle, one of the main principles of quantum physics. According to this principle, the more accurately we try to measure the particle’s location, the more momentum we should add to the particle, and this momentum brings the particle out of that location.

Moreover, according to quantum physics, a free particle in an empty space—which, in Newtonian physics, would just continue going in the same direction with the same speed—actually spreads out and its location becomes more and more blurred [3, 9].

In summary, from the viewpoint of modern physics, not having a specific location—i.e., in effect, being in different places at the same time—is actually a typical behavior of particle and, more generally, of all objects for which quantum effects cannot be ignored.

## 3 How Can We Characterize God?

**Maimonides’ interpretation of the Biblical idea.** According to the medieval theologian Maimonides [6], we cannot claim that God has any positive quality, God can only be characterized by negative qualities: God is not finite, God is not mortal, etc.

Interestingly, similar ideas were developed by Islamic theologians as well [10].

**How can we interpret this in rational terms?** Maimonides’s interpretation is somewhat similar to the belief of many physicists that no physical theory is perfect, that no matter what theory we propose, eventually there will be an experiment whose results would require some modification of this theory [3, 9]. In other words, whatever property the physical world satisfies—according to modern physics—this property is not universally true, be it the original Newton’s laws or the formulas of modern physics.

Thus, negative-qualities-only objects are actually very natural: the whole physical world is like that.

But how is this related to God? Interestingly (and somewhat unexpectedly), this seemingly natural physicists’ belief has an important computational consequence—that if we use observations, we can drastically speed up the solution to many computational problems—to the extent that we can solve many instances of NP-hard problems (provably most complex problems, see [5, 7]) in feasible time [4]. Solving hard problems is, in a nutshell, what creativity is about—at least creativity of scientists and engineers—as opposed to routine activity of applying known algorithms to easier problems. From this viewpoint, the negative-quality-only sequence of observations and experimental results serves as a source of creativity, which fits well with the idea of God as an important source of creativity.

## 4 Even God’s Name Is Holy: What Does This Mean?

**Biblical idea.** According to the Bible, not only God itself is holy, God’s name is holy as well. How can this be rationally interpreted?

A person performing good deeds can be holy, a place which helps to perform good deeds can be holy, an object used in performing these deeds can be holy, but a name sounds too abstract for that.

**Our interpretation.** Let us show that this idea can be rational too. As an analogy, instead of performing good deeds, let us consider spreading knowledge—which, by the way, is often necessary to be able to perform good deeds.

In the modern world, most of the knowledge we get is from published papers. For a paper, a natural analogue of its name is this paper’s title. And, of course, the title of the paper is often very informative by itself—moreover, e.g., in mathematics, often, the title of the paper describes the exact formulation of the statement proven in this paper, so unless one is interested in the proof itself, one does not even have to read the paper—all the needed information is in the title already.

In this example, the “name” (title) of the paper conveys the information—and thus, has the same property of conveying knowledge as the paper itself. It is therefore reasonable to expect that the very name of a person can similarly convey the same meaning of holiness as the person him/herself.

## 5 What Is a Blessing?

**Biblical ideas.** The Bible is full of stories related to blessings. We still use this word, but many places of the Bible shows that in the old days, this word had a different meaning. For example, in the modern interpretation, if you bless someone, this does not make it impossible for you to also bless someone else—but such an impossibility is the main plot of the Biblical story of Isaac blessing Jacob instead of Esau.

How can we rationally explain this impossibility? What is a blessing—according to the Bible? Can we interpret the Biblical understanding of this term so that the above impossibility makes rational sense?

**What is a blessing: our analysis and the resulting interpretation.** What is a blessing? In the Bible, a blessing somehow makes the blessed person more successful. More rain comes to his/her land, fewer diseases, etc. If we take into account that, according to modern science, these events can only be predicted with some probability, we can described the results of blessing as follows. Due to the blessing, the actual values  $v_i$  of the corresponding quantities become different from their expected mean values  $m_i$ —different in the direction that makes them more beneficial to the blessed person. In these terms, the ability to bless is the ability to change the values of these random quantities.

According to statistics (see, e.g., [8]), in general, if we have several independent random quantities  $v_i$ , then, with very high certainty, all possible combinations  $v = (v_1, \dots, v_n)$  are characterized by the inequality

$$\sum_{i=1}^n \frac{(v_i - m_i)^2}{\sigma_i^2} \leq \chi^2, \quad (1)$$

where  $\sigma_i$  is the corresponding standard deviation and the exact value of  $\chi^2 \approx n$  depends on the desired degree of certainty. From this viewpoint, if we make the blessed person to be very successful, i.e., if we increase some of the differences  $v_i - m_i$  way beyond the random-explained standard deviation  $\sigma_i$ , we thus restrict the possibility to increase other differences  $v_j - m_j$ , since, according to the formula (1), the weighted sum of the squares of these differences is bounded from above.

In this interpretation, blessing is a kind of a new physical field that somewhat changes the probabilities of random events—and in this interpretation, the person's ability to bless is indeed limited.

This interpretation also explains the opposite of blessing—a curse, in which, vice versa, the values of the related physical quantities make the cursed person less happy.

## 6 Sacrificing the Best Animals Verus Darwin

**Biblical idea and why it sound irrational.** According to the Bible, we should sacrifice our best animals to God. This seems to be inconsistent with selection, where we constantly improve the quality of the animals by making the best ones actively reproduce.

If instead of using the best horses, the best bulls, the best sheep to actively reproduce, we sacrifice them, this can probably lead to the effect opposite to selection—namely, to the continual degradation of the stock. This cannot be what God had in mind.

**This can also be rationally explained.** While we do have a lot of experience with selection, we have significantly more experience with computer simulations of such a selection—namely, the experience of using genetic algorithms and, more generally, evolutionary computations, a widely used and largely successful optimization technique.

This experience has shown that one of the main problems with these algorithms—as well as with many other optimization algorithms—is that we sometimes reach a local maximum and get stuck there [2]. One of the main ideas of how to avoid getting stuck in a local maximum is that if we get stuck, we get out—worsening the quality of the current solution, but hoping this way to find solutions which are even better. (One of the main techniques for doing this is known as simulated annealing; see, e.g., [2].)

This is exactly what sacrificing the best bull achieves: deletes the local maximum and thus, allows us to potentially progress to an even better cattle.

## 7 Fast-and-Feast

**Biblical idea.** The Bible pays a lot of attention to when we should fast and when we should feast.

Taking into account that in those days, hunger was an acute problem, it seems to make more sense to equally distribute whatever we have between different days—just like those who have survived in hostile environments usually do. From this viewpoint, the Biblical recommendation seems irrational. But is it?

**Our explanation.** Suppose that our goal is to increase the overall people's satisfaction. Let us describe this problem in precise terms.

The overall satisfaction can be obtained by adding up all the satisfaction levels that people get every day. Let us denote:

- the overall amount of food that we have for a certain period of  $n$  days by  $F$ ,
- the minimal amount of daily food needed to survive by  $f_0$ ,
- the amount of food consumed on day  $i$  by  $f_i$ , and
- the satisfaction of getting the amount of food  $f$  by  $s(f)$ .

In these terms, the corresponding optimization problem has the following form:

- given the values  $F$  and  $f_0$  and the function  $s(f)$ ,
- find the values  $f_1, \dots, f_n$  that maximize the overall satisfaction  $\sum_{i=1}^n s(f_i)$  under the constraints  $\sum_{i=1}^n f_i = F$  and  $f_i \geq f_0$  for all  $i$ .

In general, we can use the Lagrange multiplier method to deduce the above constraint satisfaction problem to the unconstrained problem of maximizing the value

$$\sum_{i=1}^n s(f_i) + \lambda \cdot \left( \sum_{i=1}^n f_i - F \right) \quad (2)$$

under the condition  $f_i \geq f_0$ , where  $\lambda$  is an appropriate constant (known as *Lagrange multiplier*).

According to calculus, if for some  $i$ , the maximum is attained inside the corresponding domain, i.e., for  $f_i > f_0$ , then the partial derivative of the expression (2) should be equal to 0, i.e., we should have  $s'(f_i) = -\lambda$ , where  $s'(f)$  denotes the derivative of the function  $s(f)$ .

Thus, for each day, the consumption  $f_i$  should be equal either to  $f_0$  or to the value  $f_{\text{opt}} > f_0$  for which  $s'(f_{\text{opt}}) = -\lambda$ . With the exception of two degenerate cases when  $F = n \cdot f_0$  and when  $F = n \cdot f_{\text{opt}}$ , the optimal solution has to include *both* “fast” days when  $f_i = f_0$  and “feast” days when  $f_i = f_{\text{opt}}$ . And this is exactly what the Bible recommends.

*Comment.* So why do people surviving in the hostile environments do not follow this optimal strategy? This is easy to explain: these folks do not know how many days they will be there before they are rescued.

**Similar arguments explains the emphasis on Shabbat.** Similar arguments can be applied not only to the amount of food leading to the optimal overall satisfaction, but also to the amount of daily effort leading to the optimal overall productivity. In this case, the optimal strategy is to have days when we work intensely and days when we rest and do not work at all—and this is exactly the Biblical idea of the Sabbath!

## 8 Conclusions and Future Work

**Conclusions.** In this paper, we showed that many Biblical ideas make rational sense. Our objective was to provide several such examples.

**Future work.** There are definitely many more examples in the Bible that can be rationally explained—and probably many examples that cannot be explained rationally. It would be nice to analyze other Biblical ideas from this viewpoint.

**Acknowledgements** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes), and by the AT&T Fellowship in Information Technology. It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478, and by a grant from the Hungarian National Research, Development and Innovation Office (NRDI).

## References

1. Central Conference of American Rabbis (CCAR) The Union Prayer-Book for Jewish Worship, Vol. 2, CCAR Press, Cincinnati, Ohio (2018)
2. Chong, E.K.P., Zak, S.H.: An Introduction to Optimization. Wiley, Hoboken, New Jersey (2013)
3. Feynman, R., Leighton, R., Sands, M.: The Feynman Lectures on Physics. Addison Wesley, Boston, Massachusetts (2005)
4. Kosheleva, O., Zakharevich, M., Kreinovich, V.: If many physicists are right and no physical theory is perfect, then by using physical observations, we can feasibly solve almost all instances of each NP-complete problem. *Mathematical Structures and Modeling* **31**, 4–17 (2014)
5. Kreinovich, V., Lakeyev, A., Rohn, J., Kahl, P.: Computational Complexity and Feasibility of Data Processing and Interval Computations. Kluwer, Dordrecht (1998)
6. Maimonides, M.: The Guide for the Perplexed. Dover Publications, New York (2000)
7. Papadimitriou, C.: Computational Complexity. Addison-Wesley, Reading, Massachusetts (1994)
8. Sheskin, D.J.: Handbook of Parametric and Non-Parametric Statistical Procedures. Chapman & Hall/CRC, London, UK (2011)
9. Thorne, K.S., Blandford, R.D.: Modern Classical Physics: Optics, Fluids, Plasmas, Elasticity, Relativity, and Statistical Physics. Princeton University Press, Princeton, New Jersey (2017)
10. Walker, P.E.: The Wellsprings of Wisdom: A Study of Abu Yaqub AISijistani's Kitab Al-Yanabi, Including a Complete English Translation With Commentary and Notes, University of Utah Press, Salt Lake City, Utah (1994)

# **General Computational Aspects**



# Why Model Order Reduction



Salvador Robles, Martine Ceberio, and Vladik Kreinovich

**Abstract** Reasonably recently, a new efficient method appeared for solving complex non-linear differential equations (and systems of differential equations). In this method—known as Model Order Reduction (MOR)—we select several solutions, and approximate a general solution by a linear combination of the selected solutions. In this paper, we use the known explanation for efficiency of neural networks to explain the efficiency of MOR techniques.

## 1 Formulation of the Problem

**We need to solve systems of differential equations.** In physics, in engineering, in many areas of biology, the corresponding phenomena are described by systems of differential equations. Thus, to make predictions about these phenomena, we need to solve such systems.

**Solving systems of differential equations is difficult.** In general, systems of differential equations are difficult to solve. This difficulty is easy to explain:

- In general, when we solve a system of  $N$  equations with  $N$  unknowns, the more unknowns we have, the more difficult it is to solve this system.
- In systems of differential equations, the unknowns are the functions  $s(x)$ . To exactly describe a general function, we need to describe infinitely many different numerical values—e.g., the values  $s(x_i)$  of this function at all possible points  $x_i$ .

---

S. Robles · M. Ceberio · V. Kreinovich (✉)

Department of Computer Science, University of Texas at El Paso, El Paso, Texas 79968, USA  
e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

S. Robles  
e-mail: [sroblesher1@miners.utep.edu](mailto:sroblesher1@miners.utep.edu)

M. Ceberio  
e-mail: [mceberio@utep.edu](mailto:mceberio@utep.edu)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023  
M. Ceberio and V. Kreinovich (eds.), *Decision Making Under Uncertainty and Constraints*, Studies in Systems, Decision and Control 217,  
[https://doi.org/10.1007/978-3-031-16415-6\\_35](https://doi.org/10.1007/978-3-031-16415-6_35)

233

The more accurately we want to represent a function, the more parameters we will need. To get a good approximation to the desired function, we therefore need to solve a system with a large number of unknowns—which requires a lot of computational efforts.

**Model Order Reduction.** Reasonably recently, a new method appeared—known as Model Order Reduction (MOR, for short) that helps to solve systems of differential equations; see, e.g., [1]. In this method, once we have found several different solutions  $s_1(x), \dots, s_n(x)$ , we then look for approximate solutions  $s(x)$  which are linear combinations of the known solutions, i.e., that have the form

$$s(x) = c_1 \cdot s_1(x) + \dots + c_n \cdot s_n(x)$$

for some coefficient  $c_i$ .

In this approximation, we have only  $n$  unknowns, so when  $n$  is reasonably small, we have a relatively easy-to-solve system of equations.

**This method works, but why?** The main idea of this method comes from linear systems, where, once you have several solutions, any linear combination of these solutions is also a solution. Many real-life systems are, however, non-linear. Interestingly, MOR method works very well for many non-linear systems as well.

Why it works is not clear. In this paper, we provide a possible explanation for this empirical success. This explanation is related to the explanation of another empirical success phenomena—an explanation of why neural networks (see, e.g., [2–4]) work well in many situations.

## 2 From Neural Networks to Model Order Reduction: Our Explanation

**Why neural networks: a reminder.** One of the main original motivations for neural networks came from the need to speed up computations—and from the observation of how biological neural networks process data.

Computers can now perform many tasks that humans do: e.g., they can recognize faces, control cars, etc. However, computers perform these tasks by using super-fast processing units that perform billions of operations per seconds, while we humans perform the same tasks by using neurons the fastest of which can perform at most 100 operations per second. The reason why a human brain can make important decisions in a short period of time is that in the brain, there are billions of neurons that work in parallel. As a result, during the time when one neuron processes data, all involved neurons perform billions of computational steps.

What is the fastest way to set up such parallel computations? In parallel computations, first, all the processors perform some operations, then they perform some other operations, etc. Computations are the fastest when each of these operations requires

the smallest amount of computation time, and when the number of such consequent operations is the smallest possible.

**Which operations are the fastest?** In a deterministic computer, the result of each operation is uniquely determined by its inputs, i.e., in mathematical terms, is a function of these inputs. Out of all possible functions, linear functions are the fastest to compute. However, we cannot use only linear functions: if all the processors were computing linear functions of their inputs, then all we could compute are compositions of linear functions—which are also linear, while many real-life processes are non-linear. Thus, in addition to linear functions, we should also compute some non-linear functions.

In general, the more inputs we have, the longer it takes to perform the corresponding computations. Thus, the fastest is to compute non-linear functions with the smallest number of inputs—i.e., non-linear functions  $s(x)$  with only one input  $x$ . So, to make computations faster, on each computation stage, we either compute a linear function or a non-linear function of one variable.

To make computations fast, a linear stage cannot be followed by a linear stage. Indeed, if after computing a linear function, we again compute a linear function of the first stage's output, we will still be computing a linear function of the original inputs—and this can be done in a single stage. Similarly, if we first compute a function  $y = s(x)$  of one variable, and then compute another function  $z = t(y)$  of one variable, then, in effect, we compute a composition  $z = t(s(x))$  of these functions, and this can also be done in a single stage. Thus, in fast computations, a linear stage must be followed by a non-linear stage, and a non-linear stage must be followed by a linear stage.

**How many stages do we need?** If we use only one stage, then all we can compute are either linear functions or functions of one variable, and many real-life quantities depend non-linearly on several variables. So, we need to have at least two layers. This is exactly how a neural network works: each of its processing units (neurons):

- first computes a linear combination  $y = w_1 \cdot x_1 + \dots + w_n \cdot x_n + w_0$  of its inputs  $x_1, \dots, x_n$ , and
- then applies a non-linear function  $z = s(y)$ —known as an *activation function*—to the resulting value  $y$ .

As a result, each neuron computes the value

$$z = s(w_1 \cdot x_1 + \dots + w_n \cdot x_n + w_0).$$

**From general to specific computational problems.** Neural networks are used for machine learning, when:

- we have no prior information about the dependence between the quantities, and
- we want to determine this dependence based on observation results.

In this case, it makes sense to require that a neural network be able to approximate any possible dependence. So, the activation functions are selected to make sure that

the corresponding neural networks are *universal approximators*—i.e., that they can approximate any reasonable function with any given accuracy.

For a general neural network, in principle, in addition to activation functions (that need to be computed every time), we can also use functions that have already been computed before. Using these functions will not add computation time—since these functions have already been computed before. However, since a neural network is intended to compute all possible functions from all possible domains, having a pre-computed function from, e.g., biology will probably not help in solving the next problem which may be from geosciences.

In contrast, when we solve a given system of differential equations, we are interested in very specific functions—solutions to this system of equations. Many of these solutions—e.g., corresponding to similar initial conditions—are similar, so it is reasonable to expect that knowing a solution to a similar problem can help in solving the current problem. Thus, for solving systems of differential equations, it makes sense to consider, in addition to activation functions (of one variable), also use pre-computed functions  $s_1(x), \dots, s_n(x)$ , possibly of several variables.

What can we compute this way if we use the fastest (two-stage) computations? Since a linear layer cannot be followed by a linear one and a non-linear stage cannot be followed by a non-linear one, we have two options:

- we can have a linear stage followed by a non-linear stage; we will denote this option by L-NL, and
- we can have a non-linear stage followed by a linear stage; we will denote this option by NL-L.

**L-NL option.** In this option, first, we compute some linear combinations  $T(x)$  of the inputs, and then apply an appropriate non-linear function  $s_i$ , resulting in  $s_i(T(x))$ .

The problem is that in this option, we have  $n$  different families of functions corresponding to using  $n$  different pre-computed functions  $s_i(x)$ . There is no continuous transition between these families. In this sense, we have a union of  $n$  disconnected families of functions. However, what we want to approximate is the family of all solutions, which continuously depend on initial conditions and parameters of the system. In other words, in this option, there is a discrepancy between:

- the class of functions that we want to approximate—namely, the class of all solutions corresponding to different initial conditions and different values of the parameters, and
- the class of functions that we use for approximation—in this option, the class of functions  $s_i(T(x))$  corresponding to  $i = 1, \dots, n$ .

This leaves us with the need to consider the second option.

**NL-L option.** In this case, first, we apply non-linear functions, i.e., compute the values  $y_1 = s_1(x), \dots, y_n = s_n(x)$ , and then we compute a linear combination of these values, i.e., an expression

$$c_1 \cdot y_1 + \dots + c_n \cdot y_n + c_0 = c_1 \cdot s_1(x) + \dots + c_n \cdot s_n(x) + c_0.$$

In this option, different solutions correspond to different values of  $c_i$ , so they can be easily smoothly transformed into one another.

Modulo a constant term  $c_0$ , what we get in this option is exactly the approximation used in Model Order Reduction (MOR). Thus, we have indeed explained the empirical success of the MOR techniques: they naturally appear if we are looking for the fastest-to-compute approximations.

**Acknowledgments** This work was supported in part by the National Science Foundation grants:

- 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and

- HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported:

- by the AT&T Fellowship in Information Technology, and

- by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

## References

1. Benner, P., Grivet-Talosa, S., Quarteroni, A., Rozza, G., Schilders, W., Silveira, L.M. (eds.): Model Order Reduction. de Gruyter, Berlin (2020)
2. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)
3. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge, Massachusetts (2016)
4. Kreinovich, V., Kosheleva, O.: Optimization under uncertainty explains empirical success of deep learning heuristics. In: Pardalos, P., Rasskazova, V., Vrahatis, M.N. (eds.) Black Box Optimization, pp. 195–220. Machine Learning and No-Free Lunch Theorems, Springer, Cham, Switzerland (2021)

# Bounding the Range of a Sum of Multivariate Rational Functions



Mohammad Adm, Jürgen Garloff, Jihad Titi, and Ali Elgayar

**Abstract** Bounding the range of a sum of rational functions is an important task if, e.g., the global polynomial sum of ratios problem is solved by a branch and bound algorithm. In this paper, bounding methods are discussed which rely on the expansion of a multivariate polynomial into Bernstein polynomials.

**Keywords** Multivariate rational function · Range enclosure · Bernstein polynomial

## 1 Introduction

In this paper, we consider the expansion of a multivariate polynomial into Bernstein polynomials over a box, i.e., an axis-aligned region, in  $\mathbb{R}^n$ . This expansion has many applications, e.g., in computer aided geometric design, robust control, global optimization, differential and integral equations, and finite element analysis [8, 13]. A very useful property of this expansion is that the interval spanned by the minimum and maximum of the coefficients of this expansion, the so-called *Bernstein coefficients*,

---

M. Adm (✉)

Department of Applied Mathematics and Physics, Palestine Polytechnic University, Hebron, Palestine

e-mail: [moh\\_95@ppu.edu](mailto:moh_95@ppu.edu)

J. Garloff (✉)

Institute for Applied Research, University of Applied Sciences / HTWG Konstanz, D-78462, Konstanz, Germany

e-mail: [juergen.garloff@htwg-konstanz.de](mailto:juergen.garloff@htwg-konstanz.de)

J. Titi

Wedad Nasser Al-Deen Secondary Girls School, Hebron, Palestine and Department of Applied Mathematics and Physics, Palestine Polytechnic University, Hebron, Palestine

e-mail: [jihadtiti@yahoo.com](mailto:jihadtiti@yahoo.com)

A. Elgayar

Faculty of Engineering, University of Benghazi, Benghazi, Libya

e-mail: [Ali.elgayar@uob.edu.ly](mailto:Ali.elgayar@uob.edu.ly)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

239

M. Ceberio and V. Kreinovich (eds.), *Decision Making Under Uncertainty*

and *Constraints*, Studies in Systems, Decision and Control 217,

[https://doi.org/10.1007/978-3-031-16415-6\\_36](https://doi.org/10.1007/978-3-031-16415-6_36)

provides bounds for the range of the given polynomial over the considered box, see, e.g., [11]. A simple (but by no means economic) method for the computation of the Bernstein coefficients from the coefficients of the given polynomial is the use of formula (2) below. This formula (and also similar ones for the Bernstein coefficients over more general sets like simplices and polytopes) allows the symbolic computation of these quantities when the coefficients of the given polynomial depend on parameters. Some applications are making use of this symbolic computation: In [6, Sections 3.2 and 3.3] and the many references therein, the reachability computation and parameter synthesis with applications in biological modelling are considered. In [4, 5], parametric polynomial inequalities over parametric boxes and polytopes are treated. Applications in static program analysis and optimization include dependence testing between references with linearized subscripts, dead code elimination of conditional statements, and estimation of memory requirements in the development of embedded systems. Applications which involve polynomials of higher degree or many variables require a computation of the Bernstein coefficients which is more economic than by formula (2). In [21], the second and third authors have presented a matrix method for the computation of the Bernstein coefficients which is faster than the methods developed so far and which is included in version 12 of the MATLAB toolbox INTLAB [17].

In this paper, we aim at finding bounds for the range of a sum of rational functions over a box. This problem appears when the global polynomial sum of ratios problem is solved by a branch and bound method, see, e.g., [7, 10]. The sum of ratios problem is one of the most difficult fractional programming problems encountered so far<sup>1</sup>.

After having introduced the Bernstein expansion in Sect. 2, we will extend in Sect. 3 the bounds for the range of a single rational function to a sum of rational functions. In the sequel we employ the following notation. Let  $n \in \mathbb{N}$  (set of the nonnegative integers) be the number of variables. A multi-index  $(i_1, \dots, i_n) \in \mathbb{N}^n$  is abbreviated by  $i$ . In particular, we write 0 for  $(0, \dots, 0)$ . Arithmetic operations with multi-indices are defined entry-wise; the same applies to comparison between multi-indices. For  $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ , its *monomials* are defined as

$x^i := \prod_{s=1}^n x_s^{i_s}$ . For  $d = (d_1, \dots, d_n) \in \mathbb{N}^n$  such that  $i \leq d$ , we use the compact nota-

tions  $\sum_{i=0}^d := \sum_{i_1=0}^{d_1} \cdots \sum_{i_n=0}^{d_n}$  and  $\binom{d}{i} := \prod_{s=1}^n \binom{d_s}{i_s}$ .

---

<sup>1</sup> The problem of optimizing one or several ratios of functions is called a *fractional program*. The ninth bibliography of fractional programming [18] covering mainly the period 2016–2018 lists 520 papers on fractional programming and its applications.

## 2 Bernstein Expansion

In this section, we present fundamental properties of the Bernstein expansion over a box, e.g., [8, Subsection 5.1], [11, 16], that are employed throughout the paper. For simplicity we consider the unit box  $\mathbf{u} := [0, 1]^n$ , since any compact nonempty box  $\mathbf{x}$  of  $\mathbb{R}^n$  can be mapped affinely onto  $\mathbf{u}$ . Let  $\ell \in \mathbb{N}^n$ ,  $a_j \in \mathbb{R}$ , with  $j = 0, \dots, \ell$ , such that for  $s = 1, \dots, n$ ,

$$\ell_s := \max \{q \mid a_{j_1, \dots, j_{s-1}, q, j_{s+1}, \dots, j_n} \neq 0\}.$$

Let  $p$  be an  $\ell$ -th degree  $n$ -variate polynomial with the power representation

$$p(x) = \sum_{j=0}^{\ell} a_j x^j. \tag{1}$$

We expand  $p$  into Bernstein polynomials of degree  $d$ ,  $d \geq \ell$ , over  $\mathbf{u}$  as

$$p(x) = \sum_{j=0}^d b_j^{(d)}(p) B_j^{(d)}(x),$$

where  $B_j^{(d)}$  is the  $j$ -th Bernstein polynomial of degree  $d$ , defined as

$$B_j^{(d)}(x) := \binom{d}{j} x^j (1-x)^{d-j},$$

and  $b_j^{(d)}(p)$  is the  $j$ -th Bernstein coefficient of  $p$  of degree  $d$  over  $\mathbf{u}$  which is given by

$$b_j^{(d)}(p) = \sum_{i=0}^j \frac{\binom{j}{i}}{\binom{d}{i}} a_i, \quad 0 \leq j \leq d, \tag{2}$$

with the convention that  $a_i := 0$  if  $i \geq \ell$ ,  $i \neq \ell$ .

Note that by (2) the Bernstein coefficients are linear: Let  $p_1$  and  $p_2$  be polynomials with the power representations (1) with  $\ell = \ell^{(1)}$  and  $\ell = \ell^{(2)}$ , respectively, and let  $\ell := \max \{\ell^{(1)}, \ell^{(2)}\}$ . If  $p = \alpha p_1 + \beta p_2$ ,  $\alpha, \beta \in \mathbb{R}$ , then

$$b_j^{(d)}(p) = \alpha b_j^{(d)}(p_1) + \beta b_j^{(d)}(p_2), \quad i = 0, \dots, d. \tag{3}$$



### 3 Bounds for the Range of a Sum of Rational Functions

Let  $p$  and  $q$  be two  $n$ -variate real polynomials with the Bernstein coefficients over the unit box  $\mathbf{u}$  given by  $b_i^{(d)}(p)$  and  $b_i^{(d)}(q)$ ,  $0 \leq i \leq d$ , respectively. We assume that the two polynomials have the same degree  $l$  since otherwise we can elevate the degree of the Bernstein expansion of either polynomial by component where necessary to ensure that their Bernstein coefficients are of the same order  $d \geq l$ . We consider the multivariate rational function  $f := \frac{p}{q}$  over  $\mathbf{u}$ . In the sequel we assume that all  $b_i^{(d)}(q)$ ,  $i = 0, \dots, d$ , have the same strict sign (and without loss of generality we may assume that all of them are positive). We use the notation for the *rational* Bernstein coefficients of  $f$

$$b_i^{(d)}(f) := \frac{b_i^{(d)}(p)}{b_i^{(d)}(q)}, \quad i = 0, \dots, d. \tag{4}$$

Then an enclosure for the range of  $f$  over  $\mathbf{u}$  is given by the following theorem which includes also the polynomial case ( $q = 1$ ).

**Theorem 1** [15, Theorem 3.1], [12, Proposition 3] *The range of  $f$  over  $\mathbf{u}$  can be bounded by*

$$\min_{i=0, \dots, d} b_i^{(d)}(f) \leq f(x) \leq \max_{i=0, \dots, d} b_i^{(d)}(f), \quad x \in \mathbf{u}. \tag{5}$$

(Vertex Condition) *Equality holds in the left or right inequality if and only if the minimum or the maximum of the Bernstein coefficients is attained at a vertex index  $i_s$  with  $i_s \in \{0, d_s\}$ ,  $s = 1, \dots, n$ .*

Now we extend the bounds for the range over a box of a single rational function to a sum of such functions. Without loss of generality, we consider here only the case that we have solely two rational functions,

$$f = f_1 + f_2, \quad \text{where } f_1 = \frac{p_1}{q_1}, \quad f_2 = \frac{p_2}{q_2}. \tag{6}$$

We assume that both the numerator and denominator polynomials have the common degree  $\ell$  and that all the Bernstein coefficients of each denominator polynomial have the same strict sign (but may be different for  $q_1$  and  $q_2$ ). By the additivity of the Bernstein coefficients (3) and the enclosure (5), one may conjecture that

$$\min_{i=0, \dots, d} (b_i^{(d)}(f_1) + b_i^{(d)}(f_2)) \leq f(x) \leq \max_{i=0, \dots, d} (b_i^{(d)}(f_1) + b_i^{(d)}(f_2)), \quad x \in \mathbf{u}. \tag{7}$$

However, this conjecture is not true even in the case of ratios of linear functions as the following example shows.

**Example 1** Let  $f_1(x) = \frac{2x+1}{x+1}$  and  $f_2(x) = \frac{0.2x+1}{5x+1}$ . Then  $f = f_1 + f_2$  attains its global minimum  $\approx 1.645445$  on  $[0, 1]$  at  $\approx 0.4239$ . The rational Bernstein coefficients of  $f_1$  and  $f_2$  are  $b_0^{(1)}(f_1) = 1, b_1^{(1)}(f_1) = 1.5, b_0^{(1)}(f_2) = 1, b_1^{(1)}(f_2) = 0.2$ , such that the lower bound in (7) is 1.7 which is greater than the global minimum of  $f$ .

We will return to (7) in Example 3.

### 3.1 The Naïve Bounds

To motivate the enclosure (11) below, we consider first the univariate case ( $n = 1$ ). We start with recalling a formula for the Bernstein coefficients of the product  $pr$  of two polynomials  $p$  and  $r$  of degrees  $\ell(p)$  and  $\ell(r)$  in terms of their Bernstein coefficients, see [9, formula (44)]. In the sequel, we suppress in the presentation of the Bernstein coefficients the reference to their degrees. Since for the degree  $\ell$  of the polynomial  $pr, \ell = \ell(p) + \ell(r)$  holds, we obtain for  $k = 0, 1, \dots, \ell$

$$\begin{aligned}
 b_k(pr) &= \sum_{\mu=\max\{0,k-\ell(r)\}}^{\min\{\ell(p),k\}} \frac{\binom{\ell(p)}{\mu} \binom{\ell(r)}{k-\mu}}{\binom{\ell}{k}} b_\mu(p) b_{k-\mu}(r) \\
 &\leq \max_{\mu} b_\mu(p) b_{k-\mu}(r) \frac{1}{\binom{\ell}{k}} \sum_{\mu=\max\{0,k-\ell(r)\}}^{\min\{\ell(p),k\}} \binom{\ell(p)}{\mu} \binom{\ell(r)}{k-\mu}. \tag{8}
 \end{aligned}$$

By the Vandermonde convolution, the last sum in (8) equals  $\binom{\ell}{k}$  such that we can conclude

$$b_k(pr) \leq \max_{\mu} b_\mu(p) b_{k-\mu}(r).$$

An analogous lower bound is provided by replacing the maximum by the minimum.

Returning to the two-term case in (6), we assume for simplicity that both the numerator and denominator polynomials have the common degree  $\ell$  and that the Bernstein coefficients of  $q_1$  and  $q_2$  have the same strict sign. Put

$$M := \max_{i,j=0,\dots,\ell} (b_i(f_1) + b_j(f_2))$$

and

$$s := p_1q_2 + q_1p_2 - Mq_1q_2. \tag{9}$$

Then by (3), we obtain for  $k = 0, 1, \dots, 2\ell$

$$b_k(s) = b_k(p_1q_2) + b_k(q_1p_2) - Mb_k(q_1q_2),$$

and by (8) with coefficients  $\alpha_\mu$  satisfying  $\sum_{\mu} \alpha_\mu = 1$

$$\begin{aligned}
 b_k(s) &= \sum_{\mu=\max\{0,k-\ell\}}^{\min\{\ell,k\}} \alpha_\mu (b_\mu(p_1)b_{k-\mu}(q_2) + b_\mu(q_1)b_{k-\mu}(p_2) - Mb_\mu(q_1)b_{k-\mu}(q_2)) \\
 &= \sum_{\mu=\max\{0,k-\ell\}}^{\min\{\ell,k\}} \alpha_\mu b_\mu(q_1)b_{k-\mu}(q_2) \left( \frac{b_\mu(p_1)}{b_\mu(q_1)} + \frac{b_{k-\mu}(p_2)}{b_{k-\mu}(q_2)} - M \right) \\
 &\leq 0,
 \end{aligned} \tag{10}$$

by the definition of  $M$ . Since by Theorem 1  $s(x) \leq \max_{k=0,\dots,2\ell} b_k(s)$ ,  $x \in \mathbf{u}$ , we conclude that  $s(x) \leq 0$  and therefore,

$$f_1(x) + f_2(x) \leq M, \quad x \in \mathbf{u}.$$

Similarly we obtain a lower bound for  $f_1 + f_2$  on  $\mathbf{u}$  if we replace the maximum by the minimum. The resulting enclosure for the range of  $f = f_1 + f_2$  on  $\mathbf{u}$

$$\min_{i,j=0,\dots,d} (b_i^{(d)}(f_1) + b_j^{(d)}(f_2)) \leq f(x) \leq \max_{i,j=0,\dots,d} (b_i^{(d)}(f_1) + b_j^{(d)}(f_2)), \quad x \in \mathbf{u}, \tag{11}$$

is simply the enclosure which we obtain if we form the (Minkowski) sum of the enclosure (5) for  $f_1$  and  $f_2$ . Therefore, this enclosure is obviously true also in the  $n$ -variate case which we will consider now again.

We put  $\bar{f} := \max_{x \in \mathbf{u}} f(x)$  and for  $d \geq \ell$ ,

$$\begin{aligned}
 \underline{m}^{(d)} &:= \min_{i,j=0,\dots,\ell} (b_i^{(d)}(f_1) + b_j^{(d)}(f_2)), \\
 \bar{m}^{(d)} &:= \max_{i,j=0,\dots,\ell} (b_i^{(d)}(f_1) + b_j^{(d)}(f_2)).
 \end{aligned}$$

In the sequel, we present our results mainly only for the upper bounds. Analogous results hold for the lower bounds.

**Theorem 2** *The following vertex condition holds*

$$\bar{f} = \bar{m}^{(d)} \text{ if and only if } \bar{m}^{(d)} = b_{i^*}^{(d)}(f_1) + b_{i^*}^{(d)}(f_2) \text{ for a vertex index } i^*.$$

**Proof** Assume that  $\bar{m}^{(d)}$  is attained at a vertex index  $i^*$ . Then the statement is clear because the sum of the related Bernstein coefficients is a function value of  $f$ , see [15, Remark 1]. Conversely, assume that  $\bar{f} = \bar{m}^{(d)}$ , and let  $\bar{f} = f(\hat{x})$  for some  $\hat{x} \in \mathbf{u}$ . Define the polynomial  $s$  as in (9) with  $M = \bar{m}^{(d)}$ . Then we can conclude that

$$\frac{s(\hat{x})}{q_1(\hat{x})q_2(\hat{x})} = f(\hat{x}) - \bar{m}^{(d)} = 0,$$

hence  $s(\hat{x}) = 0$ . Since  $s$  is nonpositive on  $\mathbf{u}$ , it attains its maximum at  $\hat{x}$ . On the other hand, in the multivariate case a straightforward extension of formula

(8) for the product of two polynomials in the Bernstein representation exists, see [2, Section 3.3], by which we can conclude as in (10) that  $b_i(s) \leq 0$ , for  $i = 0, \dots, 2d$ . Since  $s(x) \leq \max_{i=0, \dots, 2d} b_i(s)$ , it follows that there exists an index  $i^*$  with  $b_{i^*}(s) = 0$ , whence

$$\max_{x \in \mathbf{u}} s(x) = b_{i^*}(s).$$

By the polynomial vertex condition in Theorem 1, we can conclude that the index  $i^*$  is a vertex index.

In [12], some properties of the bounds in the case of a single rational function are presented. From Proposition 4 and Theorem 8 therein it immediately follows that also in the multi-term case the bounds are monotone, i.e., for  $l \leq d \leq k$  it holds that  $\underline{m}^{(d)} \leq \underline{m}^{(k)}$  and  $\overline{m}^{(k)} \leq \overline{m}^{(d)}$ , and that the so-called inclusion isotonicity of the interval function provided by the enclosure  $[\underline{m}^{(d)}(f, \mathbf{x}), \overline{m}^{(d)}(f, \mathbf{x})]$  is valid. However, compared to the single-term case, we are losing one order of convergence of the bounds to the range. So, degree elevation may not result in linear convergence. This is shown by the following example.

**Example 2** We choose  $n = 1$ ,  $f_1(x) = \frac{x}{2-x}$ ,  $f_2(x) = \frac{2-2x}{2-x}$ . Then  $f(x) = 1$ ,  $x \in \mathbf{u}$ . The two Bernstein coefficients for  $d = 1$  of  $f_1$  as well as of  $f_2$  are 0 and 1. So  $\overline{m}^{(1)} = 2$  which cannot be improved by degree elevation because both coefficients are function values.

To enforce convergence of the bounds to the range we apply subdivision. The convergence result (Theorem 4) will immediately follow from the linear convergence of the bounds with respect to the width of the box.

**Theorem 3** Let  $\mathbf{x} = [\underline{x}, \overline{x}]$  be any subbox of  $\mathbf{u}$ . Then

$$\max_{i,j=0, \dots, d} (b_i^{(d)}(f_1, \mathbf{x}) + b_j^{(d)}(f_2, \mathbf{x})) - \max_{x \in \mathbf{x}} f(x) \leq \delta \|\overline{x} - \underline{x}\|_\infty,$$

where  $\delta$  is a constant not depending on  $\mathbf{x}$ .

**Proof** Let  $\max_{x \in \mathbf{x}} f(x) = f(x')$ , with  $x' \in \mathbf{x}$ , and define  $\overline{f}_m := \max_{x \in \mathbf{x}} f_m(x)$ ,  $m = 1, 2$ . Then  $f(x')$  can be written as

$$f(x') = \overline{f}_1 + \overline{f}_2 + f_1(x') - \overline{f}_1 + f_2(x') - \overline{f}_2.$$

We apply the results on quadratic convergence in the single-term case [12, Theorem 6] and a standard argument involving the Mean Value Theorem, e.g., [14, Theorem 4.1.18] to  $f_1$  and  $f_2$  to obtain

$$\max_{i=0, \dots, d} b_i^{(d)}(f_1, \mathbf{x}) + \max_{j=0, \dots, d} b_j^{(d)}(f_2, \mathbf{x}) - f(x') \leq \delta_1 \|\overline{x} - \underline{x}\|_\infty^2 + \delta_2 \|\overline{x} - \underline{x}\|_\infty,$$

where  $\delta_1$  and  $\delta_2$  are constants not depending on  $\mathbf{x}$ . Since  $\|\overline{x} - \underline{x}\|_\infty \leq 1$  the proof is complete.

To simplify the presentation, we will reserve in the sequel the upper index of the Bernstein coefficients for the subdivision level. Repeated bisection of  $\mathbf{u}^{(0,1)} := \mathbf{u}$  in all  $n$  coordinate directions results at subdivision level  $1 \leq h$  in subboxes  $\mathbf{u}^{(h,v)}$  of edge length  $2^{-h}$ ,  $v = 1, \dots, 2^{nh}$ . Denote the Bernstein coefficients of  $f$  over  $\mathbf{u}^{(h,v)}$  by  $b_i^{(h,v)}(f)$ . For their computation see [21].

**Theorem 4** (*Linear convergence with respect to subdivision*) For  $1 \leq h$  it holds

$$\max_{i,j=0,\dots,l; v=1,\dots,2^{nh}} (b_i^{(h,v)}(f_1) + b_j^{(h,v)}(f_2)) - \bar{f} \leq \delta 2^{-h},$$

where  $\delta$  is a constant not depending on  $h$ .

With increasing subdivision level, the chances are becoming better and better that the vertex condition holds on subboxes.

In the subdivision process, it may be advantageous to check the vertex condition of Theorem 1 term-wise because then we will be able to detect terms for which we have already found the true minimum or maximum of the respective rational functions such that a further division of the boxes under consideration is not necessary for these terms. If the vertex condition is satisfied for the lower or the upper bounds for all terms and the individual vertex indices coincide for at least one index, then the vertex condition in Theorem 2 is fulfilled, and we already have found the true minimum or maximum of the sum of ratios.

The convergence can possibly be speeded up by employing term-wise the monotonicity and dominance tests presented in [19, Section 6.1].

**Example 3** In [1, Example 3], see also [10, (5.14)], the function  $f$

$$\begin{aligned} f := & \frac{-x_1^2 + 16x_1 - x_2^2 + 16x_2 - x_3^2 + 16x_3 - x_4^2 + 16x_4 - 214}{2x_1 - x_2 - x_3 + x_4 + 2} \\ & + \frac{-x_1^2 + 16x_1 - 2x_2^2 + 20x_2 - 3x_3^2 + 60x_3 - 4x_4^2 + 56x_4 - 586}{-x_1 + x_2 + x_3 - x_4 + 10} \\ & + \frac{-x_1^2 + 20x_1 - x_2^2 + 20x_2 - x_3^2 + 20x_3 - x_4^2 + 20x_4 - 324}{x_1^2 - 4x_4}, \end{aligned}$$

where

$$x_1 \in [6, 10], \quad x_2 \in [4, 6], \quad x_3 \in [8, 12], \quad x_4 \in [6, 8],$$

is to maximize. We have chosen the precision  $\epsilon = 10^{-5}$  and have used an HP OMEN laptop with Intel®Core™ i7-10750H with CPU 2.20-5.0 GHz and 16 GB RAM. The method presented in Section 3 results in 0.043 ms at subdivision level  $h = 7$  in the upper bound 16.16667 for  $\bar{f}$  attained at (6, 6, 10.05502, 8). The upper bound is very close to the bounds presented in [1] (computed with precision  $10^{-2}$ ) and [10] (computed with precision  $10^{-4}$ , according to a private communication). Interestingly, the conjectured bound (7) provides nearly the same bound attained at the same place

but for  $h = 94$ . The much higher subdivision level is not surprising because we cannot employ a vertex condition which is very useful to speed up the subdivision process.

We noticed a similar situation for the minimum. Our algorithm finds in 0.015 ms in only one subdivision step ( $h = 1$ ) the lower bound 0.976190 attained at  $(6, 4, 12, 6)$  for the minimum of  $f$ . Since this bound is attained at a vertex index, the vertex condition in Theorem 2 holds, and we know that we already have found the minimum of  $f$ . The same lower bound is provided by (7) at the same place but for  $h = 72$  which confirms our experience that (7) is true in many cases.

### 3.2 Improved Bounds

In the single-term case, the bounds converge quadratically if subdivision is applied [12, Theorem 7]. Therefore, it appears advantageous to reduce the multi-term case to the single-term case by extending all ratios to the same denominator to obtain a single rational function which is to optimize. In Example 2, this gives the exact range  $\{1\}$  of  $f$ . But such a procedure is not appropriate for a larger number of terms because the degrees of the resulting numerator and denominator polynomials become potentially large. However, we may partition the totality of the terms into groups of two or three terms and apply the procedure to each group. Finally, we form the (Minkowski) sum of all resulting enclosures. This procedure requires to compute the Bernstein coefficients of a product of two polynomials given the Bernstein coefficients of both polynomials. For this task it is beneficial to use one of the methods which are presented in [22, Section 4].

In passing, we note that most of the results presented in this paper easily extend to the Bernstein expansion over simplices [15, Remark 6], [19, 20, 23] which allow more general regions over which a sum of ratios is to optimize.

## 4 Future Work

To fight the increase of the degrees inherent in the method described in Sect. 3.2, one can use the least common multiple of the denominators. To compute this, one employs the greatest common divisor of the polynomials. A method which appears suitable for this task is the method for the division of two polynomials in Bernstein form presented in [3]. However, the focus herein is on the univariate case. Division algorithms for the multivariate case and analogues in the multivariate Bernstein setting of Gröbner bases are also discussed but have to be adapted to our problem. An important point here is that the methods allow all the computations to be performed using only Bernstein coefficients such that no conversion to the monomial coefficients is required.

**Acknowledgements** This article was made possible with the support and within the interdisciplinary setting of the Arab-German Young Academy of Sciences and Humanities (AGYA). AGYA draws on financial support of the German Federal Ministry of Education and Research (BMBF) grant 01DL20003. The second author gratefully acknowledges support from the University of Applied Sciences / HTWG Konstanz through the SRP program.

## References

1. Benson, H.P.: Using concave envelopes to globally solve the nonlinear sum of ratios problem. *J. Global Optim.* **22**, 343–364 (2002)
2. Berchtold, J., Bowyer, A.: Robust arithmetic for multivariate Bernstein-form polynomials. *Comput. Aided Design* **32**, 681–689 (2000)
3. Busé, L., Goldman, R.: Division algorithms for Bernstein polynomials. *Comput. Aided Geom. Design* **25**, 850–865 (2008)
4. Clauss, P., Chupaeva, I.Yu.: Application of symbolic approach to the Bernstein expansion for program analysis and optimization. In: Duesterwald, E. (ed.) *Compiler Construction. Lecture Notes in Computer Science*, vol. 2985, pp. 120–133. Springer, Berlin, Heidelberg (2004)
5. Clauss, P., Fernández, F.J., Garbervetsky, D., Verdoolaege, S.: Symbolic polynomial maximization over convex sets and its application to memory requirement estimation. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **17**(8), 983–996 (2009)
6. Dang, T., Dreossi, T., Fanchon, É., Maler, O., Piazza, C., Rocca, A.: Set-based analysis for biological modelling. In: Liò, P., Zuliani, P. (eds.) *Automated Reasoning for Systems Biology and Medicine, Series Computational Biology*, vol. 30, pp. 157–189. Springer Nature (2019)
7. Dür, M., Horst, R., Thoai, N.V.: Solving sum-of-ratios fractional programs using efficient points. *Optimization* **49**(5–6), 447–466 (2001)
8. Farouki, R.T.: The Bernstein polynomial basis: A centennial retrospective. *Comput. Aided Geom. Design* **29**, 379–419 (2012)
9. Farouki, R.T., Rajan, V.T.: Algorithms for polynomials in Bernstein form. *Comput. Aided Geom. Design* **5**, 1–26 (1988)
10. Gao, L., Mishra, S.K., Shi, J.: An extension of branch-and-bound algorithm for solving sum-of-nonlinear-ratios problem. *Optim. Lett.* **6**, 221–230 (2012)
11. Garloff, J.: Convergent bounds for the range of multivariate polynomials. In: Nickel, K. (ed.) *Interval Mathematics 1985. Lecture Notes in Computer Science*, vol. 212, pp. 37–56. Springer, Berlin, Heidelberg (1986)
12. Garloff, J., Hamadneh, T.: Convergence and inclusion isotonicity of the tensorial rational Bernstein form. In: Nehmeier, M., Wolff von Gudenberg, J., Tucker, W. (eds.) *Scientific Computing, Computer Arithmetic, and Validated Numerics, Lecture Notes in Computer Science*, vol. 9553, pp. 171–179. Springer (2014)
13. Garloff, J., Smith, A.P. (eds.): Special issue on the use of Bernstein polynomials in reliable computing: A centennial anniversary, *Reliab. Comput.* **17** (2012)
14. Mayer, G.: *Interval Analysis and Automatic Result Verification*. de Gruyter Stud. Math., vol. 65. de Gruyter, Berlin, Boston (2017)
15. Narkawicz, A., Garloff, J., Smith, A.P., Muñoz, C.A.: Bounding the range of a rational function over a box. *Reliab. Comput.* **17**, 34–39 (2012)
16. Rivlin, T.J.: Bounds on a polynomial. *J. Res. Nat. Bur. Standards* **74**(B):47–54 (1970)
17. Rump, S.M.: INTLAB-INTErval LABoratory. In: Csendes, T. (ed.) *Developments in Reliable Computing*, pp. 77–104. Kluwer Academic Publishers, Dordrecht (1999)
18. Stancu-Minasian, I.M.: A ninth bibliography of fractional programming. *Optimization* **68**(11), 2125–2169 (2019)

19. Titi, J., Garloff, J.: Fast determination of the tensorial and simplicial Bernstein forms of multivariate polynomials and rational functions. *Reliab. Comput.* **25**, 24–37 (2017)
20. Titi, J., Garloff, J.: Matrix methods for the simplicial Bernstein representation and for the evaluation of multivariate polynomials. *Appl. Math. Comput.* **315**, 246–258 (2017)
21. Titi, J., Garloff, J.: Matrix methods for the tensorial Bernstein form. *Appl. Math. Comput.* **346**, 254–271 (2019)
22. Titi, J., Garloff, J.: Symbolic-numeric computation of the Bernstein coefficients of a polynomial from those of its partial derivatives and of the product of two polynomials. In: Boulier, F., England, M., Sadykov, T.M., Vorozhtsov, E.V. (eds.) *Computer Algebra in Scientific Computing, CASC 2020. Lecture Notes in Computer Science*, vol. 12291, pp. 583–599. Springer, Cham (2020)
23. Titi, J., Hamadneh, T., Garloff, J.: Convergence of the simplicial rational Bernstein form. In: Le Thi, H.A., Tao, P.D., Thanh, N.N. (eds.) *Modelling, Computation and Optimization in Information Systems and Management Sciences. Advances in Intelligent Systems and Computing*, vol. 359, pp. 433–441. Springer, Cham (2015)



# Fourier Transform and Other Quadratic Problems Under Interval Uncertainty



Oscar Galindo, Christopher Ibarra, Vladik Kreinovich, and Michael Beer

**Abstract** In general, computing the range of a quadratic function on given intervals is NP-hard. Recently, a feasible algorithm was proposed for computing the range of a specific quadratic function—square of the modulus of a Fourier coefficient. For this function, the rank of the quadratic form—i.e., the number of nonzero eigenvalues—is 2. In this paper, we show that this algorithm can be extended to all the cases when the rank of the quadratic form is bounded by a constant.

## 1 Formulation of the Problem

**Need for data processing.** Computers are used to estimate the current values of physical quantities and to predict their future values (e.g., to predict tomorrow's temperature). In all these cases, we need to process data.

**Need to take uncertainty into account.** The inputs  $x_1, \dots, x_n$  for such data processing come from measurements (or from expert estimates). Both measurements and expert estimates are not absolutely accurate. Measurement results  $\tilde{x}_i$  are, in general, somewhat different from the actual (unknown) values  $x_i$  of the corresponding quantities. These differences  $\tilde{x}_i - x_i$  are called *measurement errors*. Because of these differences, the result  $\tilde{y} = f(\tilde{x}_1, \dots, \tilde{x}_n)$  of data processing is also somewhat different from the value  $y = f(x_1, \dots, x_n)$  that we would have obtained if we knew the exact values  $x_i$  of the inputs.

---

O. Galindo (✉) · C. Ibarra · V. Kreinovich  
Department of Computer Science, University of Texas at El Paso El Paso, Texas 79968, USA  
e-mail: [ogalindomo@miners.utep.edu](mailto:ogalindomo@miners.utep.edu)

C. Ibarra  
e-mail: [caibarra5@miners.utep.edu](mailto:caibarra5@miners.utep.edu)

V. Kreinovich  
e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

M. Beer  
Institute for Risk and Reliability, Leibniz University Hannover, 30167 Hannover, Germany  
e-mail: [beer@irz.uni-hannover.de](mailto:beer@irz.uni-hannover.de)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023  
M. Ceberio and V. Kreinovich (eds.), *Decision Making Under Uncertainty and Constraints*, Studies in Systems, Decision and Control 217,  
[https://doi.org/10.1007/978-3-031-16415-6\\_37](https://doi.org/10.1007/978-3-031-16415-6_37)

251

**Need for interval uncertainty.** In many practical situations, the only information that we have about measurement uncertainty is the upper bound  $\Delta_i$  on the absolute value of each measurement error. In such situations, if the measurement result is  $\tilde{x}_i$ , then all we know about the actual value  $x_i$  of the corresponding quantity is that this value is in the interval  $[\tilde{x}_i - \Delta_i, \tilde{x}_i + \Delta_i]$ . Under such interval uncertainty, it is desirable to know the range of possible value of  $y$ . Estimating such a range is known as *interval computation*; see, e.g., [2, 4, 5].

**Interval uncertainty: what is known and what we do.** In general, computing such a range is NP-hard already for quadratic functions  $f(x_1, \dots, x_n)$ ; see, e.g., [3]. Recently, a feasible algorithm was proposed for a practically important quadratic problem: the problem of estimating the absolute value (modulus) of Fourier coefficients [1].

In this paper, we show that this feasible algorithm can be extended to a reasonable general class of quadratic problems.

## 2 Class of Quadratic Expressions for Which the Range Can Be Feasibly Computed

A general quadratic function has the form

$$f = \sum_{i=1}^n \sum_{j=1}^n c_{i,j} \cdot x_i \cdot x_j + \sum_{i=1}^n c_i \cdot x_i + c_0.$$

An important characteristic of the matrix  $c_{i,j}$  is its *rank*—the number of non-zero eigenvalues. When we compute the square of the modulus of the Fourier coefficient, the rank of the corresponding matrix is 2. The general case is when the matrix  $c_{i,j}$  has rank  $k$ , i.e., that it has  $k$  non-zero eigenvalues  $\lambda_j$ ,  $j = 1, \dots, k$ . We will denote the corresponding unit eigenvectors by  $(e_{j,1}, \dots, e_{j,n})$ .

## 3 Our Result

We prove that for any fixed  $k$ , there is a feasible algorithm for estimating the range of the corresponding quadratic expression. This algorithm takes time  $O(n^k)$  in the homogeneous case and  $O(n^{k+1})$  in the general case.

So, as  $k$  increases, the time grows fast, and for  $k \approx n$ , we get exponential time. This makes sense: since the problem is NP-hard, we cannot expect lower-than-exponential computation time.

## 4 Facts from Calculus: Reminder

Computing the minimum of  $f$  is equivalent to computing the maximum of  $-f$ . Thus, it is sufficient to be able to compute the maximum.

According to calculus, the maximum with respect to each variable  $x_i \in [\underline{x}_i, \bar{x}_i]$  is attained:

- either for  $x_i = \underline{x}_i$ , then  $\frac{\partial f}{\partial x_i} \leq 0$ —otherwise, if we had  $\frac{\partial f}{\partial x_i} > 0$ , a small increase in  $x_i$  would lead to the larger value of the function;
- or for  $x_i = \bar{x}_i$ , then  $\frac{\partial f}{\partial x_i} \geq 0$ —otherwise, if we had  $\frac{\partial f}{\partial x_i} < 0$ , a small decrease in  $x_i$  would lead to the larger value of the function;
- or for  $x_i \in (\underline{x}_i, \bar{x}_i)$ , then  $\frac{\partial f}{\partial x_i} = 0$ .

## 5 Let Us Apply These Facts to Our Problem

We start with the quadratic expression

$$f = \sum_{i=1}^n \sum_{j=1}^n c_{i,j} \cdot x_i \cdot x_j + \sum_{i=1}^n c_i \cdot x_i + c_0.$$

In terms of eigenvalues and eigenvectors, the quadratic expression takes the form

$$f = \sum_{j=1}^k \lambda_j \cdot \left( \sum_{i=1}^n e_{j,i} \cdot x_i \right)^2 + \sum_{i=1}^n c_i \cdot x_i + c_0.$$

Its partial derivative w.r.t.  $x_i$  is equal to:

$$\frac{\partial f}{\partial x_i} = 2 \sum_{j=1}^k \lambda_j \cdot \left( \sum_{\ell=1}^n e_{j,\ell} \cdot x_\ell \right) \cdot e_{j,i} + c_i.$$

This expression can be described in terms of the following  $(k+1)$ -dimensional vectors:

$$e_i = (e_{1,i}, \dots, e_{k,i}, c_i) \text{ and } e_i^* = (2\lambda_1 \cdot e_{1,i}, \dots, 2\lambda_k \cdot e_{k,i}, 0).$$

Namely, in terms of the dot (scalar) product, we get  $\frac{\partial f}{\partial x_i} = e_i \cdot S$ , where:

$$S \stackrel{\text{def}}{=} \sum_{\ell=1}^n x_\ell \cdot e_\ell^* + (0, \dots, 0, 1).$$

Thus, all the  $(k + 1)$ -dimensional points  $e_i$  for which  $\frac{\partial f}{\partial x_i} = 0$  are located on a  $k$ -dimensional plane  $\{e : e \cdot S = 0\}$ .

Let us first consider the non-degenerate case, when every group of  $k + 1$  vectors  $e_i$  is linearly independent. We can have no more than  $k$  linearly independent vectors on the same  $k$ -dimensional plane. Thus, we can have no more than  $k$  indices  $i$  for which partial derivative is 0.

For points on one side of the plane, we have  $\frac{\partial f}{\partial x_i} < 0$ , so—according to the above calculus-related facts—the maximum is attained for  $x_i = \underline{x}_i$ . For points on the other side of the plane, where  $\frac{\partial f}{\partial x_i} > 0$ , maximum is attained for  $x_i = \bar{x}_i$ .

If there are fewer than  $k$  points at which the derivative is 0, we can move the plane a little bit until it reaches exactly  $k$  points. So, we arrive at the following algorithm.

## 6 Resulting Algorithm: Non-degenerate Case

We are considering the following problem:

- given a quadratic expression with matrix of rank  $k$ :

$$f = \sum_{i=1}^n \sum_{j=1}^n c_{i,j} \cdot x_i \cdot x_j + \sum_{i=1}^n c_i \cdot x_i + c_0$$

and intervals  $[\underline{x}_i, \bar{x}_i]$ ,

- find: the range  $[\underline{y}, \bar{y}]$  of the expression  $f$ .

To solve this problem, we take all possible selections  $1 \leq i_1 < \dots < i_j < \dots < i_k \leq n$  of  $k$  different indices. There are  $O(n^k)$  such selections. For each selection, we solve a system of  $k$  linear equations with  $k$  unknowns  $S_1, \dots, S_k$ :

$$\sum_{j'=1}^k e_{j',i_j} \cdot S_{j'} + c_{i_j} = 0, \quad j = 1, \dots, k.$$

We then consider all  $3^k$  possible divisions of the set  $\{1, \dots, k\}$  into 3 subsets  $L$  (lower),  $U$  (upper), and  $I$  (inside). For each division:

- we set  $x_i = \underline{x}_i$  if  $e_i \cdot S < 0$ ;
- we set  $x_i = \bar{x}_i$  if  $e_i \cdot S > 0$ ;
- we set  $x_{i_j} = \underline{x}_{i_j}$  for  $j \in L$  and  $x_{i_j} = \bar{x}_{i_j}$  for  $j \in U$ ;
- we find the remaining values  $x_{i_j}$  for  $j \in I$ , from the system of equations:

$$\frac{\partial f}{\partial x_{i_j}} = 2 \sum_{j'=1}^k \lambda_{j'} \cdot \left( \sum_{\ell=1}^n e_{j',\ell} \cdot x_\ell \right) \cdot e_{j',i_j} + c_{i_j} = 0, \quad j = 1, \dots, k.$$

If the resulting values  $x_{i_j}$  are in  $[\underline{x}_{i_j}, \bar{x}_{i_j}]$ , then we compute the value  $f(x_1, \dots, x_n)$ .

The largest of the corresponding values of the expression  $f$  is  $\bar{y}$ . Computing  $f$  by using eigenvectors takes time  $O(n \cdot k) = O(n)$ . We perform it for all  $O(n^k) \cdot 2 \cdot 3^k = O(n^k)$  cases, so overall time is  $O(n^{k+1})$ , which is feasible.

## 7 General Case

For each  $\delta > 0$ , we can add  $\delta$ -small random changes to the values  $c_{ij}$  and  $c_i$ . For example, we can add values uniformly distributed on the interval  $[-\delta, \delta]$ . With probability 1, the resulting system is non-degenerate.

The difference between the original and new objective functions does not exceed

$$\delta \cdot \left( \sum_{i=1}^n \sum_{j=1}^n |x_i| \cdot |x_j| + \sum_{i=1}^n |x_i| \right).$$

We can use straightforward interval computations (see, e.g., see, e.g., [2, 4, 5]) to get the bound  $B$  on the expression in parentheses. So, for any given  $\varepsilon > 0$ , if we take  $\delta = \varepsilon/B$ , we get a non-degenerate objective function which is  $\varepsilon$ -close to the original one. The bounds for the new objective function are  $\varepsilon$ -close to the bounds on the original one.

Thus, we have a feasible  $O(n^{k+1})$  algorithm for computing  $\underline{y}$  and  $\bar{y}$  with any given accuracy  $\varepsilon > 0$ .

## 8 Homogeneous Case

In the Fourier transform case,  $c_i = 0$ , so  $f = \sum_{i=1}^n \sum_{j=1}^n c_{i,j} \cdot x_i \cdot x_j + c_0$ . In such *homogeneous* case, we can consider  $k$ -dimensional vectors

$$e_i = (e_{1,i}, \dots, e_{k,i}) \text{ and } e_i^* = (2\lambda_1 \cdot e_{1,i}, \dots, 2\lambda_k \cdot e_{k,i}).$$

In non-degenerate case, we thus have  $\leq k - 1$  indices  $i$  at which the derivative is 0. So, we have a similar algorithm, but with  $k - 1$  instead of  $k$ . This algorithm requires time  $O(n^k)$ .

**Acknowledgements** This work was supported in part by the National Science Foundation grants: • 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and • HRD-1834620 and HRD-2034030 (CAHSI Includes). It was also supported: • by the AT&T Fellowship in Information Technology, and • by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478. The authors are thankful to all the participants of the 26th Annual UTEP/NMSU Workshop on Mathematics, Computer Science, and Computational Science (El Paso, Texas, November 5, 2021) for valuable discussions.

## References

1. De Angelis, M., Behrendt, M., Comerford, L., Zhang, Y., Beer, M.: Forward Interval Propagation through the Discrete Fourier Transform. In: A. Sofi, G. Muscolino, and R. L. Muhanna, Proceedings of the International Workshop on Reliable Engineering Computing REC'2021, Taormina, Italy, May 17–20, 2021, pp. 39–52
2. Jaulin, L., Kiefer, M., Didrit, O., Walter, E.: Applied Interval Analysis, with Examples in Parameter and State Estimation, Robust Control, and Robotics. Springer, London (2001)
3. Kreinovich, V., Lakeyev, A., Rohn, J., Kahl, P.: Computational Complexity and Feasibility of Data Processing and Interval Computations. Kluwer, Dordrecht (1998)
4. Mayer, G.: Interval Analysis and Automatic Result Verification. de Gruyter, Berlin (2017)
5. Moore, R.E., Kearfott, R.B., Cloud, M.J.: Introduction to Interval Analysis. SIAM, Philadelphia (2009)

# B-Matrices and Doubly B-Matrices in the Interval Setting



Matyáš Lorenc

**Abstract** In this paper we focus on generalizing B-matrices and doubly B-matrices into interval setting, including some results regarding these classes. By interval B-matrix, or doubly B-matrix we understand such an interval matrix, whose all instances are B-matrices, or doubly B-matrices respectively. We derive mainly means of recognition for their interval variants, such as characterizations, necessary conditions and sufficient ones.

**Keywords** B-matrix · Doubly B-matrix · Interval analysis · Interval matrix · P-matrix

## 1 Introduction

P-matrices are defined as those matrices, whose all principal minors are positive. They have a close connection to so called linear complementarity problem that is more thoroughly described in [1], which is one of the reasons they are studied. A connection has even been found between P-matrices and regularity of interval matrices, as shown in [4] or [11]. However, the task of verification whether a given matrix belongs to the class of P-matrices is co-NP-complete, as shown in [2]. This leads us to try to define some subclasses of P-matrices that are easily recognizable. Such classes are e.g. B-matrices (introduced in [9]) and doubly B-matrices (introduced in [10]). What more, the B-matrices and doubly B-matrices found their use in localization of eigenvalues, as shown in [9, 10].

In this work we will present some results based on [8], such as generalization of our special subclasses of P-matrices into interval settings and we will lay the foundations for recognizing the interval variants through characterization, or sufficient conditions or necessary ones.

---

M. Lorenc (✉)

Charles University, Faculty of Mathematics and Physics, Department of Applied Mathematics, Malostranské nám. 25, 11800 Prague, Czech Republic  
e-mail: [lorenc@kam.mff.cuni.cz](mailto:lorenc@kam.mff.cuni.cz)

First, let us note that by  $\mathbb{IR}$  we denote the set of all real intervals and then let us take a look at what we mean by interval matrix.

**Definition 1** (interval matrix) An *interval matrix*  $A$ , which we denote by  $A \in \mathbb{IR}^{m \times n}$ , is defined as

$$A = [\underline{A}, \overline{A}] = \{ A \in \mathbb{R}^{m \times n} \mid \underline{A} \leq A \leq \overline{A} \},$$

where  $\underline{A}, \overline{A}$  are called lower, or upper bound matrices of  $A$  respectively, and  $\leq$  is understood entrywise.

We can as well look at  $A$  as matrix, which has entries from  $\mathbb{IR}$ , hence  $\forall i \in [m], \forall j \in [n] : a_{ij} = [\underline{a}_{ij}, \overline{a}_{ij}]$ , where  $[n] = \{1, 2, \dots, n\}$  and analogously for  $[m]$ .

An interval matrix  $A \in \mathbb{IR}^{n \times n}$  is called an interval P-matrix if every  $A \in A$  is a P-matrix. Similarly we define interval B-matrices and interval doubly B-matrices, or e.g. even a class of Z-matrices, which are matrices with non-positive off-diagonal elements. In this manner we can even define some basic properties, such as regularity, which is again more thoroughly studied in the following works: [3, 5–7], among many others.

## 2 B-Matrices

### 2.1 Real B-Matrices

Let us start by introducing real B-matrices and a few facts about them, which are introduced by Peña in [9]. Then we proceed to state and prove one new fact.

**Definition 2** (B-matrix) Let  $A \in \mathbb{R}^{n \times n}$ . We say that  $A$  is a *B-matrix*, if  $\forall i \in [n]$  the following holds:

- (a)  $\sum_{j=1}^n a_{ij} > 0$
- (b)  $\forall k \in [n] \setminus \{i\} : \frac{1}{n} \sum_{j=1}^n a_{ij} > a_{ik}$

**Remark 1** From Definition 2 it can be deduced that every B-matrix  $A$  must fulfill following condition for all  $i \in [n]$ :

$$a_{ii} > r_i^+,$$

where  $r_i^+ = \max\{0, a_{ij} \mid j \neq i\}$ .

**Proposition 1** *B-matrices are P-matrices as well.*



**Proposition 2** Let  $A \in \mathbb{R}^{n \times n}$ . It holds that  $A$  is a B-matrix if and only if  $\forall i \in [n]$  the following holds:

$$\sum_{j=1}^n a_{ij} > n \cdot r_i^+$$

**Proposition 3** Let  $A \in \mathbb{R}^{n \times n}$ . It holds that  $A$  is a B-matrix if and only if  $\forall i \in [n]$  the following holds:

$$a_{ii} - r_i^+ > \sum_{j \neq i} (r_i^+ - a_{ij})$$

**Proposition 4** Let  $A \in \mathbb{R}^{n \times n}$ . If  $A$  is a Z-matrix, then the following is equivalent:

- (1)  $A$  is a B-matrix,
- (2) The row sums are positive.
- (3)  $A$  is strictly diagonally dominant by rows with positive diagonal entries.

**Proposition 5** Let us have two B-matrices  $A, B \in \mathbb{R}^{n \times n}$  and let  $C \in \mathbb{R}^{n \times n}$ . If  $C$  satisfies the following:

$$\forall i \in [n] : C_{i*} = A_{i*} \quad \vee \quad C_{i*} = B_{i*},$$

then  $C$  is a B-matrix.

**Proof** In Definition 2 we can see that there are no conditions intertwining the rows. So by combining some rows, which satisfy the conditions, while ensuring that elements that were diagonal still are, then we get a B-matrix.

## 2.2 Interval B-Matrices

Now we will progress to generalize the class of B-matrices into the interval setting.

The interval B-matrices are defined just as is mentioned above, at the end of Sect. 1, but that definition gives us little to verify whether a given matrix is an interval B-matrix, therefore we formulate the following characterization:

**Theorem 1** Let  $A \in \mathbb{IR}^{n \times n}$ . It holds that  $A$  is an interval B-matrix if and only if  $\forall i \in [n]$  the following two properties hold:

$$(a) \quad \sum_{j=1}^n \underline{a}_{ij} > 0$$

$$(b) \quad \forall k \in [n] \setminus \{i\} : \sum_{j \neq k} \underline{a}_{ij} > (n - 1) \cdot \bar{a}_{ik}$$

**Proof** As shown in Definition 2, square real matrix  $A$  is a B-matrix, if for its every row  $i$  holds that the row sums are positive (marked as condition (a)) and every non-diagonal element of the  $i$ th row is bounded above by the corresponding row mean ((b) condition).

The (a) condition is surely satisfied  $\forall A \in \mathbf{A}$ , because of the (a) condition of Proposition, whereas the (a) condition of Proposition always holds true for an interval B-matrix  $\mathbf{A}$  because  $\underline{A} \in \mathbf{A}$ , thus  $\underline{A}$  is a B-matrix and fulfills the condition (a) of Definition 2.

Now let's take a look at conditions (b). The (b) condition of the Definition 2 can be for every  $k \neq i$  rewritten as follows:

$$\frac{1}{n} \sum_{j=1}^n a_{ij} > a_{ik} \iff \sum_{j \neq k} a_{ij} > (n - 1) \cdot a_{ik}$$

In the last inequality, we can see there is no element twice. Consequently, if we use intervals in this inequality, by substitution (of specific values from the intervals) we obtain exact values, not a superset. So now we can see that the condition (b) of real case applies for every  $A \in \mathbf{A}$  iff it holds for  $\underline{a}_{ij}$  on the left side and  $\bar{a}_{ik}$  on the right side, which is exactly the (b) condition of Proposition.

This characterization has time complexity  $O(n^2)$ , which is the same as the complexity of the characterization for real case from Definition 2.

Now let us take a look at a few properties of interval B-matrices that are rather direct corollaries of this characterization.

**Corollary 1** *Let  $A \in \mathbb{IR}^{n \times n}$ . Then  $\mathbf{A}$  is an interval B-matrix iff  $\mathbf{A}$  with the diagonal fixed on lower bounds ( $a_{ii} = \underline{a}_{ii}$ ) is an interval B-matrix.*

**Proof** In the characterization given in Theorem 1 we see that every time any  $a_{ii}$  occurs, it occurs in form of  $\underline{a}_{ii}$ , hence we are not interested in any other value of  $a_{ii}$ . (So the reduced matrix has to fulfill exactly the same conditions as the matrix  $\mathbf{A}$ .)

**Corollary 2** *Let  $\mathbf{A} \in \mathbb{IR}^{n \times n}$  and let*

$$S = \left\{ (i, j) \mid i, j \in [n] : \exists k \in [n] \setminus \{i, j\} : \underline{a}_{ij} \leq \underline{a}_{ik} \wedge \bar{a}_{ij} \leq \bar{a}_{ik} \right\}.$$

*We have that  $\mathbf{A}$  is an interval B-matrix iff  $\mathbf{A}$  with every element, whose indices are in  $S$ , set to its lower bound ( $\forall (i, j) \in S : a_{ij} = \underline{a}_{ij}$ ) is an interval B-matrix.*

**Proof** The only time, when for every  $i$  and  $j \neq i$  the  $\bar{a}_{ij}$  occurs in Theorem 1, is the (b) condition. Let us show that in the case that  $(i, j) \in S$  this condition is not necessary and is substituted by one of the others.

Let  $(i, j) \in S$  arbitrary and let  $k \in [n] \setminus \{i\} : \underline{a}_{ij} \leq \underline{a}_{ik} \wedge \bar{a}_{ij} \leq \bar{a}_{ik}$ . Because  $(i, j) \in S$ , then surely such  $k$  exists. Then:

$$\sum_{m \neq j} \underline{a}_{im} \geq \sum_{m \neq k} \underline{a}_{im} > (n - 1) \cdot \bar{a}_{ik} \geq (n - 1) \cdot \bar{a}_{ij}$$

The first inequality is obtained from  $\underline{a}_{ij} \leq \underline{a}_{ik}$  and the third from  $\bar{a}_{ij} \leq \bar{a}_{ik}$ . The second one holds, if condition (b) holds for  $(i, k)$ , so we see that if the condition holds for  $(i, k)$ , then it holds for  $(i, j)$  as well. Thus the implication “ $\Leftarrow$ ” holds.

The second implication is trivial, because  $A$  is a superset of the reduced matrix.

Although the next corollary is obtained rather straightforwardly from the previous theorem, we will state it, as it will prove to be a useful step in the derivation of other characterizations of interval B-matrices.

**Corollary 3** *Let  $A \in \mathbb{IR}^{n \times n}$ . It holds that  $A$  is an interval B-matrix if and only if  $\forall i \in [n]$  the following holds:*

$$\forall k \in [n] \setminus \{i\} : \sum_{j=1}^n \underline{a}_{ij} > \max\{0, (n - 1) \cdot \bar{a}_{ik} + \underline{a}_{ik}\}$$

**Proof** “ $\Rightarrow$ ”

$A$  is interval B-matrix, so  $A$  satisfies both conditions from Theorem 1. Thus for arbitrary  $k \neq i$ :

$$\sum_{j \neq k} \underline{a}_{ij} > (n - 1) \cdot \bar{a}_{ik} \Leftrightarrow \sum_{j=1}^n \underline{a}_{ij} > (n - 1) \cdot \bar{a}_{ik} + \underline{a}_{ik}$$

And combined with condition (a) from Theorem 1 we get that this implication clearly holds.

“ $\Leftarrow$ ”

We will show that if matrix fulfills condition stated in this corollary, then it fulfills the conditions of Theorem 1 as well. The condition (a) holds trivially. As for the (b) condition:

$\forall k \neq i :$

$$\begin{aligned} \sum_{j=1}^n \underline{a}_{ij} > \max\{0, (n - 1) \cdot \bar{a}_{ik} + \underline{a}_{ik}\} &\geq (n - 1) \cdot \bar{a}_{ik} + \underline{a}_{ik} \Rightarrow \\ \Rightarrow \sum_{j \neq k} \underline{a}_{ij} > (n - 1) \cdot \bar{a}_{ik} \end{aligned}$$

So the (b) condition holds too. Thus this implication also holds.

By realignment of the previous corollary, we get the subsequent one.

**Corollary 4** Let  $A \in \mathbb{IR}^{n \times n}$ . It holds that  $A$  is an interval B-matrix if and only if  $\forall i \in [n]$  the following holds:

$$(a) \sum_{j=1}^n \underline{a}_{ij} > 0$$

$$(b) \forall k \in [n] \setminus \{i\} : \underline{a}_{ii} - \underline{a}_{ik} > \sum_{\substack{j \neq i \\ j \neq k}} (\bar{a}_{ik} - \underline{a}_{ij})$$

**Proof** Obtained by realignment of inequalities from Corollary 3.

Or we might realign it in a different way and obtain our future connection to interval doubly B-matrices. (Even though there is quite a clear bond through the real cases, which translates pretty straightforwardly into interval setting, we believe that this corollary illustrates this connection even more.)

**Corollary 5** Let  $A \in \mathbb{IR}^{n \times n}$ . It holds that  $A$  is an interval B-matrix if and only if  $\forall i \in [n]$  the following holds:

$$(a) \sum_{j=1}^n \underline{a}_{ij} > 0$$

$$(b) \forall k \in [n] \setminus \{i\} : \underline{a}_{ii} - \bar{a}_{ik} > \sum_{\substack{j \neq i \\ j \neq k}} (\bar{a}_{ik} - \underline{a}_{ij})$$

Next we should mention two simple corollaries of the definition of interval B-matrices:

**Corollary 6** Every interval B-matrix is an interval P-matrix.

**Proof** It holds for every instance, hence it holds for whole interval matrix.

**Corollary 7** Let us have two interval B-matrices  $A, B \in \mathbb{R}^{n \times n}$  and let  $C \in \mathbb{R}^{n \times n}$ . If  $C$  satisfies the following:

$$\forall i \in [n] : C_{i*} = A_{i*} \vee C_{i*} = B_{i*},$$

then  $C$  is an interval B-matrix.

**Proof** From our definition and from Proposition 5 we can see that it holds for every instance, thus it holds for whole interval matrix.

Now let us generalize Remark 1.

**Remark 2** Because for every interval B-matrix  $A$  holds that  $\forall A \in \mathbf{A}$ :  $A$  is a B-matrix, thus even matrix  $A'$ , defined as

$$a'_{ij} = \begin{cases} \underline{a}_{ii} & \text{if } i = j, \\ \bar{a}_{ij} & \text{otherwise.} \end{cases}$$

is a B-matrix, thus it must hold (from Remark 1) that

$$\forall i \in [n] : \underline{a}_{ii} > \max\{0, \bar{a}_{ij} | j \neq i\}.$$

Let us finish this section about B-matrices by stating a sufficient condition for interval Z-matrices and a properties the entries of interval B-matrices must fulfill.

**Proposition 6** *Let  $A \in \mathbb{IR}^{n \times n}$  be an interval Z-matrix. Then the following are equivalent:*

- (1)  $A$  is an interval B-matrix,
- (2)  $\forall i \in [n] : \sum_{j=1}^n \underline{a}_{ij} > 0$ ,
- (3)  $\forall i \in [n] : \underline{a}_{ii} > \sum_{j \neq i} |\underline{a}_{ij}|$ .
- (4)  $\underline{A}$  is a B-matrix.

**Proof** “(1)  $\Rightarrow$  2)”: From Theorem 1  
 “(2)  $\Leftrightarrow$  (3)  $\Leftrightarrow$  (4)”: From Proposition 4.  
 “(3)  $\Rightarrow$  (1)”:  $\forall A \in \mathbf{A} : \forall i \in [n]$ :

$$a_{ii} \geq \underline{a}_{ii} \quad \wedge \quad \forall j \in [n] \setminus \{i\} : |a_{ij}| \leq |\underline{a}_{ij}|$$

$\Rightarrow$

$$a_{ii} \geq \underline{a}_{ii} > \sum_{j \neq i} |\underline{a}_{ij}| \geq \sum_{j \neq i} |a_{ij}| \geq 0$$

So  $A$  is strictly diagonally dominant with positive diagonal. Thus, according to Proposition 4,  $A$  is a B-matrix.

**Proposition 7** *Let  $A \in \mathbb{IR}^{n \times n}$  be an interval B-matrix. Then  $\forall i \in [n]$  the following two properties hold:*

- (1)  $\underline{a}_{ii} > \sum_{j \in S} |\underline{a}_{ij}|$ , where  $S = \{j \in [n] \mid \underline{a}_{ij} < 0\}$  and
- (2)  $\forall j \in [n] \setminus \{i\} : \underline{a}_{ii} > \max\{|\bar{a}_{ij}|, |\underline{a}_{ij}|\}$ .

**Proof** (1) Let us distinguish the following two cases for arbitrary  $i \in [n]$ :

- I.  $\forall j \in [n] \setminus \{i\} : \underline{a}_{ij} \leq 0$   
 Then it follows directly from Theorem 1, condition (a). (Because it holds that  $\forall j \in [n] \setminus \{i\} : -\underline{a}_{ij} = |\underline{a}_{ij}|$ .)
- II.  $\exists j \in [n] \setminus \{i\} : \underline{a}_{ij} > 0$

Let us take  $k \in \operatorname{argmax} \{ \underline{a}_{ij} \mid j \neq i \}$ . Then, according to Corollary 4, the following applies:

$$\underline{a}_{ii} - \underline{a}_{ik} > \sum_{j \neq i} (\bar{a}_{ik} - \underline{a}_{ij}).$$

And because

$$\underline{a}_{ii} > \underline{a}_{ii} - \underline{a}_{ik} \quad \wedge \quad \forall j \neq i : \bar{a}_{ik} - \underline{a}_{ij} \geq 0$$

(because of the presumption of this case and definition of  $k$ ), then

$$\underline{a}_{ii} > \underline{a}_{ii} - \underline{a}_{ik} > \sum_{j \neq i} (\bar{a}_{ik} - \underline{a}_{ij}) \geq \sum_{j \in S} (\bar{a}_{ik} - \underline{a}_{ij}) > \sum_{j \in S} -\underline{a}_{ij} = \sum_{j \in S} |\underline{a}_{ij}|.$$

(2) For arbitrary  $j \neq i$ , let us distinguish three cases:

- I.  $|\underline{a}_{ij}| > |\bar{a}_{ij}|$   
 $\Rightarrow \underline{a}_{ij} \leq 0$ , thus from property 1. of this proposition:

$$\underline{a}_{ii} > \sum_{k \in S} |\underline{a}_{ik}| \geq |\underline{a}_{ij}|,$$

because  $j \in S$ .

- II.  $|\bar{a}_{ij}| > |\underline{a}_{ij}|$   
 $\Rightarrow \bar{a}_{ij} > 0$ , thus from Remark 2  $\Rightarrow \underline{a}_{ii} > \bar{a}_{ij} = |\bar{a}_{ij}|$ .
- III.  $|\underline{a}_{ij}| = |\bar{a}_{ij}|$

Then we have either a degenerated interval ( $\underline{a}_{ij} = \bar{a}_{ij}$ ), or it holds that  $\underline{a}_{ij} = -\bar{a}_{ij}$ . For both it holds that  $\underline{a}_{ij} \leq 0 \vee \bar{a}_{ij} > 0$ . Therefore we can use the same argumentation as in the first two cases with sharp inequalities.

Hence both properties holds.

### 3 Doubly B-Matrices

#### 3.1 Real Doubly B-Matrices

Let us start by introducing real doubly B-matrices and a few facts about them, some are showed and proved by Peña in [10], most of them will be proved here.

**Definition 3** (Doubly B-matrix) Let  $A \in \mathbb{R}^{n \times n}$ . We say that  $A$  is a *doubly B-matrix*, if  $\forall i \in [n]$  the following holds:

(a)  $a_{ii} > r_i^+$

(b)  $\forall j \neq i : (a_{ii} - r_i^+) (a_{jj} - r_j^+) > \left( \sum_{k \neq i} (r_i^+ - a_{ik}) \right) \left( \sum_{k \neq j} (r_j^+ - a_{jk}) \right)$ .

**Proposition 8** *Let  $A \in \mathbb{R}^{n \times n}$ . If  $A$  is a B-matrix, then  $A$  is a doubly B-matrix.*

**Proof** If  $A$  is a B-matrix, then, from Proposition 3, the following holds for every  $i \in [n]$ :

$$a_{ii} - r_i^+ > \sum_{j \neq i} (r_i^+ - a_{ij}) \geq 0$$

From that follows that  $\forall i, j \in [n], j \neq i$ :

$$(a_{ii} - r_i^+) (a_{jj} - r_j^+) > \left( \sum_{k \neq i} (r_i^+ - a_{ik}) \right) \left( \sum_{k \neq j} (r_j^+ - a_{jk}) \right),$$

which is exactly the (b) part of the Definition 3. The (a) part of the definition is obtained from Remark 1. Therefore  $A$  is a doubly B-matrix.

**Remark 3** We can now show that in general the opposite implication does not hold. We can take e.g. matrix

$$A = \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix}$$

as a counterexample. The matrix  $A$  is a doubly B-matrix, but it is not a B-matrix.

**Proposition 9** *Doubly B-matrices are P-matrices as well.*

**Proposition 10** *Let  $A \in \mathbb{R}^{n \times n}$ . If  $A$  is a doubly B-matrix, then exactly one of the following applies:*

- (a) *Either  $A$  is a B-matrix, or*
- (b) *there exists a unique  $j \in [n]$  that*

$$a_{jj} - r_j^+ \leq \sum_{m \neq j} (r_j^+ - a_{jm})$$

*and for every other  $i \in [n] \setminus \{j\}$ :*

$$a_{ii} - r_i^+ > \sum_{m \neq i} (r_i^+ - a_{im}).$$

*(I.e. there is only one row that does not satisfy the condition stated in Corollary 3.)*

**Proof** Let (a) hold, so  $A$  is a B-matrix and thus from Corollary 3  $\forall i \in [n]$ :

$$a_{ii} - r_i^+ > \sum_{m \neq i} (r_i^+ - a_{im}),$$

thus (b) does not hold.

Now let (a) not apply, so  $A$  is not a B-matrix. Then it contains a row, which does not fulfill the condition stated in Corollary 3. (Otherwise it would fulfill the characterization stated ibidem, thus it would be a B-matrix, hence we obtain a contradiction.) We will show that there cannot exist two such rows.

For contradiction, let there be two such rows  $j$  and  $j'$  that

$$a_{jj} - r_j^+ \leq \sum_{m \neq j} (r_j^+ - a_{jm})$$

and

$$a_{j'j'} - r_{j'}^+ \leq \sum_{m \neq j'} (r_{j'}^+ - a_{j'm}).$$

(It should be noted that because  $A$  is a doubly B-matrix, then from Definition 3 we get that  $0 < a_{jj} - r_j^+$  and  $0 < a_{j'j'} - r_{j'}^+$ .) Then

$$(a_{jj} - r_j^+) (a_{j'j'} - r_{j'}^+) \leq \left( \sum_{m \neq j} (r_j^+ - a_{jm}) \right) \left( \sum_{m \neq j'} (r_{j'}^+ - a_{j'm}) \right),$$

but that leads us to a contradiction with the Definition 3, because  $A$  should have been a doubly B-matrix. Therefore such a row which breaks the condition from Corollary 3 is exactly one, all the others have to satisfy this condition.

The next proposition shows us easily testable class of both B- and doubly B-matrices.

**Proposition 11** *Let  $A \in \mathbb{R}^{n \times n}$ . If  $A$  is a circulant matrix, then the following are equivalent:*

- (1)  $A$  is a B-matrix.
- (2)  $A$  is a doubly B-matrix.
- (3)  $a_{11} - r_1^+ > \sum_{j \neq 1} (r_1^+ - a_{1j})$

**Proof** “(1)  $\Rightarrow$  (2)”: See Proposition 8.

“(2)  $\Rightarrow$  (3)”:  $A$  is a doubly B-matrix, hence for arbitrary  $j \neq 1$ :



$$\begin{aligned}
 (a_{11} - r_1^+) (a_{jj} - r_j^+) &> \left( \sum_{k \neq 1} (r_1^+ - a_{1k}) \right) \left( \sum_{k \neq j} (r_j^+ - a_{jk}) \right) \Leftrightarrow \\
 \Leftrightarrow (a_{11} - r_1^+)^2 &> \left( \sum_{k \neq 1} (r_1^+ - a_{1k}) \right)^2 \Leftrightarrow \\
 \Leftrightarrow (a_{11} - r_1^+) &> \left( \sum_{k \neq 1} (r_1^+ - a_{1k}) \right)
 \end{aligned}$$

The first equivalence holds, because the  $A$  is circulant, whereas the second one comes from the fact that both sides of the resulting inequality are non-negative, which is based on following:

For left side:  $A$  is doubly B-matrix  $\Rightarrow \forall i \in [n] : a_{ii} > r_i^+$  (From condition  $a$ ) of Definition 3.)

For right side: From definition of  $r_i^+ : \forall i \in [n] \forall j \neq i : r_i^+ \geq a_{ij}$ .

Therefore the implication holds.

“(3)  $\Rightarrow$  (1)”: Because  $A$  is circulant, the following implication holds:

$$a_{11} - r_1^+ > \sum_{k \neq 1} (r_1^+ - a_{1k}) \Rightarrow a_{ii} - r_i^+ > \sum_{k \neq i} (r_i^+ - a_{ik})$$

Thus from Proposition 3  $A$  is a B-matrix.

### 3.2 Interval Doubly B-Matrices

Now we shall proceed to generalize the class of doubly B-matrices into the interval setting.

The interval doubly B-matrices are defined as mentioned above, at the end of Sect. 1, but again that definition gives us no tool to check whether a given matrix belongs to the class of interval doubly B-matrices, hence we introduce the following characterization. But first, let us state direct corollary of the definition of interval doubly B-matrices:

**Corollary 8** Every interval doubly B-matrix is an interval P-matrix.

**Proof** It holds for every instance, thus it holds for whole interval matrix.

**Theorem 2** Let  $A \in \mathbb{IR}^{n \times n}$ . Then  $A$  is an interval doubly B-matrix if and only if the following two properties hold:

- (a)  $\forall i \in [n] : \underline{a}_{ii} > \max\{0, \bar{a}_{ij} | j \neq i\}$  and
- (b)  $\forall i, j \in [n], j \neq i, \forall k, l \in [n], k \neq i, l \neq j:$

$$\begin{aligned}
 & I. (\underline{a}_{ii} - \bar{a}_{ik}) \cdot (\underline{a}_{jj} - \bar{a}_{jl}) > \\
 & \quad \left( \max \left\{ 0, \sum_{\substack{m \neq i \\ m \neq k}} (\bar{a}_{ik} - \underline{a}_{im}) \right\} \right) \cdot \left( \max \left\{ 0, \sum_{\substack{m \neq j \\ m \neq l}} (\bar{a}_{jl} - \underline{a}_{jm}) \right\} \right) \\
 & II. \underline{a}_{ii} \cdot (\underline{a}_{jj} - \bar{a}_{jl}) > \left( \max \left\{ 0, - \sum_{m \neq i} \underline{a}_{im} \right\} \right) \cdot \left( \max \left\{ 0, \sum_{\substack{m \neq j \\ m \neq l}} (\bar{a}_{jl} - \underline{a}_{jm}) \right\} \right) \\
 & III. \underline{a}_{ii} \cdot \underline{a}_{jj} > \left( \max \left\{ 0, - \sum_{m \neq i} \underline{a}_{im} \right\} \right) \cdot \left( \max \left\{ 0, - \sum_{m \neq j} \underline{a}_{jm} \right\} \right)
 \end{aligned}$$

**Proof** “ $\Rightarrow$ ”  $\forall A \in \mathbf{A}$ :  $A$  is (real) doubly B-matrix, hence:

Our “interval” condition (a) holds because of “real” condition (a) from Definition 3 for matrix  $A' \in \mathbf{A}$  with all diagonal elements set on their lower bounds and all the off-diagonal elements set on their upper bounds.

As for “interval” condition (b) let us fix arbitrary  $i, j \in [n], j \neq i$ , and arbitrary  $k \neq i, l \neq j$ . Then:

I. Let  $A \in \mathbf{A}$ , such that

$$a_{m_1 m_2} = \begin{cases} \bar{a}_{ik} & \text{if } (m_1, m_2) = (i, k), \\ \bar{a}_{jl} & \text{if } (m_1, m_2) = (j, l), \\ \underline{a}_{m_1 m_2} & \text{otherwise.} \end{cases}$$

Then for this  $A$ :

$$\begin{aligned}
 & (\underline{a}_{ii} - \bar{a}_{ik})(\underline{a}_{jj} - \bar{a}_{jl}) \geq (a_{ii} - r_i^+)(a_{jj} - r_j^+) > \\
 & > \left( \sum_{m \neq i} (r_i^+ - a_{im}) \right) \left( \sum_{m \neq j} (r_j^+ - a_{jm}) \right) \geq \\
 & \geq \left( \max \left\{ 0, \sum_{\substack{m \neq i \\ m \neq k}} (\bar{a}_{ik} - \underline{a}_{im}) \right\} \right) \left( \max \left\{ 0, \sum_{\substack{m \neq j \\ m \neq l}} (\bar{a}_{jl} - \underline{a}_{jm}) \right\} \right)
 \end{aligned}$$

The first inequality comes trivially from  $\bar{a}_{ik} \leq r_i^+$  and analogically for  $j$  and  $l$ . The second one is obtained from the fact that  $A$  is a doubly B-matrix (because  $A \in \mathbf{A}$  and  $A$  is an interval doubly B-matrix). The third and last inequality is a direct result of the following facts:  $\forall m \neq i : a_{im} \leq r_i^+$  (from the definition of  $r_i^+$ ), so  $r_i^+ - a_{im} \geq 0$ , thus whole  $\sum_{m \neq i} (r_i^+ - a_{im})$  is non-negative. Another fact is that what we drop from the sum, i.e. the “ $k$  member”, is a non-negative element of the sum. And finally  $r_i^+ \geq a_{ik} = \bar{a}_{ik} \wedge a_{im} \geq \underline{a}_{ik}$ . Again, for  $j, l$  it is analogous.

II. Let  $A \in \mathbf{A}$ , such that

$$a_{m_1 m_2} = \begin{cases} \bar{a}_{jl} & \text{if } (m_1, m_2) = (j, l), \\ \underline{a}_{m_1 m_2} & \text{otherwise.} \end{cases}$$

Then for this  $A$ :

$$\begin{aligned} \underline{a}_{ii}(\underline{a}_{jj} - \bar{a}_{jl}) &\geq (a_{ii} - r_i^+)(a_{jj} - r_j^+) > \\ &> \left( \sum_{m \neq i} (r_i^+ - a_{im}) \right) \left( \sum_{m \neq j} (r_j^+ - a_{jm}) \right) \geq \\ &\geq \left( \max \left\{ 0, -\sum_{m \neq i} \underline{a}_{im} \right\} \right) \left( \max \left\{ 0, \sum_{\substack{m \neq j \\ m \neq l}} (\bar{a}_{jl} - \underline{a}_{jm}) \right\} \right) \end{aligned}$$

The second inequality comes from the same fact as above, i.e.  $A$  is a doubly B-matrix. And the last one holds as well because of similar reasons as above plus because, in case of “ $i$  part” of the expression, we drop  $n \cdot r_i^+$ , which is some non-negative quantity.

III. Let  $A = \underline{A} \in \mathbf{A}$ . Then for this  $A$ :

$$\begin{aligned} \underline{a}_{ii} \cdot \underline{a}_{jj} &\geq (a_{ii} - r_i^+)(a_{jj} - r_i^+) > \\ &> \left( \sum_{m \neq i} (r_i^+ - a_{im}) \right) \left( \sum_{m \neq j} (r_i^+ - a_{jm}) \right) \geq \\ &\geq \left( \max \left\{ 0, -\sum_{m \neq i} \underline{a}_{im} \right\} \right) \left( \max \left\{ 0, -\sum_{m \neq j} \underline{a}_{jm} \right\} \right) \end{aligned}$$

Again these inequalities hold from the reasons stated above.

“ $\Leftarrow$ ” Let  $A \in \mathbf{A}$ .

Condition (a) from Definition 3 follows trivially from our “interval” condition (a).

Let us pick arbitrary  $i, j \in [n], j \neq i$ . Now let us distinguish the following cases:

(1)  $r_i^+, r_j^k > 0$ : Then  $\exists k \neq i, \exists l \neq j : r_i^+ = a_{ik} \wedge r_j^k = a_{jl}$ . So the following holds:

$$\begin{aligned} (a_{ii} - r_i^+)(a_{jj} - r_j^+) &\geq (\underline{a}_{ii} - \bar{a}_{ik})(\underline{a}_{jj} - \bar{a}_{jl}) > \\ &> \left( \max \left\{ 0, \sum_{\substack{m \neq i \\ m \neq k}} (\bar{a}_{ik} - \underline{a}_{im}) \right\} \right) \left( \max \left\{ 0, \sum_{\substack{m \neq j \\ m \neq l}} (\bar{a}_{jl} - \underline{a}_{jm}) \right\} \right) = \end{aligned}$$

$$\begin{aligned}
&= \left( \sum_{\substack{m \neq i \\ m \neq k}} (\bar{a}_{ik} - \underline{a}_{im}) \right) \left( \sum_{\substack{m \neq j \\ m \neq l}} (\bar{a}_{jl} - \underline{a}_{jm}) \right) \geq \\
&\geq \left( \sum_{\substack{m \neq i \\ m \neq k}} (r_i^+ - a_{im}) \right) \left( \sum_{\substack{m \neq j \\ m \neq l}} (r_j^+ - a_{jm}) \right) = \\
&= \left( \sum_{m \neq i} (r_i^+ - a_{im}) \right) \left( \sum_{m \neq j} (r_j^+ - a_{jm}) \right)
\end{aligned}$$

The second inequality holds because  $A \in \mathbf{A}$  (thus because of the assumptions of the implication, more specifically because of point  $I$ . of condition (b)). The next equality follows from  $\bar{a}_{ik} \geq r_i^+ \geq a_{im} \geq \underline{a}_{im}$ , for  $m \neq i$ , because that implies that the sums are non-negative. (Analogically for  $j$  and  $l$ .) The same chain of inequalities can be used to verify the fourth inequality. And the last equality arises from the fact that  $r_i^+ = a_{ik} \wedge r_j^+ = a_{jl}$ . Thus from Definition 3,  $A$  is a doubly B-matrix.

(2)  $r_i^+ = 0 \wedge r_j^+ > 0$ : Then  $\exists l \neq j : r_j^k = a_{jl}$ . So the following holds:

$$\begin{aligned}
&(a_{ii} - r_i^+) (a_{jj} - r_j^k) = \\
&= a_{ii} (a_{jj} - r_j^k) \geq \underline{a}_{ii} (\underline{a}_{jj} - \bar{a}_{jl}) > \\
&> \left( \max \left\{ 0, - \sum_{m \neq i} \underline{a}_{im} \right\} \right) \left( \max \left\{ 0, \sum_{\substack{m \neq j \\ m \neq l}} (\bar{a}_{jl} - \underline{a}_{jm}) \right\} \right) = \\
&= \left( - \sum_{m \neq i} \underline{a}_{im} \right) \left( \sum_{\substack{m \neq j \\ m \neq l}} (\bar{a}_{jl} - \underline{a}_{jm}) \right) \geq \\
&\geq \left( \sum_{m \neq i} (r_i^+ - a_{im}) \right) \left( \sum_{\substack{m \neq j \\ m \neq l}} (r_j^k - a_{jm}) \right) = \\
&= \left( \sum_{m \neq i} (r_i^+ - a_{im}) \right) \left( \sum_{m \neq j} (r_j^k - a_{jm}) \right)
\end{aligned}$$

The reasoning for the “ $j$  part” of the expressions in the inequalities is the same as in the previous case, so let us focus on the “ $i$  part”: The third inequality holds,

because  $A \in \mathbf{A}$ , so the point *II.* of condition *b)* applies. The fourth equality comes from the following:  $\forall m \neq i : \underline{a}_{im} \leq a_{im} \leq r_i^+ = 0 \Rightarrow -\sum_{m \neq i} \underline{a}_{im} \geq 0$ .

Therefore again from Definition 3  $A$  is a doubly B-matrix.

- (3)  $r_i^+ > 0 \wedge r_j^+ = 0$ : By swapping  $i$  for  $j$  we get the previous case.
- (4)  $r_i^+, r_j^+ = 0$ :

$$\begin{aligned} & (a_{ii} - r_i^+) (a_{jj} - r_j^+) \geq \underline{a}_{ii} \cdot \underline{a}_{jj} > \\ & > \left( \max \left\{ 0, -\sum_{m \neq i} \underline{a}_{im} \right\} \right) \left( \max \left\{ 0, -\sum_{m \neq j} \underline{a}_{jm} \right\} \right) = \\ & = \left( -\sum_{m \neq i} \underline{a}_{im} \right) \left( -\sum_{m \neq j} \underline{a}_{jm} \right) \geq \\ & \geq \left( \sum_{m \neq i} (r_i^+ - a_{im}) \right) \left( \sum_{m \neq j} (r_j^+ - a_{jm}) \right) \end{aligned}$$

Now the logic behind this chain of inequalities is the same as in the previous cases. Thus once again from Definition 3  $A$  is a doubly B-matrix.

Thus both implications hold.

This characterization has time complexity  $O(n^4)$ , which is two orders of magnitude higher than for the real case, given the  $O(n^2)$  complexity of the characterization from Definition 3.

Now let us take a look at a few properties of interval doubly B-matrices that are rather direct corollaries of this characterization.

**Corollary 9** *Let  $A \in \mathbb{IR}^{n \times n}$ . We have that  $A$  is an interval doubly B-matrix iff  $A$  with diagonal fixed on lower bounds ( $a_{ii} = \underline{a}_{ii}$ ) is an interval doubly B-matrix.*

**Proof** In the characterization given in Theorem 2 we see that every time any  $a_{ii}$  occurs, it occurs in a form of  $\underline{a}_{ii}$ , hence we are not interested in any other value of  $a_{ii}$ . (So the reduced matrix has to fulfill exactly the same conditions as the matrix  $A$ )

**Corollary 10** *Let  $A \in \mathbb{IR}^{n \times n}$  and let*

$$S = \{(i, j) | i, j \in [n] : \exists k \in [n] \setminus \{i\} : \bar{a}_{ij} \leq \underline{a}_{ik}\}.$$

*We have that  $A$  is an interval doubly B-matrix iff  $A$  with every element, whose indices are in  $S$ , set to its lower bound ( $\forall (i, j) \in S : a_{ij} = \underline{a}_{ij}$ ) is an interval doubly B-matrix.*

**Proof** The only time, when for every  $i$  and  $k \neq i$  the  $\bar{a}_{ik}$  occurs in Theorem 2, are some of the inequalities “I.” in the (b) condition. (And symmetrically as  $(j, l)$ )

in “II.” and “III.” in the (b) condition, but that is analogous, so we will prove just the first case, where it pops up as  $(i, k)$ .) Let us show that in the case that  $(i, k) \in S$  the inequalities “I.” are not necessary to check because they are substituted by some of the others.

Let  $(i, k) \in S$  arbitrary and let  $k' = \operatorname{argmax}\{a_{im} | m \in [n] \setminus \{i\}\}$ . Because  $(i, k) \in S$ , then surely  $\bar{a}_{ik} \leq \underline{a}_{ik'}$ . (And thus even  $\underline{a}_{ik} \leq \underline{a}_{ik'}$  and  $\bar{a}_{ik} \leq \bar{a}_{ik'}$ .) Let us take any arbitrary  $j, l \in [n], l \neq j$ . Then:

$$\begin{aligned} & (\underline{a}_{ii} - \bar{a}_{ik})(\underline{a}_{jj} - \bar{a}_{jl}) \geq (\underline{a}_{ii} - \bar{a}_{ik'})(\underline{a}_{jj} - \bar{a}_{jl}) > \\ & > \left( \max \left\{ 0, \sum_{\substack{m \neq i \\ m \neq k'}} (\bar{a}_{ik'} - \underline{a}_{im}) \right\} \right) \left( \max \left\{ 0, \sum_{\substack{m \neq j \\ m \neq l}} (\bar{a}_{jl} - \underline{a}_{jm}) \right\} \right) \geq \\ & \geq \left( \max \left\{ 0, \sum_{\substack{m \neq i \\ m \neq k}} (\bar{a}_{ik} - \underline{a}_{im}) \right\} \right) \left( \max \left\{ 0, \sum_{\substack{m \neq j \\ m \neq l}} (\bar{a}_{jl} - \underline{a}_{jm}) \right\} \right) \end{aligned}$$

So the inequality “I.” for  $(i, k)$  holds only if the inequality “I.” for  $(i, k')$  holds. (The last inequality in the previous chain of inequalities holds, because  $\underline{a}_{ik} \leq \underline{a}_{ik'}$  and  $\bar{a}_{ik'} \geq \bar{a}_{ik}$ , so we subtract more and add less.)

The second implication is trivial, because  $A$  is a superset of the reduced matrix.

Now let us take look at rather straightforward corollary of our definition of interval doubly B-matrices and a generalization of Proposition 10.

**Proposition 12** *Let  $A \in \mathbb{IR}^{n \times n}$ . If  $A$  is an interval B-matrix, then it is an interval doubly B-matrix as well.*

**Proof** It holds for every instance, therefore it holds for whole interval matrix.

**Proposition 13** *Let  $A \in \mathbb{IR}^{n \times n}$  be an interval doubly B-matrix. Then exactly one of the following applies:*

- (a) *Either  $A$  is an interval B-matrix, or*
- (b) *there exists a unique  $j \in [n]$  that  $j$ -th row breaks the condition stated in Corollary 5 while for all others  $i \in [n] \setminus \{k\}$  this condition holds.*  
*In other words there exists a unique  $j$ , for which holds either*

$$\sum_{m=1}^n \underline{a}_{jm} \leq 0$$

or

$$\exists k \in [n] \setminus \{j\} : \underline{a}_{jj} - \bar{a}_{jk} \leq \sum_{\substack{m \neq j \\ m \neq k}} (\bar{a}_{jk} - \underline{a}_{jm}),$$

and for all the others  $i \in [n] \setminus \{j\}$  hold both

$$\sum_{m=1}^n \underline{a}_{im} > 0$$

and

$$\forall k \in [n] \setminus \{i\} : \underline{a}_{ii} - \bar{a}_{ik} > \sum_{\substack{m \neq i \\ m \neq k}} (\bar{a}_{ik} - \underline{a}_{im}),$$

**Proof** Let (a) hold, so  $A$  is an interval B-matrix and thus from Corollary 5  $\forall i \in [n]$ :

$$\sum_{m=1}^n \underline{a}_{im} > 0$$

and

$$\forall k \in [n] \setminus \{i\} : \underline{a}_{ii} - \bar{a}_{ik} > \sum_{\substack{m \neq i \\ m \neq k}} (\bar{a}_{ik} - \underline{a}_{im}),$$

thus (b) does not hold.

Now let (a) not apply, so  $A$  is not an interval B-matrix. Then it contains a row, which does not fulfill the condition stated in Corollary 5. (Otherwise it would fulfill the characterization stated ibidem, thus it would be an interval B-matrix  $\rightarrow$  contradiction.) We will show that there cannot exist two such rows. For contradiction, let there be two such rows  $j$  and  $j'$  that breaks the condition. Let us distinguish the following cases:

(1) Let it hold that

$$\sum_{m=1}^n \underline{a}_{jm} \leq 0$$

and

$$\sum_{m=1}^n \underline{a}_{j'm} \leq 0.$$

Then

$$\underline{a}_{jj} \leq - \sum_{m \neq j} \underline{a}_{jm} \quad \wedge \quad \underline{a}_{j'j'} \leq - \sum_{m \neq j'} \underline{a}_{j'm}$$

and because  $A$  is an interval doubly B-matrix, then

$$\forall i \in [n] : \underline{a}_{ii} > \max\{0, \bar{a}_{im} | m \neq i\} \geq 0$$

$$\left( \Rightarrow 0 < \underline{a}_{jj} \leq - \sum_{m \neq j} \underline{a}_{jm} \quad \wedge \quad 0 < \underline{a}_{j'j'} \leq - \sum_{m \neq j'} \underline{a}_{j'm} \right)$$

(see Theorem 2, part (a)) and so the following is true:

$$\underline{a}_{jj} \cdot \underline{a}_{j'j'} \leq \left( - \sum_{m \neq j} \underline{a}_{jm} \right) \left( - \sum_{m \neq j'} \underline{a}_{j'm} \right) =$$

$$= \left( \max \left\{ 0, - \sum_{m \neq j} \underline{a}_{jm} \right\} \right) \left( \max \left\{ 0, - \sum_{m \neq j'} \underline{a}_{j'm} \right\} \right)$$

But that is a contradiction with the assumption that  $A$  is an interval doubly B-matrix, because it violates the (b) condition, part III. of characterization of interval doubly B-matrices stated in Theorem 2.

(2) Let it hold that

$$\sum_{m=1}^n \underline{a}_{jm} \leq 0$$

and

$$\exists k \in [n] \setminus \{j'\} : \underline{a}_{j'j'} - \bar{a}_{j'k} \leq \sum_{\substack{m \neq j' \\ m \neq k}} (\bar{a}_{j'k} - \underline{a}_{j'm}).$$

Then

$$\underline{a}_{jj} \leq - \sum_{m \neq j} \underline{a}_{jm}$$

and because  $A$  is an interval doubly B-matrix, then

$$\forall i \in [n] : \underline{a}_{ii} > \max\{0, \bar{a}_{im} | m \neq i\}$$

$$\left( \Rightarrow 0 < \underline{a}_{jj} \leq - \sum_{m \neq j} \underline{a}_{jm} \quad \wedge \quad 0 < \underline{a}_{j'j'} - \bar{a}_{ik} \leq \sum_{\substack{m \neq j' \\ m \neq k}} (\bar{a}_{j'k} - \underline{a}_{j'm}) \right)$$

(see Theorem 2, part (a)) and so the following is true:



$$\begin{aligned} \underline{a}_{jj}(\underline{a}_{j'j'} - \bar{a}_{j'k}) &\leq \left(-\sum_{\substack{m \neq j \\ m \neq k}} \underline{a}_{jm}\right) \left(\sum_{\substack{m \neq j' \\ m \neq k}} (\bar{a}_{j'k} - \underline{a}_{j'm})\right) = \\ &= \left(\max\left\{0, -\sum_{\substack{m \neq j \\ m \neq k}} \underline{a}_{jm}\right\}\right) \left(\max\left\{0, \sum_{\substack{m \neq j' \\ m \neq k}} (\bar{a}_{j'k} - \underline{a}_{j'm})\right\}\right) \end{aligned}$$

But that is a contradiction with the assumption that  $A$  is an interval doubly B-matrix, because it violates the *b*) condition, part *II*. of characterization of interval doubly B-matrices stated in Theorem 2.

(3) Let it hold that

$$\exists k \in [n] \setminus \{j\} : \underline{a}_{jj} - \bar{a}_{jk} \leq \sum_{\substack{m \neq j \\ m \neq k}} (\bar{a}_{jk} - \underline{a}_{jm})$$

and

$$\exists k' \in [n] \setminus \{j'\} : \underline{a}_{j'j'} - \bar{a}_{j'k'} \leq \sum_{\substack{m \neq j' \\ m \neq k'}} (\bar{a}_{j'k'} - \underline{a}_{j'm}).$$

Then because  $A$  is an interval doubly B-matrix, then

$$\forall i \in [n] : \underline{a}_{ii} > \max\{0, \bar{a}_{im} | m \neq i\}$$

$$\left(\Rightarrow 0 < \underline{a}_{jj} - \bar{a}_{jk} \leq \sum_{\substack{m \neq j \\ m \neq k}} (\bar{a}_{jk} - \underline{a}_{jm}) \wedge 0 < \underline{a}_{j'j'} - \bar{a}_{j'k'} \leq \sum_{\substack{m \neq j' \\ m \neq k'}} (\bar{a}_{j'k'} - \underline{a}_{j'm})\right)$$

(see Theorem 2, part (a)) and so the following is true:

$$\begin{aligned} &(\underline{a}_{jj} - \bar{a}_{jk})(\underline{a}_{j'j'} - \bar{a}_{j'k'}) \leq \\ &\leq \left(\sum_{\substack{m \neq j \\ m \neq k}} (\bar{a}_{jk} - \underline{a}_{jm})\right) \left(\sum_{\substack{m \neq j' \\ m \neq k'}} (\bar{a}_{j'k'} - \underline{a}_{j'm})\right) = \\ &= \left(\max\left\{0, \sum_{\substack{m \neq j \\ m \neq k}} (\bar{a}_{jk} - \underline{a}_{jm})\right\}\right) \left(\max\left\{0, \sum_{\substack{m \neq j' \\ m \neq k'}} (\bar{a}_{j'k'} - \underline{a}_{j'm})\right\}\right) \end{aligned}$$

But that is a contradiction with the assumption that  $A$  is an interval doubly B-matrix, because it violates the (b) condition, part I. of characterization of interval doubly B-matrices stated in Theorem 2.

It is time to state a few necessary conditions that might help us with the verification of doubly B-matrices.

**Proposition 14** *Let  $A \in \mathbb{IR}^{n \times n}$ ,  $\forall i \in [n] : k_i \in \operatorname{argmax}\{\bar{a}_{ij} | j \neq i\}$  and let us define  $i^{A_{\max}} \in \mathbb{R}^{n \times n}$  as follows:*

$$i^{A_{\max}} = (a_{m_1 m_2}); \quad a_{m_1 m_2} = \begin{cases} \bar{a}_{m_1 k_{m_1}} & \text{if } m_1 \neq i \wedge m_2 = k_{m_1}, \\ \underline{a}_{m_1 m_2} & \text{otherwise.} \end{cases}$$

*It holds that  $A$  is an interval doubly B-matrix only if  $\underline{A}$  and  $\forall i \in [n] : i^{A_{\max}}$  are doubly B-matrices.*

**Proof** It holds that  $\underline{A} \in A \wedge \forall i \in [n] : i^{A_{\max}} \in A$ .

Proposition 14 gives us quite nice necessary condition through reduction, but to compute it, we have to verify  $n + 1$  matrices whether they are doubly B-matrices, which takes us verifying  $O(n^2)$  inequalities for each. Hence together the time complexity would be  $O(n^3)$ . So let us state an equivalent condition with better time complexity, more precisely with  $O(n^2)$  complexity.

**Proposition 15** *Let  $A \in \mathbb{IR}^{n \times n}$ . It holds that  $A$  is a doubly B-matrix only if the following hold:*

- (a)  $\forall i \in [n] : \underline{a}_{ii} > \max\{0, \bar{a}_{ij} | j \neq i\}$  and
- (b)  $\forall i, j \in [n], j \neq i, k \in \operatorname{argmax}\{\bar{a}_{im} | m \neq i\}, l \in \operatorname{argmax}\{\bar{a}_{jm} | m \neq j\}$ :

$$I. (\bar{a}_{ik} > 0 \wedge \bar{a}_{jl} > 0) \Rightarrow$$

$$(\underline{a}_{ii} - \bar{a}_{ik})(\underline{a}_{jj} - \bar{a}_{jl}) > \left( \sum_{\substack{m \neq i \\ m \neq k}} (\bar{a}_{ik} - \underline{a}_{im}) \right) \left( \sum_{\substack{m \neq j \\ m \neq l}} (\bar{a}_{jl} - \underline{a}_{jm}) \right)$$

$$II. (\bar{a}_{ik} \leq 0 \wedge \bar{a}_{jl} > 0) \Rightarrow$$

$$\underline{a}_{ii}(\underline{a}_{jj} - \bar{a}_{jl}) > \left( - \sum_{m \neq i} \underline{a}_{im} \right) \left( \sum_{\substack{m \neq j \\ m \neq l}} (\bar{a}_{jl} - \underline{a}_{jm}) \right)$$

$$III. (\bar{a}_{ik} \leq 0 \wedge \bar{a}_{jl} \leq 0) \Rightarrow$$

$$\underline{a}_{ii} \cdot \underline{a}_{jj} > \left( - \sum_{m \neq i} \underline{a}_{im} \right) \left( - \sum_{m \neq j} \underline{a}_{jm} \right)$$

**Proof** We assume that  $\forall A \in A : A$  is a doubly B-matrix. Therefore our condition (a) follows from condition (a) from the Definition 3 for  $A \in A$ :

$$A = (a_{m_1 m_2}); \quad a_{m_1 m_2} = \begin{cases} \bar{a}_{m_1 m_1} & \text{if } m_1 = m_2, \\ \underline{a}_{m_1 m_2} & \text{otherwise.} \end{cases}$$

To prove condition (b), let us take arbitrary  $i, j \in [n], j \neq i$  and let  $k \in \operatorname{argmax}\{\bar{a}_{im} | m \neq i\}, l \in \operatorname{argmax}\{\bar{a}_{jm} | m \neq j\}$ . Then:

I. Let  $\bar{a}_{ik} > 0, \bar{a}_{jl} > 0$ . Let us take such  $A \in \mathbf{A}$ , that

$$A = (a_{m_1 m_2}); \quad a_{m_1 m_2} = \begin{cases} \bar{a}_{ik} & \text{if } (m_1, m_2) = (i, k), \\ \bar{a}_{jl} & \text{if } (m_1, m_2) = (j, l), \\ \underline{a}_{m_1 m_2} & \text{otherwise.} \end{cases}$$

Then  $r_i^+ = \bar{a}_{ik}, r_i^+ = \bar{a}_{jl}$ , so the following holds:

$$\begin{aligned} (\underline{a}_{ii} - \bar{a}_{ik})(\underline{a}_{jj} - \bar{a}_{jl}) &= (a_{ii} - r_i^+)(a_{jj} - r_j^+) > \\ > \left( \sum_{m \neq i} (r_i^+ - a_{im}) \right) \left( \sum_{m \neq j} (r_j^+ - a_{jm}) \right) = \\ &= \left( \sum_{\substack{m \neq i \\ m \neq k}} (\bar{a}_{ik} - \underline{a}_{im}) \right) \left( \sum_{\substack{m \neq j \\ m \neq l}} (\bar{a}_{jl} - \underline{a}_{jm}) \right) \end{aligned}$$

The last equality arises from the fact that  $a_{ik} = r_i^+$ , so  $r_i^+ - a_{ik} = 0$ , and that  $\forall m \neq i : r_i^+ \geq a_{im}$ . (And of course analogies of that hold for  $j$  as well.)

II. Let  $\bar{a}_{ik} \leq 0, \bar{a}_{jl} > 0$ . Let us take such  $A \in \mathbf{A}$ , that

$$A = (a_{m_1 m_2}); \quad a_{m_1 m_2} = \begin{cases} \bar{a}_{jl} & \text{if } (m_1, m_2) = (j, l), \\ \underline{a}_{m_1 m_2} & \text{otherwise.} \end{cases}$$

Then  $r_i^+ = 0, r_j^+ = \bar{a}_{jl}$ , so the following holds:

$$\begin{aligned} \underline{a}_{ii}(\underline{a}_{jj} - \bar{a}_{jl}) &= (a_{ii} - r_i^+)(a_{jj} - r_j^+) > \\ > \left( \sum_{m \neq i} (r_i^+ - a_{im}) \right) \left( \sum_{m \neq j} (r_j^+ - a_{jm}) \right) = \\ &= \left( - \sum_{m \neq i} \underline{a}_{im} \right) \left( \sum_{\substack{m \neq j \\ m \neq l}} (\bar{a}_{jl} - \underline{a}_{jm}) \right) \end{aligned}$$

III. Let  $\bar{a}_{ik} \leq 0, \bar{a}_{jl} \leq 0$ . Let us take  $A = \underline{A} \in \mathbf{A}$ . Then  $r_i^+ = 0, r_i^+ = 0$ , so the following holds:

$$\begin{aligned}
 \underline{a}_{ii} \cdot \underline{a}_{jj} &= (a_{ii} - r_i^+) (a_{jj} - r_j^+) > \\
 &> \left( \sum_{m \neq i} (r_i^+ - a_{im}) \right) \left( \sum_{m \neq j} (r_j^+ - a_{jm}) \right) = \\
 &= \left( - \sum_{m \neq i} \underline{a}_{im} \right) \left( - \sum_{m \neq j} \underline{a}_{jm} \right)
 \end{aligned}$$

Thus if  $A$  is an interval doubly B-matrix, then our conditions hold.

**Remark 4** The above mentioned semantic equivalence between Propositions 14 and 15 can be seen from the fact that in the proof of the second one we can use the matrices defined in the first:

In proof of the point I. of the (b) condition: For given  $i, j$ , if we had taken  $i^{A_{\max}}$  for some  $x \neq i, x \neq j$  instead of the matrix that we used, it would have worked even so. (If we restrict our view on the two rows  $i$  and  $j$ , which we are interested in, the two matrices are the same.)

The same reasoning applies for the case that in the point II. of the (b) condition for given  $i, j$  we would have used  $i^{A_{\max}}$ .

And as for the last part, the point III. of the (b) condition, there we are already using one of the matrices from Proposition 14 and that is  $\underline{A}$ .

Ergo it can be seen that the conditions of Proposition 15 are together exactly just the rewritten condition of the real case (from Definition 3) for the matrices from Proposition 14.

Now let us take a closer look on various sufficient conditions for being an interval doubly B-matrix.

First, let us demonstrate, for which matrices are the previous necessary conditions sufficient ones too. Then we shall look at a link between interval B- and doubly B-matrices, which will be analogous to Proposition 8. And after that we will show that for interval Z-matrices, it is quite easy to recognize, whether they are or are not interval doubly B-matrices.

**Proposition 16** *Let  $A \in \mathbb{IR}^{n \times n}$ ,  $n \geq 3$ . If  $A$  fulfills the following condition:*

$$\forall i \in [n] \exists k_i \in [n] \setminus \{i\} \forall j \in [n] \setminus \{i, k_i\} : \bar{a}_{ij} \leq \underline{a}_{ik_i},$$

*then  $A$  is an interval doubly B-matrix if and only if it fulfills the necessary condition stated in Proposition 14.*

**Proof** “ $\Rightarrow$ ” Trivially from Proposition 14.

“ $\Leftarrow$ ”  $\forall i \in [n]$  be  $k_i$  from the assumption. Then  $\forall A \in \mathbf{A} \forall i \in [n]: (r_i^+ > 0 \Rightarrow r_i^+ \leq \bar{a}_{ik_i} \wedge r_i^+ = a_{ik_i})$ .

Let  $A \in \mathbf{A}$  arbitrary.

(a) From assumption it holds that  $\forall i \in [n] : a_{ii} \geq \underline{a}_{ii} > \max\{0, \bar{a}_{ij} | j \neq i\} \geq \max\{0, a_{ij} | j \neq i\}$ .

(b) Let us take arbitrary  $i, j \in [n], j \neq i$ . Let us distinguish the following cases:

(1)  $r_i^+ > 0 \wedge r_j^+ > 0$

Thus from assumption,  $0 < \bar{a}_{ik_i}, 0 < \bar{a}_{jk_j}, r_i^+ = a_{ik_i}, r_j^+ = a_{jk_j}$  and so, because  $i^{A_{\max}}$  for some  $x \neq i, x \neq j$  is a doubly B-matrix from the assumption of this implication, the following applies.

$$\begin{aligned} (a_{ii} - r_i^+) (a_{jj} - r_j^+) &\geq (\underline{a}_{ii} - \bar{a}_{ik_i})(\underline{a}_{jj} - \bar{a}_{jk_j}) > \\ &> \left( \sum_{\substack{m \neq i \\ m \neq k_i}} (\bar{a}_{ik_i} - \underline{a}_{im}) \right) \left( \sum_{\substack{m \neq j \\ m \neq k_j}} (\bar{a}_{jk_j} - \underline{a}_{jm}) \right) \geq \\ &\geq \left( \sum_{\substack{m \neq i \\ m \neq k_i}} (r_i^+ - a_{im}) \right) \left( \sum_{\substack{m \neq j \\ m \neq k_j}} (r_j^+ - a_{jm}) \right) = \\ &= \left( \sum_{m \neq i} (r_i^+ - a_{im}) \right) \left( \sum_{m \neq j} (r_j^+ - a_{jm}) \right) \end{aligned}$$

(2)  $r_i^+ = 0 \wedge r_j^+ > 0$

Thus from assumption,  $\underline{a}_{ik_i} \leq 0, 0 < \bar{a}_{jk_j}, r_j^+ = a_{jk_j}$  and so, because  $i^{A_{\max}}$  is a doubly B-matrix from the assumption of this implication, the following applies.

$$\begin{aligned} (a_{ii} - r_i^+) (a_{jj} - r_j^+) &\geq \underline{a}_{ii} (\underline{a}_{jj} - \bar{a}_{jk_j}) > \\ &> \left( - \sum_{m \neq i} \underline{a}_{im} \right) \left( \sum_{\substack{m \neq j \\ m \neq k_j}} (\bar{a}_{jk_j} - \underline{a}_{jm}) \right) \geq \\ &\geq \left( \sum_{m \neq i} (r_i^+ - a_{im}) \right) \left( \sum_{\substack{m \neq j \\ m \neq k_j}} (r_j^+ - a_{jm}) \right) = \\ &= \left( \sum_{m \neq i} (r_i^+ - a_{im}) \right) \left( \sum_{m \neq j} (r_j^+ - a_{jm}) \right) \end{aligned}$$

(3)  $r_i^+ = 0 \wedge r_j^+ = 0$

Thus from assumption,  $\underline{a}_{ik_i} \leq 0, \underline{a}_{jk_j} \leq 0$  and so, because  $\underline{A}$  is a doubly B-matrix from the assumption of this implication, the following applies.

$$\begin{aligned}
 & (a_{ii} - r_i^+) (a_{jj} - r_j^+) \geq \underline{a}_{ii} \cdot \underline{a}_{jj} > \\
 & > \left( -\sum_{m \neq i} \underline{a}_{im} \right) \left( -\sum_{m \neq j} \underline{a}_{jm} \right) \geq \\
 & \geq \left( \sum_{m \neq i} (r_i^+ - a_{im}) \right) \left( \sum_{m \neq j} (r_j^+ - a_{jm}) \right)
 \end{aligned}$$

Therefore we have shown that in each case the matrix  $A$  is a doubly  $B$ -matrix, thus  $A$  is an interval doubly  $B$ -matrix.

**Proposition 17** *Let  $A \in \mathbb{IR}^{n \times n}$  interval  $Z$ -matrix. It holds that  $A$  is an interval doubly  $B$ -matrix if and only if  $\underline{A}$  is a doubly  $B$ -matrix.*

**Proof** “ $\Rightarrow$ ” Trivially, because  $\underline{A} \in A$ .

“ $\Leftarrow$ ” Let  $A \in A$ . Then

(a) From assumption it holds that  $\forall i \in [n] : a_{ii} \geq \underline{a}_{ii} > \max\{0, \underline{a}_{ij} | j \neq i\} = 0 = \max\{0, \bar{a}_{ij} | j \neq i\} \geq \max\{0, a_{ij} | j \neq i\}$ , because  $\underline{A}$  is a doubly  $B$ -matrix and also a  $Z$ -matrix.

(b) Let  $i, j \in [n], j \neq i$  arbitrary.

$$\begin{aligned}
 & (a_{ii} - r_i^+) (a_{jj} - r_j^+) \geq (\underline{a}_{ii} - 0)(\underline{a}_{jj} - 0) > \\
 & > \left( \sum_{m \neq i} (0 - \underline{a}_{im}) \right) \left( \sum_{m \neq j} (0 - \underline{a}_{jm}) \right) \geq \\
 & \geq \left( \sum_{m \neq i} (r_i^+ - a_{im}) \right) \left( \sum_{m \neq j} (r_i^+ - a_{jm}) \right)
 \end{aligned}$$

The first inequality uses just the fact that  $r_i^+ = 0, r_j^+ = 0$  and that  $\forall i, j \in [n] : a_{ij} \geq \underline{a}_{ij}$ , the second one holds, because  $\underline{A}$  is a doubly  $B$ -matrix and because it is a  $Z$ -matrix too. And the last inequality is again completely trivial using the same facts, as the first one.

**Proposition 18** *Let  $A \in \mathbb{IR}^{n \times n}, \forall i \in [n] : k_i \in \operatorname{argmax}\{\bar{a}_{ij} | j \neq i\}$  and  $\forall i \in [n] : k'_i \in \operatorname{argmax}\{\underline{a}_{ij} | j \neq i\}$ . Let us define  $\tilde{A} \in \mathbb{IR}^{n \times n}$  as follows:*

$$\tilde{A} = (\tilde{a}_{m_1 m_2}); \quad \tilde{a}_{m_1 m_2} = \begin{cases} \bar{a}_{m_1 k_{m_1}} & \text{if } m_2 = k_{m_1}, \\ \underline{a}_{m_1 m_1} & \text{if } m_2 = m_1, \\ \min\{\underline{a}_{m_1 m_2}, \underline{a}_{m_1 k_{m_1}}\} & \text{otherwise.} \end{cases}$$

*If  $\forall i \in [n] : \underline{a}_{ik'_i} \geq 0$  and  $\tilde{A}$  is a doubly  $B$ -matrix, then  $A$  is an interval doubly  $B$ -matrix.*

**Proof** Let  $A \in \mathbf{A}$ ,  $i, j \in [n]$ ,  $j \neq i$  arbitrary. Then  $\bar{a}_{ik_i} \geq 0$ , because  $\bar{a}_{ik_i} \geq \underline{a}_{ik'_i} \geq 0$  from the assumption and the definition of  $k_i$  (analogously for  $j$ ). And so the *a*) condition of the Definition 3 is satisfied trivially ( $\underline{a}_{ii} > \max\{0, \bar{a}_{ik_i}\} = \bar{a}_{ik_i}$ ) and as for the *b*) condition, let  $\forall i \in [n] : l_i \in \operatorname{argmax}\{a_{im} | m \neq i\}$ , then, because  $\max\{a_{im} | m \neq i\} \geq \underline{a}_{ik'_i} \geq 0$ , it holds that  $a_{il_i} = r_i^+$ . Hence:

$$\begin{aligned} (a_{ii} - r_i^+) (a_{jj} - r_j^+) &\geq (\underline{a}_{ii} - \bar{a}_{ik_i})(\underline{a}_{jj} - \bar{a}_{jk_j}) = \\ &= (\tilde{a}_{ii} - \tilde{r}_i^+) (\tilde{a}_{jj} - \tilde{r}_j^+) > \left( \sum_{m \neq i} (\tilde{r}_i^+ - \tilde{a}_{im}) \right) \left( \sum_{m \neq j} (\tilde{r}_j^+ - \tilde{a}_{jm}) \right) = \\ &= \left( \sum_{\substack{m \neq i \\ m \neq k_i}} (\bar{a}_{ik_i} - \min\{\underline{a}_{im}, \underline{a}_{ik_i}\}) \right) \left( \sum_{\substack{m \neq j \\ m \neq k_j}} (\bar{a}_{jk_j} - \min\{\underline{a}_{jm}, \underline{a}_{jk_j}\}) \right) \geq \\ &\geq \left( \sum_{\substack{m \neq i \\ m \neq l_i}} (a_{il_i} - a_{im}) \right) \left( \sum_{\substack{m \neq j \\ m \neq l_j}} (a_{jl_j} - a_{jm}) \right) = \\ &= \left( \sum_{m \neq i} (r_i^+ - a_{im}) \right) \left( \sum_{m \neq j} (r_j^+ - a_{jm}) \right) \end{aligned}$$

(Where  $\tilde{r}_i^+$  is  $r_i^+$  due to the matrix  $\tilde{A}$ .)

The third inequality holds, because of the assumption that the  $\tilde{A}$  is a doubly B-matrix. The fifth inequality is quite trivial too, it relies only on the following two facts:  $\bar{a}_{ik_i} \geq a_{il_i} \geq 0$  and for  $m = l_i : a_{im} = \min\{\underline{a}_{im}, \underline{a}_{ik_i}\} \leq \underline{a}_{ik_i} \leq a_{ik_i}$ . And the last equality holds, because  $\forall i : a_{il_i} = r_i^+$ .

Now, what we could be interested in, is for what matrices is the sufficient condition from Proposition 18 characterization as well. So let us take a look at that.

**Proposition 19** Let  $A \in \mathbb{IR}^{n \times n}$  such that  $\forall i \in [n] \exists k_i \in [n] \setminus \{i\} : \bar{a}_{ik_i} = \max\{\bar{a}_{ij} | j \neq i\} \wedge \underline{a}_{ik_i} = \max\{0, \underline{a}_{ij} | j \neq i\}$ . Then the sufficient condition stated in Proposition 18 is a characterization for  $A$ .

**Proof** “ $\tilde{A}$  is a doubly B-matrix, then  $A$  is an interval doubly B-matrix”: It follows from Proposition 18.

“ $A$  is an interval doubly B-matrix, then  $\tilde{A}$  is a doubly B-matrix”: From construction of  $\tilde{A}$  and from assumptions of this proposition it follows that  $\tilde{A} \in \mathbf{A}$ .

Next we will show that if the lower and upper bound matrices of an interval matrix are circulant, then some nice properties applies.

**Proposition 20** Let  $A \in \mathbb{IR}^{n \times n}$  such that  $\underline{A}$  and  $\bar{A}$  are circulant. Then the following are equivalent:

- (1)  $A$  is an interval doubly B-matrix
- (2)  $A$  is an interval B-matrix
- (3) It holds that

$$(a) \underline{a}_{11} > - \sum_{j \neq 1} \underline{a}_{1j}$$

$$(b) \forall k \in [n] \setminus \{1\} : \underline{a}_{11} - \bar{a}_{1k} > \sum_{\substack{j \neq 1 \\ j \neq k}} (\bar{a}_{1k} - \underline{a}_{1j})$$

**Proof** “(1)  $\Rightarrow$  (2)”  $A$  is a doubly B-matrix, thus it satisfies the characterization given in Theorem 2. Let  $i \in [n]$  and  $k \neq i$  arbitrary. Se let us choose

$$j = \begin{cases} i + 1 & \text{if } i < n, \\ 1 & \text{if } i = n. \end{cases} \text{ and } l = \begin{cases} k + 1 & \text{if } k < n, \\ 1 & \text{if } k = n. \end{cases}$$

(Then  $\underline{a}_{ii} = \underline{a}_{jj}$  and  $\bar{a}_{ik} = \bar{a}_{jl}$ , because both  $\underline{A}$  and  $\bar{A}$  are circulant.) Hence, because  $A$  is an interval doubly B-matrix:

$$\begin{aligned} & \underline{a}_{ii} \cdot \underline{a}_{jj} > \left( \max \left\{ 0, - \sum_{m \neq i} \underline{a}_{im} \right\} \right) \left( \max \left\{ 0, - \sum_{m \neq j} \underline{a}_{jm} \right\} \right) \Leftrightarrow \\ & \Leftrightarrow \underline{a}_{ii}^2 > \left( \max \left\{ 0, - \sum_{m \neq i} \underline{a}_{im} \right\} \right)^2 \Rightarrow \\ & \Rightarrow \underline{a}_{ii} = |\underline{a}_{ii}| > \left| \max \left\{ 0, - \sum_{m \neq i} \underline{a}_{im} \right\} \right| \geq - \sum_{m \neq i} \underline{a}_{im} \Rightarrow \\ & \Rightarrow \sum_{m=1}^n \underline{a}_{im} > 0 \end{aligned}$$

Therefore the (a) condition of Corollary 5 is satisfied. Let us take a look at the second one, the (b) condition:

$$\begin{aligned} & (\underline{a}_{ii} - \bar{a}_{ik})(\underline{a}_{jj} - \bar{a}_{jl}) > \\ & > \left( \max \left\{ 0, \sum_{\substack{m \neq i \\ m \neq k}} (\bar{a}_{ik} - \underline{a}_{im}) \right\} \right) \left( \max \left\{ 0, \sum_{\substack{m \neq j \\ m \neq l}} (\bar{a}_{jl} - \underline{a}_{jm}) \right\} \right) \Leftrightarrow \end{aligned}$$



$$\begin{aligned} \Leftrightarrow (\underline{a}_{ii} - \bar{a}_{ik})^2 &> \left( \max \left\{ 0, \sum_{\substack{m \neq i \\ m \neq k}} (\bar{a}_{ik} - \underline{a}_{im}) \right\} \right)^2 \Rightarrow \\ \Rightarrow \underline{a}_{ii} - \bar{a}_{ik} = |\underline{a}_{ii} - \bar{a}_{ik}| &> \left| \max \left\{ 0, \sum_{\substack{m \neq i \\ m \neq k}} (\bar{a}_{ik} - \underline{a}_{im}) \right\} \right| = \\ &= \max \left\{ 0, \sum_{\substack{m \neq i \\ m \neq k}} (\bar{a}_{ik} - \underline{a}_{im}) \right\} \geq \sum_{\substack{m \neq i \\ m \neq k}} (\bar{a}_{ik} - \underline{a}_{im}) \Rightarrow \\ \Rightarrow \underline{a}_{ii} - \bar{a}_{ik} &> \sum_{\substack{m \neq i \\ m \neq k}} (\bar{a}_{ik} - \underline{a}_{im}) \end{aligned}$$

Hence  $A$  fulfills the characterization of an interval B-matrix as stated in Corollary 5, thus it is an interval B-matrix.

“(2)  $\Rightarrow$  (3)”  $A$  is an interval B-matrix, so it satisfies the characterization given in Corollary 5. Hence for  $i = 1$  it follows from condition (a) of the corollary that our condition a) holds, and the same goes for the (b) conditions.

“(3)  $\Rightarrow$  (2)” From our (a) condition we know that row sum of the first row of matrix  $A$  is positive. And because  $A$  is circulant, all the row sums of this matrix are the same, thus positive.  $\Rightarrow \forall i \in [n] : \sum_{j=1}^n a_{ij} > 0$

And because both  $A$  and  $\bar{A}$  are circulant and from our condition (b), we get

$$\forall i \in [n] \forall k \neq 1 \exists k_i \neq i :$$

$$\underline{a}_{11} - \bar{a}_{1k} = \underline{a}_{ii} - \bar{a}_{ik_i} \wedge \sum_{\substack{j \neq 1 \\ j \neq k}} (\bar{a}_{1k} - \underline{a}_{1j}) = \sum_{\substack{j \neq i \\ j \neq k_i}} (\bar{a}_{ik_i} - \underline{a}_{ij}).$$

$$\Rightarrow \forall i \in [n] \forall k \neq i : \underline{a}_{ii} - \bar{a}_{ik} > \sum_{\substack{j \neq i \\ j \neq k}} (\bar{a}_{ik} - \underline{a}_{ij})$$

Therefore  $A$  is an interval B-matrix, because it fulfills the conditions of characterization shown in Corollary 5.

“(2)  $\Rightarrow$  (1)” Trivial (see Proposition 12).

## References

1. Cottle, R.W., Pang, J.S., Stone, R.E.: The Linear Complementarity Problem. SIAM, Philadelphia, PA, revised ed. of the 1992 original edn. (2009)
2. Coxson, G.E.: The P-matrix problem is co-NP-complete. Math. Program. **64**(1), 173–178 (1994)

3. Garloff, J., Adm, M., Titi, J.: A survey of classes of matrices possessing the interval property and related properties. *Reliab. Comput.* **22**, 1–10 (2016)
4. Hladík, M.: On relation between P-matrices and regularity of interval matrices. In: Bebiano, N. (ed.) *Applied and Computational Matrix Analysis*, Springer Proceedings in Mathematics & Statistics, vol. 192, pp. 27–35. Springer (2017)
5. Hladík, M.: An overview of polynomially computable characteristics of special interval matrices. In: Kosheleva, O., et al. (eds.) *Beyond Traditional Probabilistic Data Processing Techniques: Interval, Fuzzy etc. Methods and Their Applications*, Studies in Computational Intelligence, vol. 835, pp. 295–310. Springer, Cham (2020)
6. Horáček, J., Hladík, M., Černý, M.: Interval linear algebra and computational complexity. In: Bebiano, N. (ed.) *Applied and Computational Matrix Analysis*, Springer Proceedings in Mathematics & Statistics, vol. 192, pp. 37–66. Springer (2017)
7. Kreinovich, V., Lakeyev, A., Rohn, J., Kahl, P.: *Computational Complexity and Feasibility of Data Processing and Interval Computations*. Kluwer, Dordrecht (1998)
8. Lorenc, M.: Special classes of P-matrices in the interval setting. Bachelor's thesis, Department of Applied Mathematics, Charles University (2021)
9. Peña, J.M.: A class of P-matrices with applications to the localization of the eigenvalues of a real matrix. *SIAM J. Matrix Anal. Appl.* **22**(4), 1027–1037 (2001)
10. Peña, J.M.: On an alternative to Gerschgorin circles and ovals of Cassini. *Numer. Math.* **95**(2), 337–345 (2003)
11. Rohn, J.: On Rump's characterization of P-matrices. *Optim. Lett.* **6**(5), 1017–1020 (2012)

# Commonsense “And”-Operations



Javier Tellez, Wenbo Xie, and Vladik Kreinovich

**Abstract** In many practical situations, we need to estimate our degree of belief in a statement  $A \& B$  when the only thing we know are the degrees of belief  $a$  and  $b$  in combined statements  $A$  and  $B$ . An algorithm for this estimation is known as an “and”-operation, or, for historical reasons, a t-norm. Usually, “and”-operations are selected in such a way that if one of the statements  $A$  or  $B$  is false, our degree of belief in  $A \& B$  is 0. However, in practice, this is sometimes not the case: for example, an ideal faculty candidate must satisfy many properties—be a great teacher, *and* be a wonderful researcher, *and* be a great mentor, etc.—but if one of these requirements is not satisfied, this candidate may still be hired. In this paper, we show how to describe the corresponding commonsense “and”-operations.

## 1 Why “and”-Operations

In many practical applications, a certain effect appears if several conditions  $C_1, C_2, \dots$  are satisfied. For each of these conditions  $C_i$ , we can elicit, from the experts, the degree  $d_i \in [0, 1]$  to which this condition is satisfied.

However, there are many possible conditions. It is not possible to extract, from the experts, a degree to which each possible “and”-combination  $C_1 \& C_2 \& \dots$  is satisfied. Thus, we need to be able:

- given degrees of confidence  $a$  and  $b$  in statements  $A$  and  $B$ ,
- to estimate the degree to which the “and”-combination  $A \& B$  is satisfied.

---

J. Tellez · W. Xie · V. Kreinovich (✉)

Department of Computer Science, University of Texas at El Paso El Paso, Texas 79968, USA  
e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

J. Tellez  
e-mail: [jdtellez@miners.utep.edu](mailto:jdtellez@miners.utep.edu)

W. Xie  
e-mail: [wxie@miners.utep.edu](mailto:wxie@miners.utep.edu)

This estimate is usually denoted by  $f_{\&}(a, b)$ . The algorithm for computing this estimate is known as an “and”-operation or, for historical reason, a *t-norm*.

## 2 How Usual “and”-Operations are Obtained

In some situations, about each of the combined statements, we are absolutely certain either that this statement is true, or that this statement is false. Then, the “and”-operation should return the true value of the corresponding “and”-statement. So we should have  $f_{\&}(0, 0) = f_{\&}(0, 1) = f_{\&}(1, 0) = 0$  and  $f_{\&}(1, 1) = 1$ .

We want to extend these values to all possible combinations of  $a \in [0, 1]$  and  $b \in [0, 1]$ . A reasonable idea is to use linear interpolation over each variable (see, e.g., [3]), i.e., to assume that:

- for every  $a$ , the mapping  $b \mapsto f_{\&}(a, b)$  is linear, and
- for every  $b$ , the mapping  $a \mapsto f_{\&}(a, b)$  is linear.

As a result, we conclude that the desired function is bilinear, i.e., that it has the form

$$f_{\&}(a, b) = c_0 + c_a \cdot a + c_b \cdot b + c_{ab} \cdot a \cdot b$$

for some coefficients  $c_i$ .

Taking into account the above conditions for  $a, b \in \{0, 1\}$ , we conclude that  $f_{\&}(a, b) = a \cdot b$ . This is indeed one of the most frequently used “and”-operations; [1, 2, 4–7].

Similarly, linear interpolation enables us to similarly determine that an appropriate “or”-operation (historically also known as *t-conorm*) has the form

$$f_{\vee}(a, b) = a + b - a \cdot b.$$

## 3 Need to go Beyond the Usual “and”-Operations

In some cases, when we say “and”, we mean exactly the logical “and”: all conditions must be absolutely satisfied.

However, in many practical problems, “and” is “softer” than that. For example, if you ask a person who is planning to buy a house what house he/she wants, the person will say:

- not too far away
- *and* spacey
- *and* not very expensive
- *and* reasonably well thermo-isolated
- *and* in a nice neighborhood, etc.

However, this “and” does not mean literal “and”. If this person finds a house that satisfied most of these conditions, he/she will gladly buy it.

How can we describe such commonsense “and”-operations?

## 4 Our Solution

In this paper, we consider the case when we only have two conditions. For a commonsense “and”-operation  $F_{\&}(a, b)$ , it is reasonable to still have  $F_{\&}(0, 0) = 0$  and  $F_{\&}(1, 1) = 1$ . However:

- if only one of the conditions  $A$  and  $B$  is satisfied,
- then the statement  $A \& B$  should also be to some extent true.

In other words, we should have  $F_{\&}(0, 1) = F_{\&}(1, 0) = \alpha$  for some small  $\alpha > 0$ .

In this case, we get  $F_{\&}(a, b) = \alpha \cdot (a + b) + (1 - 2\alpha) \cdot a \cdot b$ . Equivalently,

$$F_{\&}(a, b) = (1 - \alpha) \cdot a \cdot b + \alpha \cdot (a + b - a \cdot b) = (1 - \alpha) \cdot f_{\&}(a, b) + \alpha \cdot f_{\vee}(a, b).$$

In other words, this operation is a convex combination of the usual “and”- and “or”-operations.

## 5 Discussion

The usual “and”-operation is associative. Thus, we can define  $f_{\&}(a, b, c)$  as, e.g.,  $f_{\&}(a, f_{\&}(b, c))$  or as  $f_{\&}(f_{\&}(a, b), c)$ —and the result will not change.

In contrast, the commonsense “and”-operation is not associative. With the commonsense “and”-operation, we will have two different results.

So, e.g., for three inputs, we get a more general formula

$$F_{\&}(a, b, c) = \alpha \cdot (a + b + c) + \beta \cdot (a \cdot b + b \cdot c + a \cdot c) + (1 - 3\alpha - 3\beta) \cdot a \cdot b \cdot c.$$

**Acknowledgements** This work was supported in part by the National Science Foundation grants: • 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and

- HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported:

- by the AT&T Fellowship in Information Technology, and
  - by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.
- The authors are thankful to all the participants of the 26th Annual UTEP/NMSU Workshop on Mathematics, Computer Science, and Computational Science (El Paso, Texas, November 5, 2021) for valuable discussions.

## References

1. Belohlavek, R., Dauben, J.W., Klir, G.J.: *Fuzzy Logic and Mathematics: A Historical Perspective*. Oxford University Press, New York (2017)
2. Klir, G., Yuan, B.: *Fuzzy Sets and Fuzzy Logic*. Prentice Hall, Upper Saddle River, New Jersey (1995)
3. Kreinovich, V., Quijas, J., Gallardo, E., C. De Sa Lopes, O. Kosheleva, and S. Shahbazova: Simple linear interpolation explains all usual choices in fuzzy techniques: membership functions, t-norms, t-conorms, and defuzzification. In: *Proceedings of the Annual Conference of the North American Fuzzy Information Processing Society NAFIPS'2015 and 5th World Conference on Soft Computing*, Redmond, Washington, August 17–19 (2015)
4. Mendel, J.M.: *Uncertain Rule-Based Fuzzy Systems: Introduction and New Directions*. Springer, Cham, Switzerland (2017)
5. Nguyen, H.T., Walker, C.L., Walker, E.A.: *A First Course in Fuzzy Logic*. Chapman and Hall/CRC, Boca Raton, Florida (2019)
6. Novák, V., Perfilieva, I., Močkoř, J.: *Mathematical Principles of Fuzzy Logic*. Kluwer, Boston, Dordrecht (1999)
7. Zadeh, L.A.: Fuzzy sets. *Information and Control* **8**, 338–353 (1965)

# Extension to Multidimensional Problems of a Fuzzy-Based Explainable and Noise-Resilient Algorithm



Javier Viaña, Stephan Ralescu, Kelly Cohen, Anca Ralescu,  
and Vladik Kreinovich

**Abstract** While Deep Neural Networks (DNNs) have shown incredible performance in a variety of data, they are *brittle* and *opaque*: easily fooled by the presence of noise, and difficult to understand the underlying reasoning for their predictions or choices. This focus on accuracy at the expense of interpretability and robustness caused little concern since, until recently, DNNs were employed primarily for scientific and limited commercial work. An increasing, widespread use of artificial intelligence and growing emphasis on user data protections, however, motivates the need for robust solutions with explainable methods and results. In this work, we extend a novel fuzzy based algorithm for regression to multidimensional problems. Previous research demonstrated that this approach outperforms neural network benchmarks while using only 5% of the number of the parameters.

**Keywords** Multidimensional problem · Explainable fuzzy AI · Noise-resilience · Regression

---

J. Viaña · S. Ralescu · K. Cohen · A. Ralescu  
University of Cincinnati, Cincinnati, OH 45219, USA  
e-mail: [vianajr@mail.uc.edu](mailto:vianajr@mail.uc.edu)

S. Ralescu  
e-mail: [ralescs@mail.uc.edu](mailto:ralescs@mail.uc.edu)

K. Cohen  
e-mail: [cohenky@ucmail.uc.edu](mailto:cohenky@ucmail.uc.edu)

A. Ralescu  
e-mail: [ralescal@ucmail.uc.edu](mailto:ralescal@ucmail.uc.edu)

V. Kreinovich (✉)  
University of Texas, El Paso, TX 79968, USA  
e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

# 1 Introduction

## 1.1 A Subsection Sample

A key factor in the success of deep neural networks ability to treat a variety of data is the depth of the model and the addition of multiple hidden layers. While DNNs demonstrate high accuracy in several tasks, their excellence relies on difficult-to-understand abstract representations in the hidden layers that obscure their decision-making process. Furthermore, despite the fact that DNNs can discover patterns in the features of the data, minor changes in the inputs, such as noise imperceptible to human senses, can cause the DNNs to misclassify an object or make a false prediction [1]. The fragile black-box nature of these networks complicates their application to fields such as medicine, autonomous cars, national security, or any field where safety and accountability must be guaranteed. As the use of artificial intelligence has become more widespread with an increasing emphasis on data privacy and protection, the need for interpretable solutions with explainable methods and results has emerged as an essential problem.

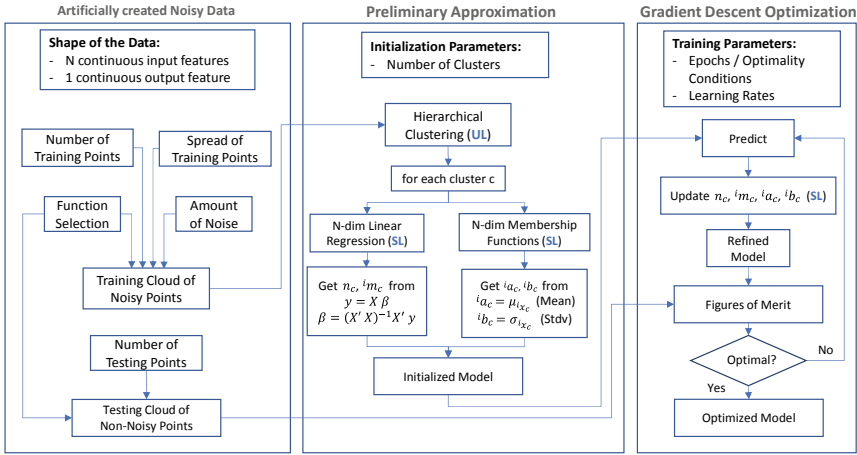
In [2], a novel fuzzy based regression algorithm was introduced for one-dimensional input problems. In the benchmark of the authors, their method proved greater noise-resilience and explainability than the neural networks considered for the same task. Given the success of such prior work, this paper is an extension of [2] to multidimensional problems. The architecture of the algorithm remains the same as the one described in [2], but the learning rules are slightly adapted to be used in an N-dimensional scenario.

Section 1.2 is a description of the algorithm's phases and the corresponding mathematical formulation. Section 2 contains the empirical evaluation obtained for a 2-input 1-output problem, proving the applicability of the method. Section 3 offers a discussion of the algorithm's properties, and finally in Sect. 4, the authors cover their conclusions and ideas for future work.

## 1.2 The Algorithm

As opposed to Deep Learning architectures, this algorithm uses a single layer of predictors (each specialized in a specific region of the input space). The Takagi–Sugeno–Kang method is used to merge the outcomes of all the individual predictors, similar to how an ensemble system works (where each expert is trained in a particular subspace of the data). Nevertheless, the main difference with the latter is the collective training and the ability to share information among the systems. Furthermore, each predictor has no more than  $3N + 1$  parameters ( $N$  being the number of input dimensions), whereas the most popular ensemble systems often use an entire neural network for each expert. The fact that the algorithm has a single layer, together with the reduced number of parameters needed for the prediction, makes it very





**Fig. 1** Block diagram of the proposed algorithm for multidimensional function approximation given a noisy cloud of training datapoints and non-noisy testing data. UL and SL stand for Unsupervised and Supervised Learning, respectively

explainable and easy to visualize. The block diagram of the algorithm is shown in Fig. 1.

The first step is the application of a Hierarchical Clustering algorithm (agglomerative, MAX-linkage/complete-linkage) to divide the joint input–output space in clusters of data. These clusters are not necessarily isolated groups of datapoints, but rather points that were close enough to be modeled in conjunction. Each cluster  $c$  (bottom right index in the formulation) is approximated with an N-dimensional hyperplane ( $r_c$ ). The N slopes and the intercept of the hyperplane are identified with  $m$  and  $n$ , respectively.

$x^q$  and  $i x^q$  represent the input vector of the  $q$ th datapoint and the  $i$ th feature or entry of this vector. In other words, the upper left index refers to the dimension and upper right index determines the instance of the data.

$$r_c(x^q) = n_c + \sum_{i=1}^N m_c^i x^q \tag{1}$$

Additionally, the clusters can be seen as fuzzy, where some datapoints belong to the cluster with a higher degree of membership. To model the membership function ( $\mu$ ) of a given fuzzy set ( $c$ ), an N-dimensional Cauchy distribution's [3] density function is chosen, (2). The  $i a_c$  parameters represent the location of the function's center, which is initialized with the mean of the cluster in each dimension. Similarly, the initialized  $i b_c$  parameters match the standard deviation.

$$\mu_c(\mathbf{x}^q) = \left[ 1 + \sum_{i=1}^N \left( \frac{i x^q - i a_c}{i b_c} \right)^2 \right]^{-1} \quad (2)$$

Finally, the prediction is obtained using a Takagi–Sugeno–Kang approach. This is calculated as the weighted mean considering the information of all the membership functions and hyperplanes, as shown in Eq. (3).

$$\hat{y}^q = \left[ \sum_{c=1}^C \mu_c(\mathbf{x}^q) r_c(\mathbf{x}^q) \right] \left[ \sum_{c=1}^C \mu_c(\mathbf{x}^q) \right]^{-1} \quad (3)$$

For the training of the parameters, Gradient Descent (GD) learning is used. This is possible given the analytical definition of the algorithm, which in many other popular fuzzy-based predictors is not necessarily always true [4]. Nevertheless, there is a plethora of gradient-free learning algorithms for that type of systems that might rely on the experimental inference, [5–7]. In the present case, the GD was carried out minimizing the Mean Squared Total Loss  $J$  shown in Eq. (4).

$$J = \sum_{q=1}^Q J^q = \frac{1}{2} \sum_{q=1}^Q (y^q - \hat{y}^q)^2 \quad (4)$$

The resulting learning rules for the parameters are

$$\Delta n_c = \eta_n \sum_{q=1}^Q u_c^q \quad (5)$$

$${}^i \Delta m_c = \eta_m \sum_{q=1}^Q u_c^q x^q \quad (6)$$

$${}^i \Delta a_c = -\eta_a \sum_{q=1}^Q v_c^q (i x^q - i a_c) \quad (7)$$

$${}^i \Delta b_c = -\eta_b \sum_{q=1}^Q v_c^q (i x^q - i a_c)^2, \quad (8)$$

where  $\eta$  represents the learning rate for each parameter and the auxiliary variables  $u_c^q$  and  $v_c^q$  are

$$u_c^q = (y^q - \hat{y}^q) \left[ \sum_{k=1}^C \mu_k(\mathbf{x}^q) \right]^{-1} \mu_c(\mathbf{x}^q) \quad (9)$$

$$v_c^q = u_c^q [\hat{y}^q - r_c(\mathbf{x}^q)] \mu_c(\mathbf{x}^q) \tag{10}$$

## 2 Results

In order to prove that the prior formulation is correct and applicable to the multi-dimensional case, a set of 2-input 1-output functions is selected to visually assess the performance of the algorithm. These functions,  $F1$ ,  $F2$  and  $F3$  are defined in Eqs. (11–13).

$$F1(^1x, ^2x) = 3(^1x)^2 + 5^1x^2x(1 + 4\sin(^1x)) + 4, ^1x \in [4, 14], ^2x \in [1, 6] \tag{11}$$

$$F2(^1x, ^2x) = \sin(^1x)\cos(^2x), ^1x \in [-4, 4], ^2x \in [-4, 4] \tag{12}$$

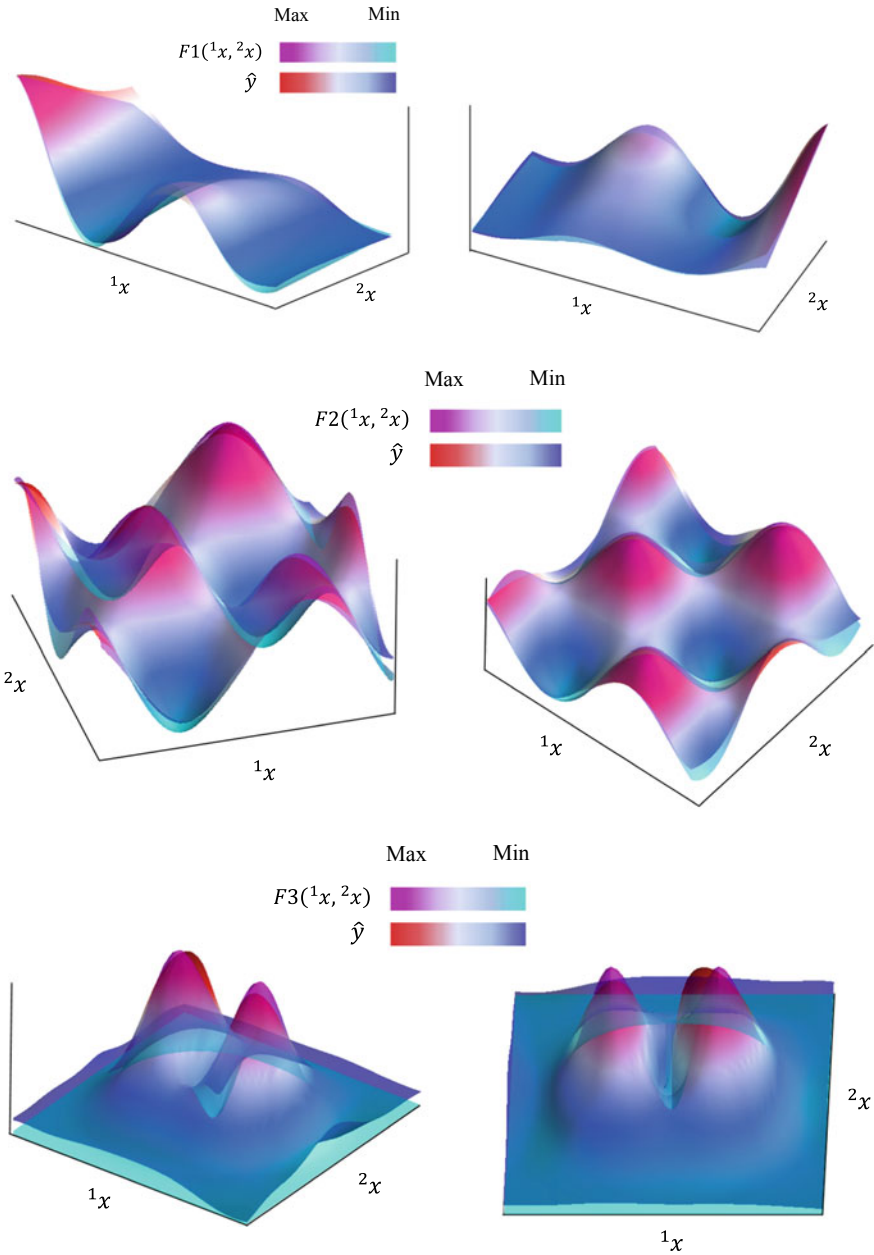
$$F3(^1x, ^2x) = \left[ (^1x)^2 + 3(^2x)^2 \right] e^{-(^1x)^2 - (^2x)^2}, ^1x \in [-3, 3], ^2x \in [-3, 3] \tag{13}$$

For training, a total of 60, 150 and 150 datapoints were used for each function, respectively. These instances were randomly generated within the input spaces. In previous research [2], this algorithm exhibited better performance than the neural networks when the training data was corrupted with artificial noise. In order to be coherent, in this bi-dimensional input space problem, a random amount of noise was injected in the training instances as well (the amount of noise was obtained using uniform distributions with boundaries  $\pm 100$ ,  $\pm 0.025$ , and  $\pm 0.025$  respectively). For all of the three functions, the learning rate was 0.001. The training stopped when the evolution of the mean squared error (with respect to the corrupted training data) converged (which translates into 2000, 800 and 1600 epochs for each surface). The number of clusters considered were 5, 12 and 8. Figure 2 illustrates the approximations obtained against the non-noisy ground truth across the entire domain of the input space.

## 3 Discussion

As it can be seen in Fig. 2, the solution provided by the algorithm matches the real non-noisy surface despite using the corrupted training data. This not only proves the applicability of the formulation to multidimensional problems but it also emphasizes the noise-resilience of the proposed algorithm.

The choice of the 2-dimensional functions was made so that the results could be displayed in the figures included in the paper, which enables a visual assessment of the algorithm’s applicability to the multidimensional problems.



**Fig. 2** Approximations vs non-noisy ground truth plots with standardized variables. Top, middle, and bottom rows refer to functions  $F1$ ,  $F2$  and  $F3$  respectively. Both columns show the same information but with different perspectives

One of the advantages of this algorithm is the possibility of direct integration of expert knowledge in the system as a new cluster (a theoretical or experimental formulation of the problem studied, or even an approximation of the it). This is not an option in a conventional neural network, where the weights do not have a physical or mathematical meaning as clear as the slope of a linear relationship. Also, adding a new neuron in a network would unbalance the weights, destroying the predictive power of the system, and ultimately requiring retraining.

Furthermore, the weights of a trained neural network depend on the initialized weights, which often vary in every execution. Thus, the parameters of the model are different after every training process. On the contrary, the method covered in this research converges always to the same values (since the hierarchical clustering is entirely deterministic).

## 4 Conclusions

We have introduced an extension of [2] to multidimensional prediction tasks that is resilient to noise and explainable. Our algorithm in [2] is simple and uses an order of magnitude fewer parameters than the benchmark neural network. We achieve explainability by leveraging the interpretable nature of the fuzzy inferencing system, without resorting to the use of hidden layers. Each fuzzy set represents a linear approximation, which could refer to a theoretical or physical formulation of the problem. Since our algorithm generates a weighted prediction, where all regressions have an influence in the outcome, it does not rely on a single predictor, resulting in resilience to noise.

In future work, we plan to apply our algorithm to real-world high-dimensional data sets covering a variety of applications characterized by a need for resilient and explainable results. These include the evaluation of our algorithm against the state-of-the-art for robust and interpretable DNNs, particularly when ground truth functions are unavailable, as well as the integration of the proposed approach into other popular methodologies such as image processing and feature recognition problems in conjunction with Convolutional Neural Networks.

## References

1. Heaven, D.: Why deep-learning AIs are so easy to fool. *Nature* **574**, 163–166 (2019)
2. Viaña, J., Cohen, K.: Fuzzy-based, noise-resilient, explainable algorithm for regression. In: *Proceedings of the Annual Conference of the North American Fuzzy Information Processing Society 2021, NAFIPS*. Springer, Purdue University, IN (2021)
3. Carrillo, R.E., Aysal, T.C., Barner, K.E.: A generalized cauchy distribution framework for problems requiring robust behavior. *EURASIP J. Adv. Signal. Process.* 1–19 (2010)
4. Viaña, J., Cohen, K.: ExTree—Explainable genetic feature coupling tree using fuzzy mapping for dimensionality reduction with application to NACA 0012 airfoils self-noise data set. In:

- Proceedings of the Annual Conference of the North American Fuzzy Information Processing Society 2020, NAFIPS. Springer, Redmond, WA (2020)
5. Pickering, L., Cohen, K.: Genetic fuzzy based tetris player. In: Proceedings of the Annual Conference of the North American Fuzzy Information Processing Society 2020, NAFIPS. Springer, Redmond, WA (2020)
  6. Viaña, J., Cohen, K.: Fast training algorithm for genetic fuzzy controllers and application to an inverted pendulum with free cart. In: Proceedings of the Annual Conference of the North American Fuzzy Information Processing Society 2020, NAFIPS. Springer, Redmond, WA (2020)
  7. Tsai, S., Chen, Y.: A novel fuzzy identification method based on ant colony optimization algorithm. *IEEE Access* **4**, 3747–3756 (2016)

# Additional Spatial Dimensions Can Help Speed Up Computations



Luc Longpré, Olga Kosheleva, and Vladik Kreinovich

**Abstract** While we currently only observe 3 spatial dimensions, according to modern physics, our space is actually at least 10-dimensional. In this paper, on different versions of the multi-D spatial models, we analyze how the existence of the additional spatial dimensions can help computations. It turns out that in all the versions, there is some speed up—moderate when the extra dimensions are actually compactified, and drastic if extra dimensions are separated by a potential barrier.

## 1 Computations and Space Dimensions: How They Are Related and What Are the Remaining Open Problems

**Many computational problems require too much computation time.** It is known that many practical computational problems are NP-hard; see, e.g., [4, 6]. This means, crudely speaking, that unless it turns out that  $P = NP$  (which most computer scientists do not believe to be possible), any algorithm that always solves the corresponding problem will require, at least for some inputs of reasonably large size, an unrealistically long time to solve—e.g., time larger than the lifetime of the Universe.

**Parallelization can help—at least to some extent.** If for a person, some task takes too much time, this person can (and does) ask for help. When two or more people work on some task, they can perform it faster. Similarly, when a certain computational

---

L. Longpré · O. Kosheleva · V. Kreinovich (✉)  
Department of Computer Science, University of Texas at El Paso, 500 W. University El Paso,  
Austin, TX 79968, USA  
e-mail: [vladik@utep.edu](mailto:vladik@utep.edu)

L. Longpré  
e-mail: [longpre@utep.edu](mailto:longpre@utep.edu)

O. Kosheleva  
e-mail: [olgak@utep.edu](mailto:olgak@utep.edu)

task requires too much time on a single computer, a natural way to speed up computations is to divide the original task between several computers—i.e., to parallelize computations. Many modern high-performance computers consist of thousands of processors working in parallel on the same task, and for many computational tasks, this indeed leads to a drastic speed-up.

**Fundamental limitations of parallelization speed-up.** In spite of the numerous successes of parallel computations, in general, parallelization is not a panacea: this idea has limitations. Some of these limitations are technical. These limitations will hopefully be overcome in the future. However, as we will show, there are also fundamental limitations on how much speed-up can be achieved by parallelization; see, e.g., [5].

Indeed, let us assume that we have a parallel computer that finishes its computations in time  $T_{\text{par}}$ . Let us show how we can simulate its computations sequentially. According to modern physics (see, e.g., [1, 8]), the speed of all processes is bounded by the speed of light  $c$ . During the time  $T_{\text{par}}$ , the information from the processors must reach the user. This means that the processors that participate in this computation must be located within the distance  $R \stackrel{\text{def}}{=} c \cdot T_{\text{par}}$ , i.e., in geometric terms, inside the sphere of radius  $R$  centered at the user location.

The overall volume of this area is equal to

$$V = \frac{4}{3} \cdot \pi \cdot R^3 = \frac{4}{3} \cdot \pi \cdot c^3 \cdot T_{\text{par}}^3.$$

Thus, if we denote by  $\Delta V$  the smallest possible volume of a single processor, then the number of processor  $N_{\text{proc}}$  that can fit inside this sphere cannot exceed the value

$$N_{\text{proc}} \leq N_{\text{max}} \stackrel{\text{def}}{=} \frac{V}{\Delta V} = \frac{4}{3 \cdot \Delta V} \cdot \pi \cdot c^3 \cdot T_{\text{par}}^3. \quad (1)$$

Whatever we can compute in parallel on  $N_{\text{proc}}$  processors, we can also compute sequentially, if we first simulate all the first steps of all the processor, then all the second steps of all the processors, etc. This way, each step of the parallel computer requires  $N_{\text{proc}}$  steps of the sequential computer. Thus, what was computed on a parallel computer in time  $T_{\text{par}}$  can be computed on a sequential computer in time  $T_{\text{seq}} = N_{\text{proc}} \cdot T_{\text{par}}$ .

Due to the inequality (1), we have

$$T_{\text{seq}} \leq \frac{4}{3 \cdot \Delta V} \cdot \pi \cdot c^3 \cdot T_{\text{par}}^3 \cdot T_{\text{par}} = C \cdot T_{\text{par}}^4, \quad (2)$$

where we denoted

$$C \stackrel{\text{def}}{=} \frac{4}{3 \cdot \Delta V} \cdot \pi \cdot c^3.$$



So, if the fastest time that it takes for a sequential computer to solve a problem is  $T$ , the fastest time  $T_{\text{par}}$  that this same problem can be solved on a parallel computer is bounded by the inequality  $T \leq T_{\text{seq}} \leq C \cdot T_{\text{par}}^4$ , thus

$$T_{\text{par}} \geq C^{-1/4} \cdot T^{1/4}. \quad (3)$$

This implies that by using parallelization, we can speed up, at best, to the 4-th root of the sequential time. This is good, but not ideal: if the original sequential time  $T$  was exponential—as for NP-hard problems—the parallel time is still exponential.

**Extra dimensions: a brief reminder.** The above argument assumes that we live in a 3-dimensional space. However, according to modern physics, the requirement that quantum field theory is consistent implies that the dimension of space is at least 10; see, e.g., [2, 7, 8].

**Resulting challenge and what we do in this paper.** A natural question is: how does the presence of these extra spatial dimensions affect computations?

This is the question that we study in this paper.

## 2 First (Naive) Idea and Why It Is Naive

**A seemingly natural idea.** At first glance, the situation is straightforward: if instead of 3 spatial dimensions we have  $d > 3$  dimensions, then the volume of the area inside the sphere of radius  $R$  is equal to  $V = c_d \cdot R^d$  for some constant  $c_d$ . Taking into account that  $R = c \cdot T_{\text{par}}$ , we conclude that  $V = c_d \cdot c^d \cdot T_{\text{par}}^d$ . Thus, the number  $N_{\text{proc}}$  of processors is bounded by the number

$$N_{\text{proc}} \leq N_{\text{max}} \stackrel{\text{def}}{=} \frac{V}{\Delta V} = \frac{c_d}{\Delta V} \cdot c^d \cdot T_{\text{par}}^d.$$

Hence, this parallel computation can be simulated on a sequential computer in time

$$T_{\text{seq}} \leq N_{\text{proc}} \cdot T_{\text{par}} = \frac{c_d}{\Delta V} \cdot c^d \cdot T_{\text{par}}^d \cdot T_{\text{par}} = C_d \cdot T_{\text{par}}^{d+1},$$

where this time

$$C_d \stackrel{\text{def}}{=} \frac{c_d}{\Delta V} \cdot c^d.$$

So, instead of the previous rather-high lower bound  $T_{\text{par}} \geq \text{const} \cdot T_{\text{seq}}^{1/4}$ , we get a much better lower bound  $T_{\text{par}} \geq \text{const} \cdot T_{\text{seq}}^{1/(d+1)}$ , with  $d \geq 10$ .

**Why this idea is naive.** The above result looks good, but it is based on the simplified idea that extra spatial dimensions are similar to the current three ones. In reality,

the fact that we currently observe only three dimensions means that different spatial dimensions are different.

There are two possible approaches to how to explain that other dimensions are not yet observable. In this section, we describe these two approaches, and in the following sections, we analyze how these approaches affect computations.

**First approach: actual compactification.** The first natural approach is to conclude that since we cannot observe any changes in other spatial dimensions, this means that these dimensions are very small in size—e.g., that each of these dimensions represents not a line, but a circle of a small radius.

**Second approach.** The second natural approach is to assume that while all our processes are happening in a very small fragment of the additional dimensions, these dimensions actually have larger size—only due to some physical reasons, we cannot leave this small fragment. An analogy is when we are in a narrow valley between two mountain ranges: in principle, we can get out of this valley, but this requires climbing high mountains—and for that, we will need lots of energy and probably special equipment, which few of us have.

**What we will do now.** Let us see how both physically realistic versions of extra spatial dimensions can affect computations.

### 3 First Approach: How Actual Compactification Affects Computations

**It all boils down to computing the volume.** The above arguments about the limits to parallelization were based on computing the volume  $V$  of the inside of the sphere of radius  $R = c \cdot T_{\text{par}}$ , where  $c$  is the speed of light and  $T_{\text{par}}$  is the computation time. In the analysis of the 3-D situation, we used the formula for the volume of a sphere in the 3-D space. To see how the resulting calculations will change in the multi-D space, we need to find, for this space, what is the corresponding volume  $V$ .

**Let us compute this volume.** To find this volume, let us recall that the distance between the two points  $x = (x_1, x_2, \dots)$  and  $y = (y_1, y_2, \dots)$  in the multi-D space is equal to

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2 + (x_4 - y_4)^2 + \dots}$$

For reasonable computation time  $T_{\text{par}}$ , the radius  $R = c \cdot T_{\text{par}}$  is large, and thus, is much larger than the size  $s_e$  of each extra dimension: remember that this size is so small that we do not notice these extra spatial dimensions. So, the terms

$$(x_4 - y_4)^2, \dots$$

corresponding to differences in extra dimensions—and which are of order  $s_e^2$ —are much much smaller than the terms  $(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2$  describing the distance in the 3-D space. Hence, with high accuracy, we can safely assume that the distance between the two multi-D points is equal to the distance between their 3-D parts:

$$d(x, y) \approx d_3(x, y) \stackrel{\text{def}}{=} \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}.$$

Therefore, the set of all the points which are at distance  $\leq R$  from the user can be described as follows: we take all the points  $(x_1, x_2, x_3)$  from the corresponding 3-D sphere, and for each of these points, we consider all possible combinations  $(x_4, \dots)$  of additional spatial coordinates.

The size of each additional coordinate is  $s_e$ , and in a  $d$ -dimensional space, there are  $d - 3$  additional spatial coordinates. Thus, the overall volume of the additional part of  $s_e^{d-3}$ , and the overall volume of the sphere in  $d$ -dimensional space is equal to  $\frac{4}{3} \cdot \pi \cdot R^3 \cdot s_e^{d-3}$ .

**How many processors can we fit now?** The multi-D volume  $\Delta V$  of a processor can be obtained by multiplying its 3-D volume  $\Delta V_3$  by its volume  $\Delta V_e$  in the extra dimensions. If the size of the processor in additional dimensions is  $s_e$ , then we get the exact same number of processors as in the 3-D case, no gain at all from the existence of additional spatial dimensions.

However, if we manage to decrease the size of a processor in extra dimensions to less than  $s_e$ , so that the volume  $\Delta V_e$  of a processor in the extra dimensions is smaller than  $s_e^{d-3}$ , then, by dividing the overall multi-D volume by the volume of a single processor, we get the new value for the number of processors:

$$N_{\text{proc}} \leq N_{\text{max}} = \frac{V}{\Delta V} = \frac{\frac{4}{3} \cdot \pi \cdot R^3 \cdot s_e^{d-3}}{\Delta V_3 \cdot \Delta V_e} = \frac{4}{3 \cdot \Delta V_3} \cdot \pi \cdot R^3 \cdot \frac{s_e^{d-3}}{\Delta V_e}.$$

Since we consider the case when  $\Delta V_e < s_e^{d-3}$ , this number of processors is larger than the corresponding 3-D number of processors

$$\frac{4}{3 \cdot \Delta V_3} \cdot \pi \cdot R^3$$

by a factor of

$$C = \frac{s_e^{d-3}}{\Delta V_e} > 1.$$

**Conclusion for this approach.** In the first approach to multi-D space-time—when all extra dimensions are actually compactified—after an appropriate level of miniaturization, we will be able to get a  $C$  times increase in number of processors that we can fit into each area—and thus, in principle, a constant times computation speed-up.

*Comment.* This is not as spectacular as we could imagine based on the naive approach, but any speed up is good.

## 4 Second Approach: How It Affects Computations

**At first glance.** If we limit ourselves to the same small area of extra dimensions in which all observable processes take place, then we get the exact same situation as in the first approach—and thus, we can get the same constant times increase, where the constant depends on how successful we are in minituarizing our processors.

**But now we have another option.** However, in the second approach, we do not have to limit ourselves to the small area that contains all observable processes: there are other areas as well, it is just that these areas are difficult to reach: since going there requires a lot of energy, thus preventing usual particles from going there.

What if we apply this considerable amount of energy and reach these additional areas? What do we gain with respect to computations?

**First gain: all the promises of the naive approach turn out to be true.** If we are allowed to use a significant area in extra dimensions, then we can have all the advantages promised by the above-described naive approach: namely, instead of being able to fit  $\sim T_{\text{par}}^3$  processors into an area of radius  $R = c \cdot T_{\text{par}}$ , we can fit a much larger amount of  $\sim T_{\text{par}}^d$  processors. Thus, instead of the possibility to reduce the sequential computation time  $T_{\text{seq}}$  to  $T_{\text{par}} \sim T_{\text{seq}}^{1/4}$ , we can get a much more drastic speed-up  $T_{\text{par}} \sim T_{\text{seq}}^{1/(d+1)}$ .

**Interestingly, there is an additional speed-up.** The fact that *all* the processes are limited to a narrow area of values of extra spatial dimensions means, in effect, that this limitation is the property of the underlying space-time, not of any specific physical field. In other words, this means that the space-time is not as flat as the space-time of our usual 3D space—that would have enabled particles to easily go in all possible spatial directions—but rather curved.

According to General Relativity theory—the theory that describes curved space-time in modern physics—in a curved space-time, free particles move along geodesic lines, i.e., lines in which the resulting proper time  $\Delta s$  between the each two locations is the shortest possible. In terms of coordinate time  $t$ , this overall proper time can be computed by adding up proper times  $ds = \frac{ds}{dt} \cdot dt$  corresponding to different parts of the trajectory, i.e., as  $\Delta s = \int \frac{ds}{dt} dt$ . According to General Relativity, the ratio  $\frac{ds}{dt}$  is, in general, smaller than 1: in a gravitational field, all the processes slow down, and if this field is very strong—e.g., near a black hole—then it can slow down drastically: when the outside world measures 10 years, people near the black hole will only count several months.

In [3], we considered possible computational consequences of this effect in the 3D space. Interestingly, in the second approach to the multi-D cases, we have an additional possibility to use this effect. Namely, the fact that for all the particles, the optimal path is by going via the narrow zone of observable processes means that in this zone, the ratio  $\frac{ds}{dt}$  is much smaller than in the neighboring zones—just like the fact that the fastest way to get from two points in the US usually involves taking a freeway is an indication that the allowed speed on the freeway is larger than on all other roads.

For example, if we are in the vicinity of a gravitating body, where the ratio  $\frac{ds}{dt}$  is smaller than 1—and which is thus an analogue of a freeway—particles will tend to move close to this vicinity, which we observe as gravitational attraction. The stronger the gravitational field, the smaller the ratio  $\frac{ds}{dt}$  and thus, the more probable it is that the particles will bend towards this vicinity—so the larger the observed gravitational attraction.

In our multi-D case, the fact that in the neighborhood of our zone the value of the ratio is much larger than in the zone itself means that during the same time  $\Delta t$ , the proper time  $\Delta s$  in this neighborhood will be larger than in our zone. In other words, during the same coordinate time, the processor located in the neighborhood will be able to perform more operations than a processor that stays in our zone. Thus, we will get an additional speed-up.

**Conclusion for this approach.** In the second approach,

- not only we can have more processor working in parallel—by placing additional processors outside the narrow zone where the observable processes occur,
- but also the processors placed outside this zone will compute much faster than the ones in the zone, which will lead to an additional speedup.

**Acknowledgements** This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes).

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.

## References

1. Feynman, R., Leighton, R., Sands, M.: The Feynman Lectures on Physics. Addison Wesley, Boston, Massachusetts (2005)
2. Green, M.B., Schwarz, J.H., Witten E.: Superstring Theory, Vols. 1, 2, Cambridge University Press (1988)
3. Kosheleva, O., Kreinovich, V.: Relativistic effects can be used to achieve a universal square-root (or even faster) computation speedup. In: Blass, A., Cegielsky, P., Dershowitz, N., Droste, M., Finkbeiner, B. (eds.) Fields of Logic and Computation III, pp. 179–189. Springer (2020)

4. Kreinovich, V., Lakeyev, A., Rohn, J., Kahl, P.: Computational Complexity and Feasibility of Data Processing and Interval Computations. Kluwer, Dordrecht (1998)
5. Morgenstein, D., Kreinovich, V.: Which algorithms are feasible and which are not depends on the geometry of space-time. *Geoinformatics* **4**(3), 80–97 (1995)
6. Papadimitriou, C.: Computational Complexity. Addison-Wesley, Reading, Massachusetts (1994)
7. Polchinski, J.: String Theory, Vols. 1, 2, Cambridge University Press (1998)
8. Thorne, K.S., Blandford, R.D.: Modern Classical Physics: Optics, Fluids, Plasmas, Elasticity, Relativity, and Statistical Physics. Princeton University Press, Princeton, New Jersey (2017)