Fields Institute Communications 82

The Fields Institute for Research in Mathematical Sciences

Nicolas Fillion
Robert M. Corless
Ilias S. Kotsireas
Editors

Algorithms and Complexity in Mathematics, Epistemology, and Science

Proceedings of 2015 and 2016 ACMES Conferences

FIELDS

Springer

Fields Institute Communications

Volume 82

Fields Institute Editorial Board:

Ian Hambleton, *Director of the Institute*

Huaxiong Huang, *Deputy Director of the Institute*

James G. Arthur, *University of Toronto*

Kenneth R. Davidson, *University of Waterloo*

Lisa Jeffrey, *University of Toronto*

Barbara Lee Keyfitz, *Ohio State University*

Thomas S. Salisbury, *York University*

Juris Steprans, *York University*

Noriko Yui, *Queen's University*

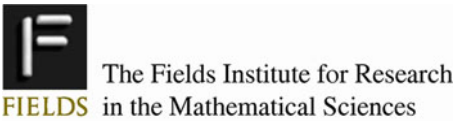
The Communications series features conference proceedings, surveys, and lecture notes generated from the activities at the Fields Institute for Research in the Mathematical Sciences. The publications evolve from each year's main program and conferences. Many volumes are interdisciplinary in nature, covering applications of mathematics in science, engineering, medicine, industry, and finance.

More information about this series at http://www.springer.com/series/10503

Nicolas Fillion • Robert M. Corless
Ilias S. Kotsireas
Editors

Algorithms and Complexity in Mathematics, Epistemology, and Science

Proceedings of 2015 and 2016 ACMES Conferences



Editors
Nicolas Fillion
Department of Philosophy
Simon Fraser University Philosophy
Burnaby, BC, Canada

Robert M. Corless
The Rotman Institute of Philosophy,
The Ontario Research Center for Computer Algebra and The School of Mathematical and Statistical Sciences
The University of Western Ontario
London, ON, Canada

Ilias S. Kotsireas
Department of Physics & Computer Science
Wilfrid Laurier University
Waterloo, ON, Canada

ISSN 1069-5265 ISSN 2194-1564 (electronic)
Fields Institute Communications
ISBN 978-1-4939-9050-4 ISBN 978-1-4939-9051-1 (eBook)
https://doi.org/10.1007/978-1-4939-9051-1

Library of Congress Control Number: 2019930430

Mathematics Subject Classification (2010): 00

© Springer Science+Business Media, LLC, part of Springer Nature 2019
This work is subject to copyright. All rights are reserved 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, express 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.

Cover illustration: Drawing of J.C. Fields by Keith Yeomans

This Springer imprint is published by the registered company Springer Science+Business Media, LLC, part of Springer Nature.
The registered company address is: 233 Spring Street, New York, NY 10013, U.S.A.

This proceedings is dedicated to the memories of Jon Borwein and Ann Johnson, distinguished participants of the ACMES meetings. Their contributions shall not be forgotten.

Preface

Mathematicians, scientists, and philosophers have historically been involved in a multidisciplinary endeavor of seeking to understand and represent aspects of reality with theories and models. This practice is nowadays more fruitful than at any other point before, thanks to an expanding understanding of the methods that allow us to extract information from the theories and models we advance. Indeed, as Alan Turing remarked, "the assumption that as soon as a fact is presented to a mind all consequences of that fact spring into the mind simultaneously with it [...] is a very useful assumption under many circumstances, but one too easily forgets that it is false." In other words, understanding the implications of scientific models and theories typically requires ingenious and skillful computation. Indeed, from a computational point of view, to have theoretical knowledge worthy of the name, we typically need to have an efficient computational algorithm to answer a range of questions, within specified constraints on time and computational resources. In some cases, the required computations are straightforward, and the theories are, in a sense, inferentially transparent, but this is the exception rather than the rule. Moreover, in order to assess the validity of results obtained thusly, one needs to interpret the computational task as a component of a broader modeling practice, which includes other important themes in applied mathematics, such as robustness under perturbation, inference from data, and uncertainty quantification.

This volume brings together papers that each contribute to this broad task in one way or another from disciplinary perspectives that include mathematics, computer science, philosophy of science, and history of science. The volume is based on the contributions made at the two ACMES conferences organized at Western University in 2015 and 2016, in which about 150 academics from multiple countries participated. ACMES (Algorithms and Complexity in Mathematics, Epistemology, and Science) is a multidisciplinary conference that covers a combination of numerical analysis and its underlying philosophy, computer algebra, reliability and uncertainty quantification, computation and complexity theory, error analysis, perturbation theory, experimental mathematics, scientific epistemology, machine learning, and foundations of mathematics, with the aim of furthering our understanding of the multifaceted role of mathematics in modern science. By bringing contributions from

researchers who approach the mathematical sciences from different perspectives, the volume aims to do so in a way that is informed by the state of the art in mathematics, scientific computing, and current modeling technique, with the hope that it leads to a self-reflective outlook on modern applied mathematics that draws from theory and practice and situates it in its proper philosophical, sociological, and historical context. In line with its multidisciplinary commitment, the conferences have been co-funded primarily by the Fields Institute for Research in Mathematical Sciences and by the Rotman Institute of Philosophy.

The collection features the work of distinguished scientists and philosophers. Its most important features are the depth of individual works and breadth of topics in computational mathematics and its underlying philosophy. The depth of the research contributions included in this paper goes hand in hand with their accessibility, as authors were asked to write for a multidisciplinary audience. Several papers are, in part, presented as tutorials for readers in other areas. Even though research productivity requires academics to focus on more or less narrow research areas, setting the core methods and results of a subdiscipline in a broader context also benefits researchers across disciplines.

Burnaby, BC, Canada — Nicolas Fillion
London, ON, Canada — Robert M. Corless
Waterloo, ON, Canada — Ilias S. Kotsireas

Contents

Contributors

James Robert Brown Department of Philosophy, University of Toronto, Toronto, ON, Canada

Shaoshi Chen KLMM, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China

Robert M. Corless The Rotman Institute of Philosophy, The Ontario Research Center for Computer Algebra and The School of Mathematical and Statistical Sciences, The University of Western Ontario, London, ON, Canada

Ernest Davis Department of Computer Science, New York University, New York, NY, USA

Nicolas Fillion Department of Philosophy, Simon Fraser University Philosophy, Burnaby, BC, Canada

Max Gunzburger Department of Scientific Computing, Florida State University, Tallahassee, FL, USA

Benjamin C. Jantzen Department of Philosophy, Virginia Tech, Blacksburg, VA, USA

Daniel Lichtblau Wolfram Research, Champaign, IL, USA

Yuri Matiyasevich St. Petersburg Department of V. A. Steklov Mathematical Institute, Saint Petersburg, Russia

Robert H.C. Moir Department of Computer Science, The University of Western Ontario, London, ON, Canada

Mark Wilson University of Pittsburgh, Pittsburgh, PA, USA

James Woodward University of Pittsburgh, Pittsburgh, PA, USA

Ethics and the Continuum Hypothesis

James Robert Brown

Abstract Mathematics and ethics are surprisingly similar. To some extent this is obvious, since neither looks to laboratory experiments nor sensory experience of any kind as a source of evidence. Both are based on reason and something commonly call "intuition." This is not all. Interestingly, mathematics and ethics both possess similar distinctions between pure and applied. I explore some of the similarities and draw methodological lessons from them. We can use these lessons to explore how and why Freiling's refutation of the continuum hypothesis might be justified.

Three decades ago Christopher Freiling [6] published a remarkable result. He showed that the continuum hypothesis (CH) is false. His way of achieving this is far removed from normal mathematical reasoning and much more closely resembled a thought experiment that one might find in physics. Of course, it had to be different from normal mathematical reasoning, since the continuum hypothesis is demonstrably not provable, nor is its negation.

What I want to do in this article is relate Freiling's argument to typical reasoning inside ethics. It will turn out that there are a couple of different features of ethics that relate in interesting ways to mathematics in general and the continuum hypothesis in particular.

First, I will begin with a brief account of Freiling's work. Second, I will discuss the distinction between pure and applied mathematics as seen by philosophers and as seen by mathematicians. They are interestingly different accounts of the distinction. Third, I will look at the pure and applied distinction when it comes to ethics. Surprisingly, the distinction there is remarkably similar to the pure-applied distinction made by mathematicians, though not by philosophers of mathematics. I will try to relate this to the techniques used by Freiling, with the aim of legitimizing and better understanding those techniques. Finally, there's still an open question

J. R. Brown (✉)
Department of Philosophy, University of Toronto, Toronto, ON, Canada
e-mail: jrbrown@chass.utoronto.ca

© Springer Science+Business Media, LLC, part of Springer Nature 2019

N. Fillion et al. (eds.), *Algorithms and Complexity in Mathematics, Epistemology, and Science*, Fields Institute Communications 82,
https://doi.org/10.1007/978-1-4939-9051-1_1

about why this technique works, assuming that it does. I will attempt the beginning of an explanation by appeal to another feature of contemporary ethics, the distinction between so-called thick and thin concepts.

1 The Continuum Hypothesis

CH is the famous claim made by Cantor that the real numbers are the first uncountable infinite set, $|R| = \aleph_1$. Since $|R| = 2^{\aleph_0}$, we can express CH as the claim that $2^{\aleph_0} = \aleph_1$. (The Generalized Continuum Hypothesis, which won't concern us here, is the claim that $2^{\aleph_n} = \aleph_{n+1}$.)

CH was the first of Hilbert's famous problems. Many great mathematicians have tried but failed to prove (or refute) it. Gödel [7, 8] showed that CH is consistent with the rest of set theory, so it cannot be refuted. Cohen [3] showed that the negation of CH is also consistent with standard set theory. These two results taken together give us the independence of CH (assuming the consistency of ZFC). Some might make the additional assumption that all of mathematics can be captured by standard set theory; if so, then CH is independent from the rest of mathematics. A remarkable fact.

We now go through the various steps of Freiling's argument. First, we assume Zermelo-Frankel set theory with the Axiom of Choice (ZFC). We will further assume that CH is true; we are aiming for a *reductio ad absurdum*.

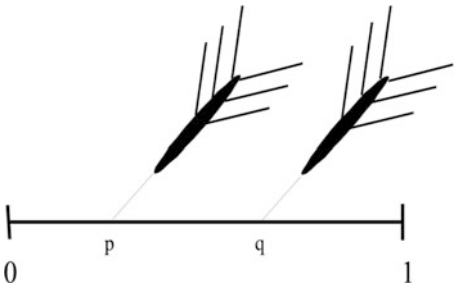
ZFC implies that every set can be well ordered. Thus, the real interval [0, 1] can be well ordered. Since we have assumed CH, it follows that the size of this well ordered set is $\aleph_1$.

The real numbers in [0, 1] are paired up with the ordinals, though not, of course, in the standard ordering, but instead in a well ordering:

$$0, 1, 2, 3, \ldots, \omega, \omega + 1, \ldots, \omega 2, \omega 2 + 1, \ldots, \omega 3, \ldots, \omega^2, \ldots, \omega^\omega, \ldots, \omega^{\omega^\omega}, \ldots, \epsilon_0, \ldots, \Omega$$

Recall that ω is the set of all finite ordinals and Ω is the set of all countable ordinals. The finite ordinals correspond to the finite cardinals; ω corresponds to the cardinal number $\aleph_0$, as do all the others ωs up to, but not including, Ω, which corresponds to $\aleph_1$. CH is the assumption that $|R|$, or $|[0, 1]|$ which is the set we're concerned with, has length $\aleph_1$, or, in other words, that it can be paired with the ordinals up to Ω.

Imagine two people, Alice and Bob, toss darts at the line [0, 1]. Alice's dart hits the real number p and Bob's hits q. Which of these numbers is earlier in the well ordering? We make three important assumptions about the tosses. They are *random* and each point in [0, 1] is as likely to be hit as any other. The tosses are *symmetric* in that it is a matter of indifference whether Alice or Bob tossed first. And finally, the tosses are *independent* of one another.



Alice can now make the following simple argument that Bob's number q must be later in the well ordering than the number p that she picked out with her dart. p corresponds to some ordinal that is an initial segment of the ordinal Ω. But any such initial segment is countable. According to measure theory, the measure of such a set is 0. The link between measure theory and probability theory gives us the probability of that event to be 0. And the probability of being later in the well ordering is 1. Of course, it is logically possible for Bob to hit a real number earlier in the well ordering, but the chance is zero.

Because of the randomness, symmetry, and independence of the two darts, Bob can make the same argument, namely, that Alice's dart hit a number p that must be later than his number q.

Now we have two good arguments, one says p is earlier than q with probability one and the other says q is earlier than p with probability one. This is absurd, yet in any pair of tosses, one must be earlier in the well ordering. (In case the two darts hit the same point, $p = q$, we simply toss again.)

Aside for those worried about non-measurable sets. Instead of saying the probability of being later in the well ordering is one, say the probability of being earlier is zero. Since the set of earlier points are countable, they are measurable, and, of course, have measure zero. This is enough for the absurdity.

We blame CH for leading to this absurdity. Thus, $|R| > \aleph_1$. I'll leave to readers to imagine how three darts might be used to undermine a new hypothesis that $|R| = \aleph_2$, then the effects of four darts on the hypothesis $|R| = \aleph_3$, and so on.

This is Freiling's remarkable result. Of course, in any *reductio ad absurdum* argument, different premisses might be blamed. Freiling (in private correspondence) now thinks well ordering should be discarded instead of CH. This is certainly a plausible alternative, but for my purposes it does not matter, since my concern here is with the use of a thought experiment to justify, in principle, a mathematical result. Whether the dart thought experiment justifies the rejection of CH rather than the rejection of well ordering or vice versa is a secondary detail.

Like any thought experiment, there will be objections in principle, usually based on how realistic we demand the thought experiment to be. In this case, we assume the darts are infinitely thin and can pick out a single point. We are also here prepared to talk about probabilities of hitting individual points, whereas in realistic cases we would demand intervals. These are typical of the perennial concerns about thought

experiments and what counts as legitimate reasoning in various disciplines. Keep in mind some famous examples: Einstein running as fast as a beam of light; Galileo sailing on a ship at sea that encounters no tossing and turning; Newton's bucket in an otherwise empty universe. Unrealistic thought experiments can be wonderfully productive. Freiling's dart example is one of these.

Prominent mathematicians such as David Mumford [15] and Yuri Manin [14] have endorsed the dart argument, but for the most part the mathematical community has not been won over. This reluctance, I think, is a mistake. Perhaps by viewing the result in a new light, it will seem more plausible. I will return to CH after major a detour.

2 Pure Versus Applied Mathematics

When talking about mathematics, the distinction between pure and applied inevitably arises. It's a curious fact that philosophers make a distinction that is perfectly clear and objective, but it is quite different from that offered by mathematicians. Typically, philosophers would cite a simple example: They might say, for instance, that "$2 + 2 = 4$" is pure mathematics, while "2 apples + 2 apples = 4 apples" is applied. More generally, they might claim, mathematics is pure when it makes no reference to anything nonmathematical; as soon as it involves the physical or financial realm, it is applied. As I said, the philosophers' distinction is perfectly clear and objective. It is not at all like the typical mathematicians' account, as we shall soon see. By the way, being more objective is not the same as being better. That is an entirely different matter.

A working mathematician is much more likely to say that what makes mathematics pure is the kind of interest we have in it. Physics makes extensive use of mathematics that is often dull and boring, but on occasion it also makes use of it in ways that are mathematically interesting. When the latter happens, it is pure mathematics, even though there is lots of nonmathematical stuff involved. The singularity theorems of Hawking and Penrose concern black holes, but are mathematically of great interest to differential geometers, who are as likely as not to be indifferent to the physics involved. The travelling salesman problem is tedious, if your concern is finding the shortest route for the salesman to cover a territory, but to a mathematician concerned with computational complexity it is highly interesting. Such examples can be generated *ad nauseam*.

There have been many provocative pronouncements about applied mathematics coming from champions of the pure sort. Paul Halmos remarks, "...there is a sense in which applied mathematics is just bad mathematics. It's a good contribution. It serves humanity. But just the same, much too often it is bad, ugly, badly arranged, sloppy, untrue, undigested, unorganized, and unarchitected mathematics." (1991, 18) G.H. Hardy, in his famous self-portrait, *A Mathematician's Apology*, declared that applied mathematics is "repulsive, ugly and interminably dull." He also famously remarked: "I have never done anything 'useful.' No discovery of mine has

made, or is likely to make, directly or indirectly, for good or ill, the least difference to the amenity of the world. . . " (1944, 150) Halmos and Hardy are amusing snobs, but we should not lose sight of the fact that they are indeed snobs and they should not be emulated.

Hardy had trouble maintaining the pure-applied distinction, so he switched to "real mathematics," implicitly conceding that the pure-applied distinction is not a happy one. As for real mathematics, according to Hardy, it includes number theory, of course, and classical analysis, but it also includes relativity and quantum mechanics, which a typical philosopher would call applied. In other words, according to Hardy, real mathematics is aesthetically pleasing; it is the fun stuff, whether or not it involves nonmathematical entities. Minus the snobbery, this is what mathematicians would call pure.

We might cheerfully use terms such as "mathematically interesting" and "mathematically important," but we can't get away from the fact that these are much more subjective notions being used to make the pure-applied distinction than the philosophers' characterization. It's not, however, wholly subjective to the point where we can say nothing about it. We can, in fact, make partial sense following Hardy, himself. "The 'seriousness' of a mathematical theorem lies, not in its practical consequences, which are usually negligible, but in the significance of the mathematical ideas which it connects. We may say, roughly, the mathematical idea is significant if it can be connected, in a natural and illuminating way, with a large complex of other mathematical ideas." (Hardy [10], 89) This helps a bit, but can we do better?

There is a wide consensus that the Riemann hypothesis is both interesting and significant. Why? I'm not sure how to justify the "interesting" claim—though I don't doubt it—but the reason that the Riemann hypothesis is called "important" or "significant" is straightforward. It implies a great many things elsewhere in mathematics, such as facts related to the distribution of prime numbers. But if asked why the distribution of primes is important, I would resort to saying it's a brute fact or that it is connected to something else that is highly interesting and significant. So, we're back where we started with the subjective idea of mathematical interest.

The fact that we are stuck with an apparently subjective notion, however, does not mean it is hopelessly subjective or useless. Though I would be hard pressed to justify the claim, I am quite sure some pieces of music are objectively very much better than others. In any case, we don't have to solve this problem. (I just walk away when someone plays bad music. We can do the same with mathematics that bores us.) Subjective or not, the mathematicians' distinction between pure and applied is clear and coherent enough for us to use. This much at least is evident, since there is a stable consensus on examples within the mathematical community.

So we have two perfectly legitimate senses of pure and applied, the philosophers' distinction and the mathematicians' distinction. They are not really rival conceptions; they merely have different aims. Platonists, who think the physical real and the mathematical realm are separate, will consider the philosophers' distinction between pure and applied exactly right. Those interested in mathematical methodology will

rightly see that the mathematicians' distinction is highly fruitful. Let's keep both in mind as we move from mathematics to ethics.

Before leaving this section, a confession of sorts is in order. I have described the mathematicians' distinction between pure and applied. It would be fair to ask: Is the distinction I have described between pure and applied mathematics that of the pure mathematician or of the applied mathematician? Clearly, the answer is pure. I think it might also be the account an applied mathematician might also give, but I am less confident. Significant qualification would likely be in order. If the account is different, it would make things a bit messier, but I think it would not undermine what I have said about the pure mathematicians' distinction between pure and applied, nor, most importantly, would it undermine the dart thought experiment.

3 Ethics for Mathematicians

There are several branches of ethics: metaethics, normative ethics, descriptive ethics, theoretical ethics, practical ethics, applied ethics. I will distinguish pure and applied. Though this is an important distinction, the methods of reasoning are the same in each. I will illustrate with a pair of well-known examples (well-known to philosophers, that is, as well known as elementary calculus is to mathematicians).

Before continuing, I should mention that empiricism, the doctrine that all knowledge comes through the senses, has had trouble with two things: mathematics and ethics. This is frequently noted, but seldom developed in any detail.[1] The remarkable thing is the depth of similarity between these two distinct disciplines.

Judith Thomson [17] wrote a famous paper defending a woman's right to abortion. She began with a standard anti-abortion argument and made the case for rejecting it. The standard argument runs like this:

1. A fetus is an innocent person.
2. Innocent persons have a right to life.
3. Abortion kills a fetus.
4. Therefore, abortion is morally wrong.

This is a widely accepted argument, free of any religious considerations, so potentially acceptable to anyone. A standard way to overturn such an argument is to construct a parallel argument that is obviously faulty. Thomson did this with her famous violinist thought experiment.

Imagine a famous violinist who is near death. A society of music lovers has searched through medical records and found that you are the only person in the world with the right body type who could save the violinist's life. While you are asleep one night they sneak into your bedroom and attach the violinist, who by

[1] There are a few notable exceptions, such as some of the authors in [12], and especially [2], [5], and [13]. These three take a different view from mine.

this time has fallen into a coma and knows nothing of this. When you wake in the morning you are shocked to find yourself attached. The music lovers explain why they did this and inform you that the cure this will take about 9 months. Just as you are about to unhook the violinist, the music lovers make the following argument:

1. A violinist is an innocent person.
2. Innocent persons have a right to life.
3. Unhooking the violinist will kill him.
4. Therefore, unhooking him is morally wrong.

At this point we get near universal agreement that it is morally permissible for you to unhook the violinist. You have no obligation to remain hooked up. It would be very generous of you to go through with the 9 month process and all that it entails, but you are not immoral for failing to be supererogatory. In short, it is your body and you do not have to share it.

The final step is obvious. We note that the two arguments have the same form. Since the violinist version is clearly faulty, the fetus version must be similarly faulty. Thus, we conclude: the standard anti-abortion argument is not a good argument.

The thought experiment brought out something crucial: the concept "right to life" is not the same as "right to what is needed to sustain life," which the initial argument had implicitly assumed. The fetus, like the violinist, has the former but not the latter, namely, use of the mother's body.

As you might imagine, debate did not come to a halt after the violinist thought experiment. There are lots of other arguments for and against abortion. And, of course, there are critics of Thomson's thought experiment. I have no interest here in describing the debate details; I use Thomson's work as an example of a kind of reasoning that is fairly typical in applied ethics.

Now, another example that is just as famous, the trolley problem [4]. Imagine a runaway trolley that is heading down the track toward five people who cannot get out of the way. If you throw a switch, you will send the trolley down a siding, sparing the lives of the five people. Unfortunately, there is one person on the siding who cannot get out of the way. Should you throw the switch? The near universal answer is that yes, you should throw the switch. It is better to kill one in order to save five.

Now we are going to complicate things a bit. Imagine the runaway trolley again with five people on the track who will be killed if hit by the trolley. This time there is no siding to redirect the trolley. Instead you and a very large person are on a bridge over the track. If you push the big guy off the bridge onto the track, you will stop the trolley and save the five people. He will be killed. Should you push the big guy? The moral principle seems to be the same: It is better to kill one to save five.

And yet, there is serious revulsion at the thought of pushing someone off the bridge to stop the trolley, unlike redirecting a train to someone who cannot get out of the way. Why? There is a huge literature on this; it is one of the leading topics of interest in contemporary ethics. Once again, I have no interest in trying to settle the problem here. I merely present it as an example of pure ethics and the kind of reasoning that goes into tackling issues.

Abortion and euthanasia are applied. The trolley problem is pure. I would not claim that the boundary between them is sharp, but there is a simple rule of thumb to distinguish them. Philosophical works on abortion are about abortion, but philosophical works about runaway trolleys are not about trolleys. Philosophical papers on abortion typically culminate, if only implicitly, in policy recommendations that abortion should or should not be permitted. Trolley papers are about utilitarianism and its limits, not which policies municipal councils should adopt and impose on their citizens when they see a trolley on the loose.

The moral for us about ethical reasoning is simple: The methods of ethics, whether pure or applied, are the same (at lest over a wide range of cases). Thought experiments and visual reasoning are allowed in both. So, what has this to do with mathematics?

4 Mathematical Methods

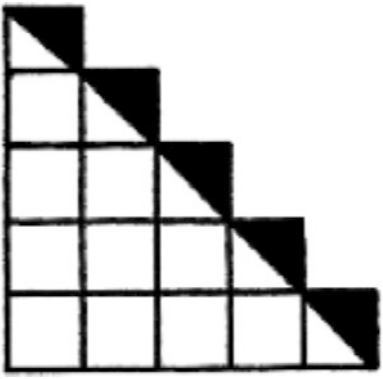
Ask anyone about evidence in mathematics, and you will very likely get the response: Proofs provide evidence. And we might also hear that *only* proofs provide evidence. This is not a good response for several reasons.

First, proofs are some sort of derivation based on axioms, but where do the axioms come from? We cannot prove axioms, except in the trivial sense that $P \rightarrow P$. Are we making them up out of whole cloth (perhaps guided by habit and utility)? Of course, we sometimes explore arbitrary formal systems, but if we thought that was true of all mathematics, many of us would lose interest. We need an account of the origin of axioms and other first principles that allows us to rationally believe them. They need not be certain, but at least plausible. Second, are the main concepts correctly defined? There was a time when the theorem "All functions are continuous" was proved. Of course, that is laughable today. What changed? The concept of a function evolved into today's arbitrary association between two sets, allowing, for instance, the Dirichlet function $f(x) = 1$ or 0, depending on whether x is rational or irrational. The concept of set, for instance, as formulated by Cantor as an arbitrary collection, led to paradoxes, and so was restricted. But there are still arguments today about how to correctly understand the concept. Third, speaking of set theory, must a canonical proof be in set theory? When in doubt we sometimes reconstruct proofs inside set theory, but it often seems artificial. Finally, are there other ways to legitimately justify theorems? If so, what might they be?

Let's consider a few interesting examples. First, a picture proof. I leave readers to figure out for themselves (if you have not seen it somewhere already) how the following picture works as evidence for the number theory theorem. After you have figured it out, ask yourself if the picture proof works, and is it as convincing as the standard proof by induction?

Theorem $1 + 2 + 3 + \ldots + n = n^2/2 + n/2$

Proof



Next, the pigeonhole principle, which says: *If there are* $n + 1$ *pigeons distributed in n pigeonholes, then at least one hole must have at least two pigeons.* The principle is obvious. It can be derived in set theory, but there is no need, since such a proof would not increase our confidence. The principle is important in combinatorial mathematics. By the pure-applied distinctions we presented above, philosophers would call it applied, since it involves non-mathematical entities, but mathematicians would call it pure, since it is not about pigeons, though expressed in picturesque language.

Mark Kac wrote a famous paper, "Can One Hear the Shape of a Drum?" [11]. If we know the shape, we can calculate the overtones. Kac is posing an inverse problem: If we know the overtones made by the drum, can we infer the shape of the drumhead? Once again, philosophers would call it applied, since it involves non-mathematical entities, while most mathematicians would call it pure, since it is about interesting mathematics. Most people who think about the problem have no particular interest in drums. The question is really about manifolds with certain elastic properties; can we infer the eigenvalues of a specific Laplacian? Kac did not know the answer at the time; it turned out to be no. Either way would have been interesting.

These examples are clearly what most mathematicians would call instances of pure mathematics; they are not about drums, pigeons, or geometric shapes, even though these figure in the examples. There are many more examples like them, which readers could easily supply for themselves.[2]

Of course, context can matter. Cases I am calling pure could be applied in the right circumstances. A musician who heard a strange drum sound might go to her friendly neighbourhood mathematician for advice on building a drum that could make the desired sound. A town with trollies regularly braking free, might consider passing a local ordinance requiring citizens to throw switches in some

[2]For more on this see [1].

circumstances. The trolly problem, it turns out, is on the edge of becoming an important applied problem for self-driving cars. How should they be programmed to swerve in emergency situations? None of this detracts, however, from the main point about these examples and how they are typically used.

5 The Moral

The important thing about these examples—both ethical and mathematical—is the moral we can draw: The mathematicians' distinction between pure and applied mathematics is the same as the ethicists' distinction between pure and applied ethics. Perhaps not everywhere, but they are certainly the same in a wide variety of cases. Mathematical work is no more about drums or pigeons than the trolley problem is about trolleys. And the methods of pure mathematics and applied mathematics are the same, just as the methods of pure ethics and applied ethics are the same. In addition to traditional proofs, intuitions, thought experiments, and visual arguments should be considered legitimate sources of evidence. This is the central moral of these examples.

But aren't such techniques sometimes misleading and produce results that are downright wrong? Yes. But we were never completely free of this unfortunate possibility. Regardless of the highest standards of rigour, there is, for instance, no way to be sure important current definitions won't change. As mentioned above, the definition of function changed, overturning the theorem that all functions are continuous. The proof was faultless.

Finally, why can't there be other legitimate ways to justify theorems? In his Gibbs lecture Gödel suggests a different outlook.

> If mathematics describes an objective world just like physics, there is no reason why inductive methods should not be applied in mathematics just the same as in physics. The fact is that in mathematics we still have the same attitude today that in former times one had toward all science, namely we try to derive everything by cogent proofs from the definitions (that is, in ontological terminology, from the essences of things). Perhaps this method, if it claims monopoly, is as wrong in mathematics as it was in physics. (1951, vol. III, p. 313)

Once we give rein to Gödel's suggestion, the range of possibilities will be enormous. Jonathan Borwein's experimental mathematics, for instance, is a hugely important example.[3] The possibilities could also include thought experiments, just as we find in physics. For instance, Galileo was able to establish that all bodies fall at the same rate, not by empirically measuring their rate of fall, but through a strikingly beautiful thought experiment.

Aristotle (and common sense) claimed that a heavy cannon ball falls faster than a light musket ball ($H > L$). Thus, a body consisting of the cannon ball and the

[3] See Borwein's home page for his several books, articles, and various projects: https://www.carma.newcastle.edu.au/jon/.

musket ball attached with string ($H + L$) would be heavier than the cannon ball alone, so should fall faster ($H + L > H$). However, the light musket ball would act as a drag on the cannon ball, slowing it down ($H > H + L$). Now we have a contradiction, which can be resolved by having all bodies fall at the same rate ($H + L = H = L$). It is a spectacular result achieved by thinking things through in a way that is similar to much mathematical reasoning.

To take Gödel's advice and do mathematics in the way we do physics would not mean that we should hold up a ruler to the diagonal of a unit square to see if the length is an irrational number. It would be more like the use of sophisticated statistical techniques, computer proofs and simulations, and thought experiments like those of Galileo, Einstein, and others.

6 Facing Problems

How do we know when a technique will work and when it won't? It would be a mistake to think that scientific or mathematical methods are presented to us on a platter. They are hard-won and can be just as controversial as the theories and theorems they are used to justify. Telescopes and microscopes were controversial until we got a good idea of the relevant optics involved and learned to distinguish what is real from artifacts of observation, due, for instance, to staining. Most of us gladly accept proofs by *reductio ad absurdum*, but not constructivists. Most accept use of the Axiom of Choice, but some are still reluctant. The use of computers is gaining ground, but there are holdouts.

Thought experiments in physics, mathematics, or ethics can be highly controversial. How do we know they are reliable? Are they influenced by culture, gender, or various other hidden biases and values? Much work is being done on this by philosophers and cognitive scientists. For instance, recall the trolley problem. One intuition says we should throw the switch, saving five, but killing one. An opposing intuition takes hold when we consider pushing the big guy into the path of the trolley, even though we would be able to save five. Joshua Greene [9] investigated this using a fMRI. Subjects were asked to think about the trolley problem, first the switch-throwing case, then pushing the big guy case. Interestingly, most moral reasoning happens in one part of the brain, but when asked to consider pushing the big guy, the emotional part of the brain lights up. Greene uses this fact to conclude that our reliable moral faculties work well in most cases, but get confused when the problem we are thinking about involves coming into physical contact with someone. The upshot, according to him, is that it is morally correct to push the single big guy to save five others. The fact that we feel otherwise can be explained away.

I am not endorsing Greene's view, which is highly controversial. I am citing it as an example of the kinds of things we could do to check on the reliability of intuition. With this in mind, we can now return to CH and the darts thought experiment.

7 How Does the Refutation of CH Work?

Freiling's refutation of CH relied on three central concepts: *randomness, symmetry*, and *independence* of the two dart throws. But these three concepts can all be defined inside standard set theory. That suggests that there should be a reconstruction of the informal dart thought experiment that would pass all normal tests of rigour. We know, however, that this is not possible, since CH is demonstrably independent of the rest of set theory. This is a puzzle. Now back to ethics, this time for the beginning of an explanation.

I'm going to borrow something else from ethics, this time the distinction between thick and thin concepts. We can start with the usual fact-value distinction. Factual concepts such as *tree, electron*, and *negative charge* are *thin concepts*; they belong wholly to the factual realm. Similarly, value concepts such as *good, beautiful*, and *wicked* are also thin, since they belong wholly to the value realm. Normally we think of concepts and statements as clearly factual or clearly evaluative, but a number of highly useful concepts have both factual and evaluative content; they are known as *thick concepts*.[4] If I say "Bob is healthy" I am saying something about Bob's biological state that is factual; I am also expressing an evaluation of that state, namely, that it is a good state to be in. Because *healthy* combines both factual and evaluative aspects, it is a thick concept. "Alice is good" is an expression of pure value, hence thin. But when I say "Alice was courageous during the war," I am reporting factually on her actions and simultaneously I am expressing an evaluation of them. Hence, *courageous* is a thick concept. So are *coward, cruel, kind, truthful, duplicitous, zealous, treacherous, brutal, grateful*, and on the list goes. Once we get the hang of it, we can spot thick moral concepts everywhere. Outside of ethics, there seem to be epistemic or methodological thick concepts, e.g., *simple, coherent, explanatory power*. And culinary thick concepts: *tasty, bland, delicious.* Slurs, such as racist, sexist, and homophobic terms, are arguably thick concepts, as well.

The key to understanding thick concepts is that they draw on two realms. Thick moral concepts involve facts and values, but thick concepts could arise elsewhere. They could arise anywhere there are distinct realms with thin concepts in each and additional concepts that somehow overlap. With this in mind, let us now ask the question: Could some of our concepts in the sciences be thick in the sense of simultaneously combining both physical and mathematical aspects? That is, could a single concept be simultaneously physical and mathematical, in spite of the two distinct realms?

In his famous textbook, *Calculus*, Michael Spivak says,

> In physics the second derivative is particularly important. If $s(t)$ is the position at time t of a particle moving along a straight line, then $s''(t)$ is called the acceleration at time t. Acceleration plays a special role in physics, because as stated in Newton's laws of motion, the force on a particle is the product of its mass and its acceleration. Consequently you can feel the second derivative when you sit in an accelerating car. [16, 159]

[4] For more detail, see [19], an early and influential source.

In the most obvious sense this would seem to be absurd. One can no more feel derivatives than one can smell infinite series or taste tangent spaces. We have no sensory contact with abstract entities. My initial inclination is to say we feel accelerations and the resistance of inertial forces, both of which are physical, not mathematical entities. We model physical acceleration with the second derivative, which is an abstract entity, not something we could actually feel.

And yet there seems to be something vaguely right about what Spivak says. Many working physicists and engineers are at home talking this way. How could this be possible? Why might Spivak be right to talk this way? Because acceleration is (or has become) a thick concept. So are velocity, force, mass, and several other notions. We have learned to employ these concepts in a way where calling (physical) acceleration a (mathematical) derivative seems wholly right and natural, just as we simultaneously describe and prescribe Alice's behaviour when we call her courageous, and just as we describe and prescribe Bob's cooking when we say it is delicious.

Just to be clear, thick concepts are drawn from two realms. In the ethical case, the realms are facts and values. In our case thick mathematical concepts are drawn (typically) from the physical realm and the pure mathematical realm, pure in the philosophers' sense of pure.

William Thurston, one of the great mathematicians of recent times, was part of a heated debate about the relations between mathematics and physics. He remarked on various ideas of a derivative.

> People have very different ways of understanding particular pieces of mathematics. To illustrate this, it is best to take an example that practicing mathematicians understand in multiple ways, but that we see our students struggling with. The derivative of a function fits well. The derivative can be thought of as:
>
> (1) Infinitesimal: the ratio of the infinitesimal change in the value of a function to the infinitesimal change in a function.
> (2) Symbolic: the derivative of x^n is nx^{n-1}, the derivative of $\sin(x)$ is $\cos(x)$, the derivative of $f \circ g$ is $f' \circ g \times g'$, etc.
> (3) Logical: $f'(x) = d$ if and only if for every ϵ there is a δ such that when $0 < |\Delta x| < \delta$, $|\frac{f(x+\Delta x)-f(x)}{\Delta x} - d| < \epsilon$.
> (4) Geometric: the derivative is the slope of a line tangent to the graph of the function, if the graph has a tangent.
> (5) Rate: the instantaneous speed of $f(t)$, when t is time.
> (6) Approximation: The derivative of a function is the best linear approximation to the function near a point.
> (7) Microscopic: The derivative of a function is the limit of what you get by looking at it under a microscope of higher and higher power.
>
> This is a list of different ways of thinking about or conceiving of the derivative, rather than a list of different logical definitions. [18, 164]

In terms of thick and thin, I would say Thurston's (1), (2), (3), and (6) involve thin concepts and are straightforwardly pure mathematics. I'm not sure what to make of (7), which seems rather metaphorical, though the metaphor might be useful

pedagogically.[5] The interesting cases are (4) and (5). I think (5) exemplifies what I mean by a thick mathematical concept, because the physical concept of speed is built into the characterization of the derivative. (4) is arguably a thick concept, too. It intertwines the algebraic concept of derivative with the geometric concept of a tangent. This is within mathematics, but applications of mathematics are certainly not confined to the natural sciences.

Those who teach physics or engineering might have a better sense of this. They teach thick concepts. That is, they take simple physical situations that are easy to understand and show how the mathematics applies. Eventually, it becomes easy and natural for the student; it is not just learned but internalized. Teaching thick mathematical concepts is similar to teaching operational definitions of physical concepts. Measuring instruments, to take a striking example, become transparent. After we learn how to use it, a thermometer does not give us a reading from which we then infer a temperature. We directly "see" the temperature. This may not be precisely like thick mathematical concepts, but there is some kinship, which I hope aids in grasping my main point. Another consideration (due to the physicist John Sipe in conversation) is that lots of concepts in theoretical physics are a 'mishmash' and so must be thick, since they are not clearly understood. In the future we might have a better grip on them, but now they are not thin or easily analyzable into thin concepts.

With the idea of thick and thin concepts under our belt, both in ethics and in mathematics, let's return to the central concepts in the refutation of CH: *random, symmetric, independent*. At this point readers can probably guess what I have in mind. I will propose that they are not thin mathematical concepts, but rather are thick. They have both mathematical and physical content. The Gödel-Cohen independence result guarantees that we cannot use thin versions of these notions to refute CH. But richer thick versions (of at least one of them) might work. I think that this is what is indeed happening. If this is correct, it is both an explanation and a justification. Of course, it is not a rock-solid explanation or justification, but it is not altogether negligible. Freiling's dart thought experiment is plausible, but one might remain sceptical. After all, how does it work, given that there is no reconstruction of the argument using mathematical concepts? If randomness, etc. are thick mathematical concepts, then that fact could provide an explanation of how the argument works. And given that there is an explanation of the inner workings of the argument, it should increase our confidence in the correctness of that argument.

Yuri Manin seems to believe something similar when he remarks, "Freiling's argument appeals directly to our physical intuition, and is best classified as a thought experiment. It is similar in nature to some classical thought experiments in physics, deducing e.g. various dynamic consequences from the impossibility of *perpetuum mobile*." [14]

[5] Rob Corless points out that with computers we can "zoom in" on a function and get a better look at the slope. So, (7) is arguably not so metaphorical, after all.

If this approach is right, it not only explains why Freiling's refutation of CH is successful, but it also supports a significant liberalization of the idea of evidence and legitimate methods in mathematics, the very thing Gödel wanted.

Acknowledgements Thanks to Paul Bartha, Philipp Berghofer, Mark Colyvan, Rob Corless, Nic Fillion, Chris Freiling, Tom Hurka, Tracy Issacs, Mary Leng, Kathleen Okruhlik, Debbie Roberts, Zvonimir Šikić, John Sipe, Alan Sokal, and Harald Wiltsche for discussions of various issues including information about ethical intuitions and the thin-thick distinction. Thanks also to an anonymous referee who provided several useful remarks. Finally, I am grateful to various audiences who heard versions of this material, especially at the ACMES conference at The University of Western Ontario, May 12–15, 2016.

References

1. Brown JR (2017) Proofs and guarantees. Math Intell 39(4):47–50. https://doi.org/10.1007/s00283-017-9730-1
2. Clark-Doane (2012) Morality and mathematics. Ethics 122:313–340
3. Cohen P (1963) Set theory and the continuum hypothesis. W.A. Benjamin, New York
4. Foot P (1958) Moral arguments. Mind 67(268):502–513
5. Franklin J (2004) On the parallel between mathematics and morals. Philosophy 79:97–119
6. Freiling C (1986) Axioms of symmetry: throwing darts at the real number line. J Symb Log 51:190–200
7. Gödel K (1938) The consistency of the axiom of choice and the generalized continuum hypothesis. Proc Natl Acad Sci U S A 24:556–557
8. Gödel K (1939) Consistency proof of the axiom of choice and the generalized continuum hypothesis. Proc Natl Acad Sci U S A 25:220–224
9. Greene JD, Sommerville RB, Nystrom LE, Darley JM, Cohen JD (2001) An fMRI investigation of emotional engagement in moral judgment. Science 293:2105–2108
10. Hardy H (1944) A mathematician's apology. Cambridge University Press, Cambridge
11. Kac M (1966) Can one hear the shape of a drum. Am Math Mon 73(4, part 2):1–23
12. Leibowitz UD, Sinclair N (eds) (2016) Explanation in ethics and mathematics: debunking and dispensability. Oxford University Press, Oxford
13. Leng M (2016) Naturalism and placement, or, what should a good Quinean say about mathematical and moral truth? Proc Aristot Soc cxvi, Part 3. https://doi.org/10.1093/arisoc/aow014
14. Manin Y (2002) Georg Cantor and his heritage. arXiv:math.AG/0209244 v1
15. Mumford D (2000) Dawning of the age of stochasticity. In: Arnold V (ed) Mathematics: frontiers and perspectives. American Mathematical Society, New York
16. Spivak M (1994) Calculus. Publish or Perish
17. Thomson JJ (1971) A defense of abortion. Philos Public Aff 1(1):47–66
18. Thurston W (1994) On proof and progress in mathematics. Bull Am Math Soc 30(2):161–174
19. Williams B (1985) Ethics and the limits of philosophy. Harvard University Press, Cambridge

Further Reading

Brown JR (1999/2008) Philosophy of mathematics: a contemporary introduction to the world of proofs and pictures, 2nd edn. Routledge, London/New York

Elstein DY, Hurka T (2009) From thick to thin: two moral reduction plans. Can J Philos 39(4):5-5-536

Gödel K (1951) Some basic theorems of the foundations of mathematics and their implications. Reprinted in Feferman, et al. Gödel: collected works, vol III, Oxford (1995)

Gödel K (1964) What is Cantor's continuum problem? Reprinted in Feferman et al. Gödel: collected works, vol II, Oxford (1995)

Hurka T (2014) British ethical theorists from Sidgwick to Ewing. Oxford University Press, Oxford

Jaffe A, Quinn F (1993) 'Theoretical mathematics': toward a cultural synthesis of mathematics and theoretical physics. Bull Am Math Soc 29(1):1–13

Moore GE (1903) Principia ethica. Cambridge University Press, Cambridge

Robinson JA (2000) Proof = Guarantee + Explanation. In: Intellectics and computational logic (to Wolfgang Bibel on the occasion of his 60th birthday. Kluwer, Dordrecht, pp 277–294

How to Generate All Possible Rational Wilf-Zeilberger Pairs?

Shaoshi Chen

Dedicated to the memory of Jonathan M. Borwein and Ann Johnson.

Abstract A Wilf–Zeilberger pair (F, G) in the discrete case satisfies the equation

$$F(n+1,k) - F(n,k) = G(n,k+1) - G(n,k).$$

We present a structural description of all possible rational Wilf–Zeilberger pairs and their continuous and mixed analogues.

1 Introduction

The Wilf–Zeilberger (abbr. WZ) theory [50, 61, 62] has become a bridge between symbolic computation and combinatorics. Through this bridge, not only classical combinatorial identities from handbooks and long-standing conjectures in combinatorics, such as Gessel's conjecture [10, 36] and q-TSPP conjecture [38], are proved algorithmically, but also some new identities and conjectures related to mathematical constants, such as π and zeta values, are discovered via computerized guessing [9, 18, 22, 53].

WZ-pair is one of leading concepts in the WZ theory that was originally introduced in [62] with a recent brief description in [59]. In the discrete case, a WZ-pair $(F(n,k), G(n,k))$ satisfies the WZ equation

S. Chen (✉)
KLMM, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China
e-mail: schen@amss.ac.cn

© Springer Science+Business Media, LLC, part of Springer Nature 2019

N. Fillion et al. (eds.), *Algorithms and Complexity in Mathematics, Epistemology, and Science*, Fields Institute Communications 82,
https://doi.org/10.1007/978-1-4939-9051-1_2

$$F(n+1,k)-F(n,k)=G(n,k+1)-G(n,k),$$

where both F and G are hypergeometric terms, i.e., their shift quotients with respect to n and k are rational functions in n and k, respectively. Once a WZ-pair is given, one can sum on both sides of the above equation over k from 0 to ∞ to get

$$\sum_{k=0}^{\infty}F(n+1,k)-\sum_{k=0}^{\infty}F(n,k)=\lim_{k\to\infty}G(n,k+1)-G(n,0).$$

If $G(n,0)$ and $\lim_{k\to\infty}G(n,k+1)$ are 0 then we obtain

$$\sum_{k=0}^{\infty}F(n+1,k)=\sum_{k=0}^{\infty}F(n,k),$$

which implies that $\sum_{k=0}^{\infty}F(n,k)$ is independent of n. Thus, we get the identity $\sum_{k=0}^{\infty}F(n,k)=c$, where the constant c can be determined by evaluating the sum for one value of n. We may also get a companion identity by summing the WZ-equation over n. For instance, the pair (F,G) with

$$F=\frac{\binom{n}{k}^2}{\binom{2n}{n}}\quad\text{and}\quad G=\frac{(2k-3n-3)k^2}{2(2n+1)(-n-1+k)^2}\cdot\frac{\binom{n}{k}^2}{\binom{2n}{n}}$$

leads to two identities

$$\sum_{k=0}^{\infty}\binom{n}{k}^2=\binom{2n}{n}\quad\text{and}\quad\sum_{n=0}^{\infty}\frac{(3n-2k+1)}{2(2n+1)\binom{2n}{n}}\binom{n}{k}^2=1.$$

Besides to prove combinatorial identities, WZ-pairs have many other applications. One of the applications can be traced back to Andrei Markov's 1890 method for convergence-acceleration of series for computing $\zeta(3)$, which leads to the Markov-WZ method [37, 45, 46]. WZ-pairs also play a central role in the study of finding Ramanujan-type and Zeilberger-type series for constants involving π in [21, 23, 24, 26, 27, 35, 39, 67], zeta values [42, 43] and their q-analogues [31, 32, 44]. Most recent applications are related to congruences and super congruences [28, 29, 41, 53, 55–57, 66].

For appreciation we select some remarkable (q)-series about $\pi,\zeta(3)$ together with (super)-congruences whose proofs can be obtained via WZ-pairs as follows (this list is surely not comprehensive):

1. Ramanujan's series for $1/\pi$: first recorded in Ramanujan's second notebook, proved by Bauer in [8], and by Ekhad and Zeilberger using WZ-pairs in [21]. For a nice survey on Ramanujan's series, see [7].

$$\frac{2}{\pi} = \sum_{k=0}^{\infty} \frac{4k+1}{(-64)^k} \binom{2k}{k}^3.$$

2. Guillera's series for $1/\pi^2$: found and proved by Guillera in 2002 using WZ-pairs [23]. For more results on Ramanujan-type series for $1/\pi^2$, see Zudilin's surveys [65, 67].

$$\frac{128}{\pi^2} = \sum_{k=0}^{\infty} (-1)^k \binom{2k}{k}^5 \frac{820k^2 + 180k + 13}{2^{20k}}.$$

3. Guillera's Zeilberger-type series for π^2: found and proved by Guillera using WZ-pairs in [25].

$$\frac{\pi^2}{2} = \sum_{k=1}^{\infty} \frac{(3k-1)16^k}{k^3 \binom{2k}{k}^3}.$$

4. Markov–Apéry's series for $\zeta(3)$: first discovered by Andrei Markov in 1890, used by Apéry for his irrationality proof, and proved by Zeilberger using WZ-pairs in [63].

$$\zeta(3) = \frac{5}{2} \sum_{k=1}^{\infty} \frac{(-1)^{k-1}}{k^3 \binom{2k}{k}}.$$

5. Amdeberhan's series for $\zeta(3)$: proved by Amdeberhan in 1996 using WZ-pairs [5].

$$\zeta(3) = \frac{1}{4} \sum_{k=1}^{\infty} (-1)^{k-1} \frac{56k^2 - 32k + 5}{k^3(2k-1)^2 \binom{2k}{k} \binom{3k}{k}}.$$

6. Bailey–Borwein–Bradley identity: experimentally discovered and proved by Bailey et al. in [6], a proof using the Markov-WZ method is given in [42, 43] and its q-analogue is presented in [44].

$$\sum_{k=0}^{\infty} \zeta(2k+2) z^{2k} = 3 \sum_{k=1}^{\infty} \frac{1}{\binom{2k}{k}(k^2 - z^2)} \prod_{m=1}^{k-1} \frac{m^2 - 4z^2}{m^2 - z^2}, \quad z \in \mathbb{C} \text{ with } |z| < 1.$$

7. van Hamme's supercongruence I: first conjectured by van Hamme [60], proved by Mortenson [47] using ${}_6F_5$ transformations and by Zudilin [66] using WZ-pairs.

$$\sum_{k=0}^{\frac{p-1}{2}} \frac{4k+1}{(-64)^k}\binom{2k}{k}^3 \equiv p(-1)^{\frac{p-1}{2}} (\mathrm{mod}\ p^3),$$

where p is an odd prime and the multiplicative inverse of $(-64)^k$ should be computed modulo p^3.

8. van Hamme's supercongruence II: first conjectured by van Hamme [60], proved by Long [41] using hypergeometric evaluation identities, one of which is obtained by Gessel using WZ-pairs in [22].

$$\sum_{k=0}^{\frac{p-1}{2}} \frac{6k+1}{256^k}\binom{2k}{k}^3 \equiv p(-1)^{\frac{p-1}{2}} (\mathrm{mod}\ p^4),$$

where $p > 3$ is a prime and the multiplicative inverse of $(256)^k$ should be computed modulo p^4.

9. Guo's q-analogue of van Hamme's supercongruence I: discovered and proved recently by Guo using WZ-pairs in [29].

$$\sum_{k=0}^{\frac{p-1}{2}} (-1)^k q^{k^2}[4k+1]_q \frac{(q;q^2)_k^3}{(q^2;q^2)_k^3} \equiv [p]_q q^{\frac{(p-1)^2}{4}} (-1)^{\frac{p-1}{2}} \ (\mathrm{mod}\ [p]_q^3),$$

where for $n \in \mathbb{N}$, $(a;q)_n := (1-a)(1-aq)\cdots(1-aq^{n-1})$ with $(a;q)_0 = 1$, $[n]_q = 1 + q + \cdots + q^{n-1}$ and p is an odd prime.

10. Hou–Krattenthaler–Sun's q-analogue of Guillera's Zeilberger-type series for π^2: inspired by a recent conjecture on supercongruence by Guo in [30], and proved using WZ-pairs in [35]. This work is also connected to other emerging developments on q-analogues of series for famous constants and formulae [31, 32, 58].

$$2\sum_{k=0}^{\infty} q^{2k^2+2k}(1+q^{2k^2+2}-2q^{4k+3})\frac{(q^2;q^2)_k^3}{(q;q^2)_{k+1}^3(-1;q)_{2k+3}} = \sum_{k=0}^{\infty} \frac{q^{2k}}{(1-q^{2k+1})^2}.$$

For applications, it is crucial to have WZ-pairs at hand. In the previous work, WZ-pairs are obtained either by guessing from the identities to be proved using Gosper's algorithm or by certain transformations from a given WZ-pair [22]. Riordan in the preface of his book [51] commented that "the central fact developed is that identities are both inexhaustible and unpredictable; the age-old dream of putting order in this chaos is doomed to failure ". As an optimistic respond to

Riordan's comment, Gessel in his talk[1] on the WZ method motivated with some examples commented that "WZ forms bring order to this chaos ", where WZ-forms are a multivariate generalization of WZ-pairs [63]. With the hope of discovering more combinatorial identities in an intrinsic and algorithmic way, it is natural and challenging to ask the following question.

Problem 1 How to generate all possible WZ-pairs algorithmically?

This problem seems quite open, but every promising project needs a starting point. In [40], Liu had described the structure of a special class of analytic WZ-functions with $F = G$ in terms of Rogers–Szegö polynomials and Stieltjes–Wigert polynomials in the q-shift case. In [54], Sun studied the relation between generating functions of $F(n, k)$ and $G(n, k)$ if (F, G) is a WZ-pair and applied this relation to prove some combinatorial identities. In this paper, we solve the problem completely for the first non-trivial case, namely, the case of rational WZ-pairs. To this end, let us first introduce some notations. Throughout this paper, let K be a field of characteristic zero and $K(x, y)$ be the field of rational functions in x and y over K. Let $D_x = \partial/\partial_x$ and $D_y = \partial/\partial_y$ be the usual derivations with respect to x and y, respectively. The shift operators σ_x and σ_y are defined respectively as

$$\sigma_x(f(x, y)) = f(x + 1, y) \quad \text{and} \quad \sigma_y(f(x, y)) = f(x, y + 1) \quad \text{for } f \in K(x, y).$$

For any $q \in K \setminus \{0\}$, we define the q-shift operators $\tau_{q,x}$ and $\tau_{q,y}$ respectively as

$$\tau_{q,x}(f(x, y)) = f(qx, y) \quad \text{and} \quad \tau_{q,y}(f(x, y)) = f(x, qy) \quad \text{for } f \in K(x, y).$$

For $z \in \{x, y\}$, let Δ_z and $\Delta_{q,z}$ denote the difference and q-difference operators defined by $\Delta_z(f) = \sigma_z(f) - f$ and $\Delta_{q,z}(f) = \tau_{q,z}(f) - f$ for $f \in K(x, y)$, respectively.

Definition 1 Let $\partial_x \in \{D_x, \Delta_x, \Delta_{q,x}\}$ and $\partial_y \in \{D_y, \Delta_y, \Delta_{q,y}\}$. A pair (f, g) with $f, g \in K(x, y)$ is called a *WZ-pair* with respect to (∂_x, ∂_y) in $K(x, y)$ if $\partial_x(f) = \partial_y(g)$.

The set of all rational WZ-pairs in $K(x, y)$ with respect to (∂_x, ∂_y) forms a linear space over K, denoted by $\mathscr{P}_{(\partial_x,\partial_y)}$. A WZ-pair (f, g) with respect to (∂_x, ∂_y) is said to be *exact*[2] if there exists $h \in K(x, y)$ such that $f = \partial_y(h)$ and $g = \partial_x(h)$. Let $\mathscr{E}_{(\partial_x,\partial_y)}$ denote the set of all exact WZ-pairs with respect to (∂_x, ∂_y), which forms a subspace of $\mathscr{P}_{(\partial_x,\partial_y)}$. The goal of this paper is to provide an explicit description of the structure of the quotient space $\mathscr{P}_{(\partial_x,\partial_y)}/\mathscr{E}_{(\partial_x,\partial_y)}$.

[1]The talk was given at the Waterloo Workshop in Computer Algebra (in honor of Herbert Wilf's 80th birthday), Wilfrid Laurier University, May 28, 2011. For the talk slides, see the link: http://people.brandeis.edu/~gessel/homepage/slides/wilf80-slides.pdf.

[2]This is motivated by the fact that a differential form $\omega = g dx + f dy$ *with* $f, g \in K(x, y)$ is exact in $K(x, y)$ if and only if $f = D_y(h)$ and $g = D_x(h)$ for some $h \in K(x, y)$.

The remainder of this paper is organized as follows. As our key tools, residue criteria for rational integrability and summability are recalled in Sect. 2. In Sect. 3, we present structure theorems for rational WZ-pairs in three different settings. This paper ends with a conclusion along with some remarks on the future research.

2 Residue Criteria

In this section, we recall the notion of residues and their (q-)discrete analogues for rational functions and some residue criteria for rational integrability and summability from [11, 13, 34].

Let F be a field of characteristic zero and $F(z)$ be the field of rational functions in z over F. Let D_z be the usual derivation on $F(z)$ such that $D_z(z) = 1$ and $D_z(c)$=0 for all $c \in F$. A rational function $f \in F(z)$ is said to be D_z-*integrable* in $F(z)$ if $f = D_z(g)$ for some $g \in F(z)$. By the irreducible partial fraction decomposition, one can always uniquely write $f \in F(z)$ as

$$f = q + \sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{a_{i,j}}{d_i^j}, \tag{1}$$

where $q, a_{i,j}, d_i \in F[z]$, $\deg_z(a_{i,j}) < \deg_z(d_i)$ and the d_i's are distinct irreducible and monic polynomials. We call $a_{i,1}$ the *pseudo* D_z-*residue* of f at d_i, denoted by $\operatorname{pres}_{D_z}(f, d_i)$. For an irreducible polynomial $p \in F[z]$, we let $\mathscr{O}_p$ denote the set

$$\mathscr{O}_p := \left\{ \frac{a}{b} \in F(z) \mid a, b \in F[z] \text{ with } \gcd(a, b) \text{ and } p \nmid b \right\},$$

and let $\mathscr{R}_p$ denote the set $\{f \in F(z) \mid pf \in \mathscr{O}_p\}$. If $f \in \mathscr{R}_p$, the pseudo-residue $\operatorname{pres}_{D_z}(f, p)$ is called the D_z-*residue* of f at p, denoted by $\operatorname{res}_{D_z}(f, p)$. The following example shows that pseudo-residues may not be the obstructions for D_z-integrability in $F(z)$.

Example 1 Let $F := \mathbb{Q}$ and $f = (1 - z^2)/(z^2 + 1)^2$. Then the irreducible partial fraction decomposition of f is of the form

$$f = \frac{2}{(z^2 + 1)^2} - \frac{1}{z^2 + 1}.$$

The pseudo-residue of f at $z^2 + 1$ is -1, which is nonzero. However, f is D_z-integrable in $F(z)$ since $f = D_z(z/(z^2 + 1))$.

The following lemma shows that D_z-residues are the only obstructions for D_z-integrability of rational functions with squarefree denominators, so are pseudo-residues if F is algebraically closed.

Lemma 1 ([13, Proposition 2.2]) *Let $f = a/b \in F(z)$ be such that $a, b \in F[z]$, $\gcd(a, b) = 1$. If b is squarefree, then f is D_z-integrable in $F(z)$ if and only if $\operatorname{res}_{D_z}(f, d) = 0$ for any irreducible factor d of b. If F is algebraically closed, then f is D_z-integrable in $F(z)$ if and only if $\operatorname{pres}_{D_z}(f, z - \alpha) = 0$ for any root α of the denominator b.*

By the Ostrogradsky–Hermite reduction [11, 33, 49], we can decompose a rational function $f \in F(z)$ as $f = D_z(g) + a/b$, where $g \in F(z)$ and $a, b \in F[z]$ are such that $\deg_z(a) < \deg_z(b)$, $\gcd(a, b) = 1$, and b is a squarefree polynomial in $F[z]$. By Lemma 1, f is D_z-integrable in $F(z)$ if and only if $a = 0$.

We now recall the (q-)discrete analogue of D_z-residues introduced in [13, 34]. Let ϕ be an automorphism of $F(z)$ that fixes F. For a polynomial $p \in F[z]$, we call the set $\{\phi^i(p) \mid i \in \mathbb{Z}\}$ the *ϕ-orbit* of p, denoted by $[p]_\phi$. Two polynomials $p, q \in F[z]$ are said to be ϕ-equivalent (denoted as $p \sim_\phi q$) if they are in the same ϕ-orbit, i.e., $p = \phi^i(q)$ for some $i \in \mathbb{Z}$. For any $a, b \in F(z)$ and $m \in \mathbb{Z}$, we have

$$\frac{a}{\phi^m(b)} = \phi(g) - g + \frac{\phi^{-m}(a)}{b}, \tag{2}$$

where g is equal to $\sum_{i=0}^{m-1} \frac{\phi^{i-m}(a)}{\phi^i(b)}$ if $m \geq 0$, and equal to $-\sum_{i=0}^{-m-1} \frac{\phi^i(a)}{\phi^{m+i}(b)}$ if $m < 0$.

Let σ_z be the shift operator with respect to z defined by $\sigma_z(f(z)) = f(z + 1)$. Note that σ_z is an automorphism of $F(z)$ that fixes F. A rational function $f \in F(z)$ is said to be *σ_z-summable* in $F(z)$ if $f = \sigma_z(g) - g$ for some $g \in F(z)$. For any $f \in F(z)$, we can uniquely decompose it into the form

$$f = p(z) + \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{\ell=0}^{e_{i,j}} \frac{a_{i,j,\ell}}{\sigma_z^\ell(d_i)^j}, \tag{3}$$

where $p, a_{i,j,\ell}, d_i \in F[z]$, $\deg_z(a_{i,j,\ell}) < \deg_z(d_i)$ and the d_i's are irreducible and monic polynomials such that no two of them are σ_z-equivalent. We call the sum $\sum_{\ell=0}^{e_{i,j}} \sigma_z^{-\ell}(a_{i,j,\ell})$ the *σ_z-residue* of f at d_i of multiplicity j, denoted by $\operatorname{res}_{\sigma_z}(f, d_i, j)$. Recently, the notion of σ_z-residues has been generalized to the case of rational functions over elliptic curves [20, Appendix B]. The following lemma is a discrete analogue of Lemma 1 which shows that σ_z-residues are the only obstructions for σ_z-summability in the field $F(z)$.

Lemma 2 ([13, Proposition 2.5]) *Let $f = a/b \in F(z)$ be such that $a, b \in F[z]$ and $\gcd(a, b) = 1$. Then f is σ_z-summable in $F(z)$ if and only if $\operatorname{res}_{\sigma_z}(f, d, j) = 0$ for any irreducible factor d of the denominator b of any multiplicity $j \in \mathbb{N}$.*

By Abramov's reduction [1, 2], we can decompose a rational function $f \in F(z)$ as

$$f = \Delta_z(g) + \sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{a_{i,j}}{b_i^j},$$

where $g \in F(z)$ and $a_{i,j}, b_i \in F[z]$ are such that $\deg_z(a_{i,j}) < \deg_z(b_i)$ and the b_i's are irreducible and monic polynomials in distinct σ_z-orbits. By Lemma 2, h is σ_z-summable in $F(z)$ if and only if $a_{i,j} = 0$ for all i, j with $1 \le i \le n$ and $1 \le j \le m_i$.

Let q be a nonzero element of F such that $q^m \ne 1$ for all nonzero $m \in \mathbb{Z}$ and let $\tau_{q,z}$ be the q-shift operator with respect to z defined by $\tau_{q,z}(f(z)) = f(qz)$. Since q is nonzero, $\tau_{q,z}$ is an automorphism of $F(z)$ that fixes F. A rational function $f \in F(z)$ is said to be $\tau_{q,z}$-*summable* in $F(z)$ if $f = \tau_{q,z}(g) - g$ for some $g \in F(z)$. For any $f \in F(z)$, we can uniquely decompose it into the form

$$f = c + zp_1 + \frac{p_2}{z^s} + \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{\ell=0}^{e_{i,j}} \frac{a_{i,j,\ell}}{\tau_{q,z}^{\ell}(d_i)^j}, \tag{4}$$

where $c \in F, s, n, m_i, e_{i,j} \in \mathbb{N}$ with $s \ne 0$, and $p_1, p_2, a_{i,j,\ell}, d_i \in F[z]$ are such that $\deg_z(p_2) < s$, $\deg_z(a_{i,j,\ell}) < \deg_z(d_i)$, and p_2 is either zero or has nonzero constant term, i.e., $p_2(0) \ne 0$. Moreover, the d_i's are irreducible and monic polynomials in distinct $\tau_{q,z}$-orbits and $z \nmid d_i$ for all i with $1 \le i \le n$. We call the constant c the $\tau_{q,z}$-*residue* of f at infinity, denoted by $\operatorname{res}_{\tau_{q,z}}(f, \infty)$ and call the sum $\sum_{\ell=0}^{e_{i,j}} \tau_{q,z}^{-\ell}(a_{i,j,\ell})$ the $\tau_{q,z}$-*residue* of f at d_i of multiplicity j, denoted by $\operatorname{res}_{\tau_{q,z}}(f, d_i, j)$. A q-analogue of Lemma 2 is as follows.

Lemma 3 ([13, Proposition 2.10]) *Let $f = a/b \in F(z)$ be such that $a, b \in F[z]$ and $\gcd(a, b) = 1$. Then f is $\tau_{q,z}$-summable in $F(z)$ if and only if $\operatorname{res}_{\tau_{q,z}}(f, \infty) = 0$ and $\operatorname{res}_{\tau_{q,z}}(f, d, j) = 0$ for any irreducible factor d of the denominator b of any multiplicity $j \in \mathbb{N}$.*

By a q-analogue of Abramov's reduction [2], we can decompose a rational function $f \in F(z)$ as

$$f = \Delta_{q,z}(g) + c + \sum_{i=1}^{n} \sum_{j=1}^{m_i} \frac{a_{i,j}}{b_i^j},$$

where $g \in F(z), c \in F$, and $a_{i,j}, b_i \in F[z]$ are such that $\deg_z(a_{i,j}) < \deg_z(b_i)$ and the b_i's are irreducible and monic polynomials in distinct σ_z-orbits and $\gcd(z, b_i) = 1$ for all i with $1 \le i \le n$. By Lemma 3, f is $\tau_{q,z}$-summable in $F(z)$ if and only if $c = 0$ and $a_{i,j} = 0$ for all i, j with $1 \le i \le n$ and $1 \le j \le m_i$.

Remark 1 Note that pseudo-residues are essentially different from residues in the differential case, but not needed in the shift and q-shift cases.

3 Structure Theorems

In this section, we present structure theorems for rational WZ-pairs in terms of some special pairs. Throughout this section, we will assume that K is an algebraically closed field of characteristic zero and let $\partial_x \in \{D_x, \Delta_x, \Delta_{q,x}\}$ and $\partial_y \in \{D_y, \Delta_y, \Delta_{q,y}\}$.

We first consider the special case that $q \in K$ is a root of unity. Assume that m is the minimal positive integer such that $q^m = 1$. For any $f \in K(x, y)$, it is easy to show that $\tau_{q,y}(f) = f$ if and only if $f \in K(x)(y^m)$. Note that $K(x, y)$ is a finite algebraic extension of $K(x)(y^m)$ of degree m. In the following theorem, we show that WZ-pairs in this special case are of a very simple form.

Theorem 1 *Let $\partial_x \in \{D_x, \Delta_x, \Delta_{q,x}\}$ and $f, g \in K(x, y)$ be such that $\partial_x(f) = \Delta_{q,y}(g)$. Then there exist rational functions $h \in K(x, y)$ and $a, b \in K(x, y^m)$ such that $\partial_x(a) = 0$ and*

$$f = \Delta_{q,y}(h) + a \quad \text{and} \quad g = \partial_x(h) + b.$$

Moreover, we have $a \in K(y^m)$ if $\partial_x \in \{D_x, \Delta_x\}$ and $a \in K(x^m, y^m)$ if $\partial_x = \Delta_{q,x}$.

Proof By Lemma 2.4 in [14], any rational function $f \in K(x, y)$ can be decomposed as

$$f = \Delta_{q,y}(h) + a, \quad \text{where } h \in K(x, y) \text{ and } a \in K(x)(y^m). \tag{5}$$

Moreover, f is $\tau_{q,y}$-summable in $K(x, y)$ if and only if $a = 0$. Then

$$\partial_x(f) = \Delta_{q,y}(\partial_x(h)) + \partial_x(a).$$

Note that $\partial_x(a) \in K(x)(y^m)$, which implies that $\partial_x(a) = 0$ because $\partial_x(f)$ is $\tau_{q,y}$-summable in $K(x, y)$. Then $\Delta_{q,y}(g) = \Delta_{q,y}(\partial_x(h))$. So $g = \partial_x(h) + b$ for some $b \in K(x, y^m)$. This completes the proof. □

From now on, we assume that q is not a root of unity. We will investigate WZ-pairs in three different cases according to the choice of the pair (∂_x, ∂_y).

3.1 The Differential Case

In the continuous setting, we consider WZ-pairs with respect to (D_x, D_y), i.e., the pairs of the form (f, g) with $f, g \in K(x, y)$ satisfying $D_x(f) = D_y(g)$.

Definition 2 A WZ-pair (f, g) with respect to (D_x, D_y) is called a *log-derivative* pair if there exists nonzero $h \in K(x, y)$ such that $f = D_y(h)/h$ and $g = D_x(h)/h$.

The following theorem shows that any WZ-pair in the continuous case is a linear combination of exact and log-derivative pairs, which was first proved by Christopher in [19] and then extended to the multivariate case in [12, 64].

Theorem 2 *Let $f, g \in K(x, y)$ be such that $D_x(f) = D_y(g)$. Then there exist rational functions $a, b_1, \ldots, b_n \in K(x, y)$ and nonzero constants $c_1, \ldots, c_n \in K$ such that*

$$f = D_y(a) + \sum_{i=1}^{n} c_i \frac{D_y(b_i)}{b_i} \quad \text{and} \quad g = D_x(a) + \sum_{i=1}^{n} c_i \frac{D_x(b_i)}{b_i}.$$

Proof The proof in the case when K is the field of complex numbers can be found in [19, Theorem 2] and in the case when K is any algebraically closed field of characteristic zero can be found in [12, Theorem 4.4.3].

Corollary 1 *The quotient space $\mathscr{P}_{(D_x,D_y)}/\mathscr{E}_{(D_x,D_y)}$ is spanned over K by the set*

$$\{(f, g) + \mathscr{E}_{(D_x,D_y)} \mid f, g \in K(x, y) \text{ such that } (f, g) \text{ is a log-derivative pair}\}.$$

Remark 2 A differentiable function $h(x, y)$ is said to be hyperexponential over $\mathbb{C}(x, y)$ if $D_x(h) = fh$ and $D_y(h) = gh$ for some $f, g \in \mathbb{C}(x, y)$. The above theorem enables us to obtain the multiplicative structure of hyperexponential functions, i.e., any hyperexponential function $h(x, y)$ can be written as $h = \exp(a) \cdot \prod_{i=1}^{n} b_i^{c_i}$ for some $a, b_i \in \mathbb{C}(x, y)$ and $c_i \in \mathbb{C}$.

3.2 The (q)-Shift Case

In the discrete setting, we consider WZ-pairs with respect to (∂_x, ∂_y) with $\partial_x \in \{\Delta_x, \Delta_{q,x}\}$ and $\partial_y \in \{\Delta_y, \Delta_{q,y}\}$, i.e., the pairs of the form (f, g) with $f, g \in K(x, y)$ satisfying $\partial_x(f) = \partial_y(g)$.

Let $\theta_x \in \{\sigma_x, \tau_{q,x}\}$ and $\theta_y \in \{\sigma_y, \tau_{q,y}\}$. For any nonzero $m \in \mathbb{Z}$, θ_x^m is also an automorphism on $K(x, y)$ that fixes $K(y)$, i.e., for any $f \in K(x, y)$, $\theta_x^m(f) = f$ if and only if $f \in K(y)$. The ring of polynomials in θ_x and θ_y over K is denoted by $K[\theta_x, \theta_y]$. For any $p = \sum_{i,j} c_{i,j}\theta_x^i\theta_y^j \in K[\theta_x, \theta_y]$ and $f \in K(x, y)$, we define the action $p \bullet f = \sum_{i,j} c_{i,j}\theta_x^i(\theta_y^j(f))$. Then $K(x, y)$ can be viewed as a $K[\theta_x, \theta_y]$-module. Let $G = \langle\theta_x, \theta_y\rangle$ be the free abelian group generated by θ_x and θ_y. Let $f \in K(x, y)$ and H be a subgroup of G. We call the set $\{c\theta(f) \mid c \in K \setminus \{0\}, \theta \in H\}$ the *H-orbit* at f, denoted by $[f]_H$. Two elements $f, g \in K(x, y)$ are said to be H-equivalent if $[f]_H = [g]_H$, denoted by $f \sim_H g$. The relation $\sim_H$ is an equivalence relation. A rational function $f \in K(x, y)$ is said to be (θ_x, θ_y)-*invariant* if there exist $m, n \in \mathbb{Z}$, not all zero, such that $\theta_x^m\theta_y^n(f) = f$. All possible (θ_x, θ_y)-invariant rational functions have been completely characterized in [4, 16, 17, 48, 52]. We summarize the characterization as follows.

Proposition 1 *Let $f \in K(x, y)$ be (θ_x, θ_y)-invariant, i.e., there exist $m, n \in \mathbb{Z}$, not all zero, such that $\theta_x^m \theta_y^n(f) = f$. Set $\bar{n} = n/\gcd(m, n)$ and $\bar{m} = m/\gcd(m, n)$. Then*

1. *if $\theta_x = \sigma_x$ and $\theta_y = \sigma_y$, then $f = g(\bar{n}x - \bar{m}y)$ for some $g \in K(z)$;*
2. *if $\theta_x = \tau_{q,x}$, $\theta_y = \tau_{q,y}$, then $f = g(x^{\bar{n}} y^{-\bar{m}})$ for some $g \in K(z)$;*
3. *if $\theta_x = \sigma_x$, $\theta_y = \tau_{q,y}$, then $f \in K(x)$ if $m = 0$, $f \in K(y)$ if $n = 0$, and $f \in K$ if $mn \neq 0$.*

We introduce a discrete analogue of the log-derivative pairs.

Definition 3 A WZ-pair (f, g) with respect to (∂_x, ∂_y) is called a *cyclic* pair if there exists a (θ_x, θ_y)-invariant $h \in K(x, y)$ such that

$$f = \frac{\theta_x^s - 1}{\theta_x - 1} \bullet h \quad \text{and} \quad g = \frac{\theta_y^t - 1}{\theta_y - 1} \bullet h,$$

where $s, t \in \mathbb{Z}$ are not all zero satisfying that $\theta_x^s(h) = \theta_y^t(h)$.

In the above definition, we may always assume that $s \geq 0$. Note that for any $n \in \mathbb{Z}$ we have

$$\frac{\theta_y^n - 1}{\theta_y - 1} = \begin{cases} \sum_{j=0}^{n-1} \theta_y^j, & n \geq 0; \\ -\sum_{j=1}^{-n} \theta_y^{-j}, & n < 0. \end{cases}$$

Example 2 Let $a \in K(y)$ and $b \in K(x)$. Then both $(a, 0)$ and $(0, b)$ are cyclic by taking $h = a, s = 1, t = 0$ and $h = b, s = 0, t = 1$, respectively. Let $p = 2x + 3y$. Then the pair (f, g) with

$$f = \frac{1}{p} + \frac{1}{\sigma_x(p)} + \frac{1}{\sigma_x^2(p)} \quad \text{and} \quad g = \frac{1}{p} + \frac{1}{\sigma_y(p)}$$

is a cyclic WZ-pair with respect to (Δ_x, Δ_y).

Let $V_0 = K(x)[y]$ and V_m be the set of all rational functions of the form $\sum_{i=1}^{I} a_i / b_i^m$, where $m \in \mathbb{Z}_+, a_i, b_i, \in K(x)[y]$, $\deg_y(a_i) < \deg_y(b_i)$ and the b_i's are distinct irreducible polynomials in the ring $K(x)[y]$. By definition, the set V_m forms a subspace of $K(x, y)$ as a vector spaces over $K(x)$. By the irreducible partial fraction decomposition, any $f \in K(x, y)$ can be uniquely decomposed into $f = f_0 + f_1 + \cdots + f_n$ with $f_i \in V_i$ and so $K(x, y) = \bigoplus_{i=0}^{\infty} V_i$. The following lemma shows that the space V_m is invariant under certain shift operators.

Lemma 4 *Let $f \in V_m$ and $P \in K(x)[\theta_x, \theta_y]$. Then $P(f) \in V_m$.*

Proof Let $f = \sum_{i=1}^{I} a_i / b_i^m$ and $P = \sum_{i,j} p_{i,j} \theta_x^i \theta_y^j$. For any $\theta = \theta_x^i \theta_y^j$ with $i, j, k \in \mathbb{Z}$, $\theta(b_i)$ is still irreducible and $\deg_y(\theta(a_i)) < \deg_y(\theta(b_i))$. Then all of the simple fractions $p_{i,j} \theta_x^i \theta_y^j(a_i) / \theta_x^i \theta_y^j(b_i)^n$ appearing in $P(f)$ are proper

in y and have irreducible denominators. If some of denominators are the same, we can simplify them by adding the numerators to get a simple fraction. After this simplification, we see that $P(f)$ can be written in the same form as f, so it is in V_m. □

Lemma 5 *Let p be a monic polynomial in $K(x)[y]$. If $\theta_x^m(p) = c\theta_y^n(p)$ for some $c \in K(x)$ and $m, n \in \mathbb{Z}$ with m, n being not both zero, then $c \in K$.*

Proof Write $p = \sum_{i=0}^{d} p_i y^i$ with $p_i \in K(x)$ and $p_d = 1$. Then

$$\theta_x^m(p) = \sum_{i=0}^{d} \theta_x^m(p_i) y^i = c \sum_{i=0}^{d} p_i \theta_y^n(y^i) = c\theta_y^n(p).$$

Comparing the leading coefficients in y yields $c = 1$ if $\theta_y = \sigma_y$ and $c = q^{-nd}$ if $\theta_y = \tau_{q,y}$. Thus, $c \in K$ because $q \in K$. □

Lemma 6 *Let $f \in K(x, y)$ be a rational function of the form*

$$f = \frac{a_0}{b^m} + \frac{a_1}{\theta_x(b^m)} + \cdots + \frac{a_n}{\theta_x^n(b^m)},$$

where $m \in \mathbb{Z}_+, n \in \mathbb{N}, a_0, a_1, \ldots, a_n \in K(x)[y]$ with $a_n \neq 0$ and $b \in K(x)[y]$ are such that $\deg_y(a_i) < \deg_y(b)$ and b is an irreducible and monic polynomial in $K(x)[y]$ such that $\theta_x^i(b)$ and $\theta_x^j(b)$ are not θ_y-equivalent for all $i, j \in \{0, 1, \ldots, n\}$ with $i \neq j$. If $\theta_x(f) - f = \theta_y(g) - g$ for some $g \in K(x, y)$, then (f, g) is cyclic.

Proof By a direct calculation, we have

$$\theta_x(f) - f = \frac{\theta_x(a_n)}{\theta_x^{n+1}(b^m)} - \frac{a_0}{b^m} + \frac{\theta_x(a_0) - a_1}{\theta_x(b^m)} + \cdots + \frac{\theta_x(a_{n-1}) - a_n}{\theta_x^n(b^m)}.$$

If $\theta_x(f) - f = \theta_y(g) - g$ for some $g \in K(x, y)$, then all of the θ_y-residues at distinct θ_y-orbits of $\theta_x(f) - f$ are zero by residue criteria in Sect. 2. Since $b^m, \theta_x(b^m), \ldots, \theta_x^n(b^m)$ are in distinct θ_y-orbits, $\theta_x^{n+1}(b^m)$ must be θ_y-equivalent to one of them. Otherwise, we get

$$a_0 = 0, \quad \theta_x(a_0) - a_1 = 0, \quad \ldots, \quad \theta_x(a_{n-1}) - a_n = 0, \quad \text{and} \quad \theta_x(a_n) = 0.$$

Since θ_x is an automorphism on $K(x, y)$, we have $a_0 = a_1 = \cdots = a_n = 0$, which contradicts the assumption that $a_n \neq 0$. If $\theta_x^{n+1}(b^m)$ is θ_y-equivalent to $\theta_x^i(b^m)$ for some $0 < i \leq n$, so is $\theta_x^{n+1-i}(b^m)$, which contradicts the assumption. Thus, $\theta_x^{n+1}(b^m) = c\theta_y^t(b^m)$ for some $c \in K(x) \setminus \{0\}$ and $t \in \mathbb{Z}$. By Lemma 5, we have $c \in K \setminus \{0\}$. A direct calculation leads to

$$\begin{aligned}\theta_x(f)-f&=\frac{\theta_x(a_n)}{\theta_x^{n+1}(b^m)}-\frac{a_0}{b^m}\\&\quad+\sum_{i=1}^{n}\frac{\theta_x(a_{i-1})-a_i}{\theta_x^i(b^m)}=\frac{\theta_x(a_n)}{c\theta_y^t(b^m)}-\frac{a_0}{b^m}+\sum_{i=1}^{n}\frac{\theta_x(a_{i-1})-a_i}{\theta_x^i(b^m)}\\&=\frac{\theta_y^{-t}\theta_x(a_n/c)-a_0}{b^m}+\sum_{i=1}^{n}\frac{\theta_x(a_{i-1})-a_i}{\theta_x^i(b^m)}+\theta_y(u)-u\end{aligned}$$

for some $u \in K(x,y)$ using the formula (2). By the residue criteria, we then get $a_0 = \theta_y^{-t}\theta_x(a_n/c), a_1 = \theta_x(a_0), \ldots,$ and $a_n = \theta_x(a_{n-1})$. This implies that $\theta_x^{n+1}(a_0) = c\theta_y^t(a_0)$ and $a_i = \theta_x^i(a_0)$ for $i \in \{1,\ldots,n\}$. So $f = \frac{\theta_x^{n+1}-1}{\theta_x-1} \bullet h$ with $h = a_0/b^m$, which leads to

$$\theta_x(f)-f=\theta_x^{n+1}(h)-h=\theta_y^t(h)-h=\theta_y(g)-g \quad \text{with} \quad g=\frac{\theta_y^t-1}{\theta_y-1}\bullet h.$$

Thus, (f, g) is a cyclic WZ-pair. □

The following theorem is a discrete analogue of Theorem 2.

Theorem 3 *Let $f, g \in K(x,y)$ be such that $\partial_x(f) = \partial_y(g)$. Then there exist rational functions $a, b_1, \ldots, b_n \in K(x,y)$ such that*

$$f=\partial_y(a)+\sum_{i=1}^{n}\frac{\theta_x^{s_i}-1}{\theta_x-1}\bullet b_i \quad \textit{and} \quad g=\partial_x(a)+\sum_{i=1}^{n}\frac{\theta_y^{t_i}-1}{\theta_y-1}\bullet b_i,$$

where for each $i \in \{1,\ldots,n\}$ we have $\theta_x^{s_i}(b_i) = \theta_y^{t_i}(b_i)$ for some $s_i \in \mathbb{N}$ and $t_i \in \mathbb{Z}$ with s_i, t_i not all zero.

Proof By Abramov's reduction and its q-analogue, we can decompose f as

$$f=\partial_y(a)+c+\sum_{j=1}^{J}f_j \quad \text{with } f_j=\sum_{i=1}^{I}\sum_{\ell=0}^{L_{i,j}}\frac{a_{i,j,\ell}}{\theta_x^\ell(b_i^j)},$$

where $a \in K(x,y), c \in K(x)$, and $a_{i,j,\ell}b_i \in K(x)[y]$ such that $c = 0$ if $\theta_y = \sigma_y$, $\deg_y(a_{i,j,\ell}) < \deg_y(b_i)$, and the b_i's are irreducible and monic polynomials belonging to distinct G-orbits where $G = \langle\theta_x, \theta_y\rangle$. Moreover, $\theta_x^{\ell_1}(b_i^j)$ and $\theta_x^{\ell_2}(b_i^j)$ are in distinct θ_y-orbits if $\ell_1 \neq \ell_2$. By applying Lemma 4 to the equation $\theta_x(f) - f = \theta_y(g) - g$, we get that $\theta_x(c) - c$ is θ_y-summable and so is $\theta_x(f_j) - f_j$ for each multiplicity $j \in \{1,\ldots,J\}$. By residue criteria for θ_y-sumability and the assumption that the b_i's are in distinct $\langle\theta_x,\theta_y\rangle$-orbits, we have $\theta_x(c) - c = 0$ and for each $i \in \{1,\ldots,I\}$, the rational function $f_{i,j} := \sum_{\ell=0}^{L_{i,j}} a_{i,j,\ell}/\theta_x^\ell(b_i^j)$ is either equal

to zero $G_{b_i} := \{\theta \in \langle\theta_x\rangle \mid \theta(b_i) \sim_{\theta_y} b_i\} = \{id\}$ or there exists $g_{i,j} \in K(x, y)$ such that $\theta_x(f_{i,j}) - f_{i,j} = \theta_y(g_{i,j}) - g_{i,j}$ if $G_{b_i} = \langle\theta_x^{(L_{i,j}+1)}\rangle$ with $L_{i,j} \in \mathbb{N}$. Then $(f_{i,j}, g_{i,j})$ is cyclic by Lemma 6 for every i, j with $1 \le i \le I$ and $1 \le j \le J$. So the pair (f, g) can be written as

$$(f, g) = (\partial_y(a), \partial_x(a)) + (c, 0) + \sum_{i=1}^{I}\sum_{j=1}^{J}(f_{i,j}, g_{i,j}).$$

This completes the proof. □

Corollary 2 *The quotient space $\mathscr{P}_{(\partial_x,\partial_y)}/\mathscr{E}_{(\partial_x,\partial_y)}$ is spanned over K by the set*

$$\{(f, g) + \mathscr{E}_{(\partial_x,\partial_y)} \mid f, g \in K(x, y) \text{ such that } (f, g) \text{ is a cyclic pair}\}.$$

3.3 The Mixed Case

In the mixed continuous-discrete setting, we consider the rational WZ-pairs with respect to $(\theta_x - 1, D_y)$ with $\theta_x \in \{\sigma_x, \tau_{q,x}\}$.

Lemma 7 *Let p be an irreducible and monic polynomial in $K(x)[y]$. Then for any nonzero $m \in \mathbb{Z}$, we have either* $\gcd(p, \theta_x^m(p)) = 1$ *or $p \in K[y]$.*

Proof Since θ_x is an automorphism on $K(x, y)$, $\theta_x^i(p)$ is irreducible in $K(x)[y]$ for any $i \in \mathbb{Z}$. If $\gcd(p, \theta_x^m(p)) \ne 1$, then $\theta_x^m(p) = cp$ for some $c \in K(x)$. Write $p = \sum_{i=0}^{d} p_i y^i$ with $p_i \in K(x)$ and $p_d = 1$. Then $\theta_x^m(p) = cp$ implies that $\theta_x^m(p_i) = cp_i$ for all i with $0 \le i \le d$. Then $c = 1$ and $p_i \in K$ for all i with $0 \le i \le d - 1$. So $p \in K[y]$. □

The structure of WZ-pairs in the mixed setting is as follows.

Theorem 4 *Let $f, g \in K(x, y)$ be such that $\theta_x(f) - f = D_y(g)$. Then there exist $h \in K(x, y)$, $u \in K(y)$ and $v \in K(x)$ such that*

$$f = D_y(h) + u \quad \text{and} \quad g = \theta_x(h) - h + v.$$

Proof By the Ostrogradsky–Hermite reduction, we decompose f into the form

$$f = D_y(h) + \sum_{i=1}^{I}\sum_{j=0}^{J_i}\frac{a_{i,j}}{\theta_x^j(b_i)},$$

where $h \in K(x, y)$ and $a_{i,j}, b_i \in K(x)[y]$ with $a_{i,J_i} \ne 0$, $\deg_y(a_{i,j}) < \deg_y(b_i)$ and b_i being irreducible and monic polynomials in y over $K(x)$ such that the b_i's are in distinct θ_x-orbits. By a direct calculation, we get

$$\theta_x(f) - f = D_y(\theta_x(h) - h) + \sum_{i=1}^{I} \left(\frac{\theta_x(a_{i,J_i})}{\theta_x^{J_i+1}(b_i)} - \frac{a_{i,0}}{b_i} + \sum_{j=1}^{J_i} \frac{\theta_x(a_{i,j-1}) - a_{i,j}}{\theta_x^{j}(b_i)} \right).$$

For all i, j with $1 \le i \le I$ and $0 \le j \le J_i + 1$, the $\theta_x^j(b_i)$'s are irreducible and monic polynomials in y over $K(x)$. We first show that for each $i \in \{1, \ldots, I\}$, we have $b_i \in K[y]$. Suppose that there exists $i_0 \in \{1, \ldots, I\}$, $b_{i_0} \notin K[y]$. Then $\gcd(\theta_x^m(b_{i_0}), b_{i_0}) = 1$ for any nonzero $m \in \mathbb{Z}$ by Lemma 7. Since $\theta_x(f) - f$ is D_y-integrable in $K(x, y)$, we have $\theta_x(a_{i_0,J_{i_0}}) = 0$ by Lemma 1. Then $a_{i_0,J_{i_0}} = 0$, which contradicts the assumption that $a_{i,J_i} \neq 0$ for all i with $1 \le i \le I$. Since $b_i \in K[y]$, f can be written as

$$f = D_y(h) + \sum_{i=1}^{I} \frac{a_i}{b_i}, \quad \text{where } a_i := \sum_{j=0}^{J_i} a_{i,j}.$$

Since $\theta_x(f) - f$ is D_y-integrable in $K(x, y)$ and since

$$\theta_x(f) - f = D_y(\theta_x(h) - h) + \sum_{i=1}^{I} \frac{\theta_x(a_i) - a_i}{b_i},$$

we have $\theta_x(a_i) - a_i = 0$ for each $i \in \{1, \ldots, I\}$ by Lemma 1. This implies that $a_i \in K(y)$ and $f = D_y(h) + u$ with $u = \sum_{i=1}^{I} a_i/b_i \in K(y)$. Since $\theta_x(f) - f = D_y(g)$, we get $D_y(g - (\theta_x(h) - h)) = 0$. Then $g = \theta_x(h) - h + v$ for some $v \in K(x)$. □

Corollary 3 *The quotient space $\mathscr{P}_{(\theta_x-1,D_y)}/\mathscr{E}_{(\theta_x-1,D_y)}$ is spanned over K by the set*

$$\{(f, g) + \mathscr{E}_{(\theta_x-1,D_y)} \mid f \in K(y) \text{ and } g \in K(x)\}.$$

4 Conclusion

We have explicitly described the structure of rational WZ-pairs in terms of special pairs. With structure theorems, we can easily generate rational WZ-pairs, which solves Problem 1 in the rational case completely. For the future research, the next direction is to solve the problem in the cases of more general functions. Using the terminology of Gessel in [22], a hypergeometric term $F(x, y)$ is said to be a *WZ-function* if there exists another hypergeometric term $G(x, y)$ such that (F, G) is a WZ-pair. In the scheme of creative telescoping, (F, G) being a WZ-pair with respect to (∂_x, ∂_y) is equivalent to that ∂_x being a telescoper for F with certificate G. Complete criteria for the existence of telescopers for hypergeometric terms and their variants are known [3, 15, 17]. With the help of existence criteria for telescopers, one

can show that $F(x, y)$ can be decomposed as the sum $F = \partial_y(H_1) + H_2$ with H_1, H_2 being hypergeometric terms and H_2 is of proper form (see definition in [22, 62]) if F is a WZ-function. So it is promising to apply the ideas in the study of the existence problem of telescopers to explore the structure of WZ-pairs.

Acknowledgements I would like to thank Prof. Victor J.W. Guo and Prof. Zhi-Wei Sun for many discussions on series for special constants, (super)-congruences and their q-analogues that can be proved using the WZ method. I am also very grateful to Ruyong Feng and Rong-Hua Wang for many constructive comments on the earlier version of this paper. I also thank the anonymous reviewers for their constructive and detailed comments.
This work was supported by the NSFC grants 11501552, 11688101 and by the Frontier Key Project (QYZDJ-SSW-SYS022) and the Fund of the Youth Innovation Promotion Association, CAS.

References

1. Abramov SA (1975) The rational component of the solution of a first order linear recurrence relation with rational right hand side. Ž Vyčisl Mat i Mat Fiz 15(4):1035–1039, 1090
2. Abramov SA (1995) Indefinite sums of rational functions. In: ISSAC '95: proceedings of the 1995 international symposium on symbolic and algebraic computation. ACM, New York, NY, pp 303–308
3. Abramov SA (2003) When does Zeilberger's algorithm succeed? Adv Appl Math 30(3):424–441
4. Abramov SA, Petkovšek M (2002) On the structure of multivariate hypergeometric terms. Adv Appl Math 29(3):386–411
5. Amdeberhan T (1996) Faster and faster convergent series for $\zeta(3)$. Electron J Combin 3(1):Research Paper 13, approx. 2
6. Bailey DH, Borwein JM, Bradley DM (2006) Experimental determination of Apéry-like identities for $\zeta(2n+2)$. Exp Math 15(3):281–289
7. Baruah ND, Berndt BC, Chan HH (2009) Ramanujan's series for $1/\pi$: a survey. Am Math Mon 116(7):567–587
8. Bauer GC (1859) Von den Coefficienten der Reihen von Kugelfunctionen einer Variablen. J Reine Angew Math 56:101–121
9. Borwein J, Bailey D, Girgensohn R (2004) Experimentation in mathematics: computational paths to discovery. A K Peters/CRC Press, Natick, MA/Boca Raton, FL
10. Bostan A, Kauers M (2010) The complete generating function for Gessel walks is algebraic. Proc Am Math Soc 138(9):3063–3078
11. Bronstein M (2005) Symbolic integration I: transcendental functions, 2nd edn. Springer, Berlin
12. Chen S (2011) Some applications of differential-difference algebra to creative telescoping. PhD Thesis, Ecole Polytechnique LIX
13. Chen S, Singer MF (2012) Residues and telescopers for bivariate rational functions. Adv Appl Math 49(2):111–133
14. Chen S, Singer MF (2014) On the summability of bivariate rational functions. J Algebra 409:320–343
15. Chen WYC, Hou Q-H, Mu Y-P (2005) Applicability of the q-analogue of Zeilberger's algorithm. J Symb Comput 39(2):155–170
16. Chen S, Feng R, Fu G, Kang J (2012) Multiplicative decompositions of multivariate q-hypergeometric terms. J Syst Sci Math Sci 32(8):1019–1032
17. Chen S, Chyzak F, Feng R, Fu G, Li Z (2015) On the existence of telescopers for mixed hypergeometric terms. J Symb Comput 68(part 1):1–26

18. Chen WYC, Hou Q-H, Zeilberger D (2016) Automated discovery and proof of congruence theorems for partial sums of combinatorial sequences. J Differ Equ Appl 22(6):780–788
19. Christopher C (1999) Liouvillian first integrals of second order polynomial differential equations. Electron J Differ Equ 49:1–7
20. Dreyfus T, Hardouin C, Roques J, Singer MF (2018) On the nature of the generating series of walks in the quarter plane. Invent Math 213(1):139–203. https://doi.org/10.1007/s00222-018-0787-z
21. Ekhad SB, Zeilberger D (1994) A WZ proof of Ramanujan's formula for π. In: Geometry, Analysis and Mechanics. World Scientific Publishing, River Edge, NJ, pp 107–108
22. Gessel IM (1995) Finding identities with the WZ method. J. Symb. Comput. 20(5–6):537–566
23. Guillera J (2002) Some binomial series obtained by the WZ-method. Adv Appl Math 29(4):599–603
24. Guillera J (2006) Generators of some Ramanujan formulas. Ramanujan J 11(1):41–48
25. Guillera J (2008) Hypergeometric identities for 10 extended Ramanujan-type series. Ramanujan J 15(2):219–234
26. Guillera J (2010) On WZ-pairs which prove Ramanujan series. Ramanujan J 22(3):249–259
27. Guillera J (2013) WZ-proofs of "divergent" Ramanujan-type series. In: Advances in combinatorics. Springer, Heidelberg, pp 187–195
28. Guo VJW (2017) Some generalizations of a supercongruence of van Hamme. Integr Transforms Spec Funct (1):1–12
29. Guo VJW (2018) A q-analogue of a Ramanujan-type supercongruence involving central binomial coefficients. J Math Anal Appl 458(1):590–600
30. Guo VJW (2018) A q-analogues of the (J.2) supercongruence of van Hamme. J Math Anal Appl 466(1):776–788
31. Guo VJW, Liu J-C (2018) q-analogues of two Ramanujan-type formulas for $1/\pi$. J Differ Equ Appl 24(8):1368–1373
32. Guo VJW, Zudilin W (2018) Ramanujan-type formulae for $1/\pi$: q-analogues. Integral Transforms Spec Funct 29(7):505–513. https://doi.org/10.1080/10652469.2018.1454448
33. Hermite C (1872) Sur l'intégration des fractions rationnelles. Ann Sci École Norm Sup (2) 1:215–218
34. Hou Q-H, Wang R-H (2015) An algorithm for deciding the summability of bivariate rational functions. Adv Appl Math 64:31–49
35. Hou Q-H, Krattenthaler C, Sun Z-W (2018) On q-analogues of some series for π and π^2. Proceedings of the American Mathematical Society. https://doi.org/10.1090/proc/14374
36. Kauers M, Koutschan C, Zeilberger D (2009) Proof of Ira Gessel's lattice path conjecture. Proc Natl Acad Sci U S A 106(28):11502–11505
37. Kondratieva M, Sadov S (2005) Markov's transformation of series and the WZ method. Adv Appl Math 34(2):393–407
38. Koutschan C, Kauers M, Zeilberger D (2011) Proof of George Andrews's and David Robbins's q-TSPP conjecture. Proc Natl Acad Sci U S A 108(6):2196–2199
39. Liu Z-G (2012) Gauss summation and Ramanujan-type series for $1/\pi$. Int J Number Theory 08(02):289–297
40. Liu Z-G (2015) A q-extension of a partial differential equation and the Hahn polynomials. Ramanujan J 38(3):481–501
41. Long L (2011) Hypergeometric evaluation identities and supercongruences. Pac J Math 249(2):405–418
42. Pilehrood KH, Pilehrood TH (2008a) Simultaneous generation for zeta values by the Markov-WZ method. Discrete Math Theor Comput Sci 10(3):115–123
43. Pilehrood KH, Pilehrood TH (2008b) Generating function identities for $\zeta(2n+2)$, $\zeta(2n+3)$ via the WZ method. Electron J Combin 15(1):Research Paper 35, 9
44. Pilehrood KH, Pilehrood TH (2011) A q-analogue of the Bailey-Borwein-Bradley identity. J Symb Comput 46(6):699–711
45. Mohammed M (2005) The q-Markov-WZ method. Ann Comb 9(2): 205–221

46. Mohammed M, Zeilberger D (2004) The Markov-WZ method. Electron J Combin 11(1): 205–221
47. Mortenson E (2008) A p-adic supercongruence conjecture of van Hamme. Proc Am Math Soc 136(12):4321–4328
48. Ore O (1930) Sur la forme des fonctions hypergéométriques de plusieurs variables. J Math Pures Appl (9) 9(4):311–326
49. Ostrogradskiĭ MV (1845) De l'intégration des fractions rationnelles. Bull de la classe physico-mathématique de l'Acad Impériale des Sciences de Saint-Pétersbourg 4:145–167, 286–300
50. Petkovšek M, Wilf HS, Zeilberger D (1996) $A = B$. A K Peters Ltd., Wellesley, MA. With a foreword by Donald E. Knuth
51. Riordan J (1968) Combinatorial identities. Wiley, Hoboken, NJ
52. Sato M (1990) Theory of prehomogeneous vector spaces (algebraic part)—the English translation of Sato's lecture from Shintani's note. Nagoya Math J 120:1–34. Notes by Takuro Shintani. Translated from the Japanese by Masakazu Muro
53. Sun Z-W (2011) Super congruences and Euler numbers. Sci China Math 54(12):2509–2535
54. Sun X (2012) Some discussions on three kinds of WZ-equations. Master thesis, Soochow University, April 2012. Supervised by Xinrong Ma
55. Sun Z-W (2012) A refinement of a congruence result by van Hamme and Mortenson. Ill J Math 56(3):967–979
56. Sun Z-W (2013) Conjectures involving arithmetical sequences. In: Kanemitsu S, Li H, Liu J (eds) Number theory: arithmetic in Shangri-La, Proceedings of the 6th China-Japan Seminar (Shanghai, August 15–17, 2011). World Scientific Publishing, Singapore, pp 244–258
57. Sun Z-W (2013) Products and sums divisible by central binomial coefficients. Electron J Combin 20(1):91–109(19)
58. Sun Z-W (2018) Two q-analogues of Euler's formula $\zeta(2) = \pi^2/6$. Preprint. arXiv:arXiv:1802.01473
59. Tefera A (2010) What is ... a Wilf-Zeilberger pair? Not Am Math Soc 57(4):508–509
60. van Hamme L (1997) Some conjectures concerning partial sums of generalized hypergeometric series. In: p-adic functional analysis (Nijmegen, 1996). Lecture notes in pure and applied mathematics, vol 192. Dekker, New York, pp 223–236
61. Wilf HS, Zeilberger D (1992a) An algorithmic proof theory for hypergeometric (ordinary and "q") multisum/integral identities. Invent Math 108(3):575–633
62. Wilf HS, Zeilberger D (1992b) Rational function certification of multisum/integral/"q" identities. Bull Am Math Soc (N.S.) 27(1):148–153
63. Zeilberger D (1993) Closed form (pun intended!). In: A tribute to Emil Grosswald: number theory and related analysis. Contemporary mathematics, vol 143. American Mathematical Society, Providence, RI, pp 579–607
64. Zoladek H (1998) The extended monodromy group and Liouvillian first integrals. J Dynam Control Syst 4(1):1–28
65. Zudilin W (2007) More Ramanujan-type formulas for $1/\pi^2$. Russ Math Surv 62(3):634–636
66. Zudilin W (2009) Ramanujan-type supercongruences. J Number Theory 129(8):1848–1857
67. Zudilin W (2011) Arithmetic hypergeometric series. Russ Math Surv 66(2):369–420

Backward Error Analysis for Perturbation Methods

Robert M. Corless and Nicolas Fillion

Abstract We demonstrate via several examples how the backward error viewpoint can be used in the analysis of solutions obtained by perturbation methods. We show that this viewpoint is quite general and offers several important advantages. Perhaps the most important is that backward error analysis can be used to demonstrate the validity of the solution, however obtained and by whichever method. This includes a nontrivial safeguard against slips, blunders, or bugs in the original computation. We also demonstrate its utility in deciding when to truncate an asymptotic series, improving on the well-known rule of thumb indicating truncation just prior to the smallest term. We also give an example of elimination of *spurious* secular terms even when genuine secularity is present in the equation. We give short expositions of several well-known perturbation methods together with computer implementations (as scripts that can be modified). We also give a generic backward error based method that is equivalent to iteration (but we believe useful as an organizational viewpoint) for regular perturbation.

1 Introduction

As the title suggests, the main idea of this paper is to use backward error analysis (BEA) to assess and interpret solutions obtained by perturbation methods. The idea will seem natural, perhaps even obvious, to those who are familiar with the way in which backward error analysis has seen its scope increase dramatically

R. M. Corless (✉)
The Rotman Institute of Philosophy, The Ontario Research Center for Computer Algebra and The School of Mathematical and Statistical Sciences, The University of Western Ontario, London, ON, Canada
e-mail: rcorless@uwo.ca

N. Fillion
Department of Philosophy, Simon Fraser University Philosophy, Burnaby, BC, Canada
e-mail: nfillion@sfu.ca

© Springer Science+Business Media, LLC, part of Springer Nature 2019

N. Fillion et al. (eds.), *Algorithms and Complexity in Mathematics, Epistemology, and Science*, Fields Institute Communications 82,
https://doi.org/10.1007/978-1-4939-9051-1_3

since the pioneering work of Wilkinson in the 60s, e.g., [30, 31]. From its first use in numerical linear and polynomial algebraic problems, BEA has become a general method fruitfully applied to problems involving root finding, interpolation, numerical differentiation, quadrature, and the numerical solutions of ODEs, BVPs, DDEs, and PDEs, see, e.g, [8, 11, 16]. This is hardly a surprise when one considers that BEA offers several interesting advantages over a purely forward-error approach, including that it then becomes more obvious that some kind of conditioning or sensitivity analysis is needed.

BEA is often used in conjunction with perturbation methods. Not only is it the case that many algorithms' backward error analyses rely on perturbation methods, but the backward error is related to the forward error by a coefficient of sensitivity known as the condition number, which is itself a kind of sensitivity to perturbation. In this paper, we examine a different idea, namely, that perturbation methods themselves can also be interpreted within the backward error analysis framework. Our examples will have a classical feel, but the analysis and interpretation is what differs, and we will make general remarks about the benefits of this mode of analysis and interpretation.

The idea of using BEA to ensure correctness of a perturbation computation is not new. For instance, Boyd mentions the residual by name [5, p. 251, 289], although he does not use it systematically. The paper [34] names it and uses it. The paper [6] goes one step further, and discusses the meaning of the residual in a modeling context. Perhaps most significantly, the works of A.J. Roberts, including the codes freely available on his website for solving center manifolds and for solving DDE by perturbation methods, use the residual systematically. This is exemplified also in his recent book [27] which also includes discussion of programming computer algebra systems (in his case, REDUCE) to use the residual systematically in computing (and verifying) perturbation expansions.

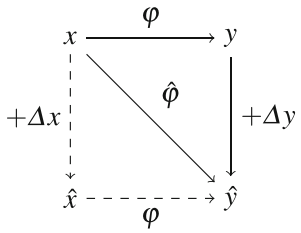
In this paper, we give an abstract framework that clarifies the systematic use of BEA for perturbation, and illustrates using examples from the literature how useful this can be. We also use computer algebra, in our case Maple. All Maple scripts used in this paper are available at www.publish.uwo.ca/~rcorless/PerturbationBEA. Our sources for examples include the venerable [3] and the wide-ranging [24]. Given that Google Scholar lists over 30,000 hits for books with either "pertubation" + "theory" or "perturbation" + "methods" in the title, our selection must necessarily be limited.

2 The Basic Method from the BEA Point of View

The basic idea of BEA is increasingly well-known in the context of numerical methods. The slogan *a good numerical method gives the exact solution to a nearby problem* very nearly sums up the whole perspective. Any number of more formal definitions and discussions exist—we like the one given in [8, chap. 1], as one might suppose is natural, but one could hardly do better than go straight to the source and

consult, e.g., [30–33]. More recently [15] has offered a good historical perspective. In what follows we give a brief formal presentation and then give detailed analyses by examples in subsequent sections.

Problems can generally be represented as maps from an input space $\mathscr{I}$ to an output space $\mathscr{O}$. If we have a problem $\varphi : \mathscr{I} \to \mathscr{O}$ and wish to find $y = \varphi(x)$ for some putative input $x \in \mathscr{I}$, lack of tractability might instead lead you to engineer a simpler problem $\hat{\varphi}$ from which you would compute $\hat{y} = \hat{\varphi}(x)$. Then $\hat{y} - y$ is the *forward error* and, provided it is small enough for your application, you can treat $\hat{y}$ as an approximation in the sense that $\hat{y} \approx \varphi(x)$. In BEA, instead of focusing on the forward error, we try to find an $\hat{x}$ such that $\hat{y} = \varphi(\hat{x})$ by considering the *backward error* $\Delta x = \hat{x} - x$, i.e., we try to find for which set of data our approximation method $\hat{\varphi}$ has exactly solved our reference problem φ. The general picture can be represented by the following commutative diagram:



We can see that, whenever x itself has many components, different backward error analyses will be possible since we will have the option of reflecting the forward error back into different selections of the components.

It is often the case that the map φ can be defined as the solution to $\phi(x, y) = 0$ for some operator ϕ, i.e., as having the form

$$x \xrightarrow{\varphi} \{y \mid \phi(x, y) = 0\} \ . \tag{1}$$

In this case, there will in particular be a simple and useful backward error resulting from computing the residual $r = \phi(x, \hat{y})$. Trivially $\hat{y}$ then exactly solves the reverse-engineered problem $\hat{\varphi}$ given by $\hat{\phi}(x, y) = \phi(x, y) - r = 0$. Thus, when the residual can be used as a backward error, this directly computes a reverse-engineered problem that our method has solved exactly. We are then in the fortunate position of having both a problem and its solution, and the challenge then consists in determining how similar the reference problem φ and the modified problems $\hat{\varphi}$ are, *and whether or not the modified problem is a good model for the phenomenon being studied.*

2.1 Regular Perturbation BEA-Style

Now let us introduce a *general framework for perturbation methods* that relies on the general framework for BEA introduced above. Perturbation methods are so numerous and varied, and the problems tackled are from so many areas, that it seems a general scheme of solution would necessarily be so abstract as to be difficult to use in any particular case. Actually, the following framework covers many methods. For simplicity of exposition, we will introduce it using the simple gauge functions $1, \varepsilon, \varepsilon^2, \ldots$, but note that extension to other gauges is usually straightforward (such as Puiseux, $\varepsilon^n \ln^m \varepsilon$, etc), as we will show in the examples. To begin with, let

$$F(x, u; \varepsilon) = 0 \tag{2}$$

be the operator equation we are attempting to solve for the unknown u. The dependence of F on the scalar parameter ε and on any data x is assumed but henceforth not written explicitly. In the case of a simple power series perturbation, we will take the mth order approximation to u to be given by the *finite* sum

$$z_m = \sum_{k=0}^{m} \varepsilon^k u_k \, . \tag{3}$$

The operator F is assumed to be Fréchet differentiable. For convenience we assume slightly more, namely, that for any u and v in a suitable region, there exists a linear invertible operator $F_1(v)$ such that

$$F(u) = F(v) + F_1(v)(u - v) + O\left(\|u - v\|^2\right) \, . \tag{4}$$

Here, $\|\cdot\|$ denotes any convenient norm. We denote the *residual* of z_m by

$$\Delta_m := F(z_m) \, , \tag{5}$$

i.e., Δ_m results from evaluating F at z_m instead of evaluating it at the reference solution u as in Eq. (2). If $\|\Delta_m\|$ is small, we say we have solved a "nearby" problem, namely, the reverse-engineered problem for the unknown u defined by

$$F(u) - F(z_m) = 0 \, , \tag{6}$$

which is exactly solved by $u = z_m$. Of course this is trivial. It is *not* trivial in consequences if $\|\Delta_m\|$ is small compared to data errors or modelling errors in the operator F. We will exemplify this point more concretely later.

We now suppose that we have somehow found $z_0 = u_0$, a solution with a residual whose size is such that

$$\|\Delta_0\| = \|F(u_0)\| = O(\varepsilon) \qquad \text{as} \qquad \varepsilon \to 0 \, . \tag{7}$$

Finding this u_0 is part of the art of perturbation; much of the rest is mechanical. Suppose now inductively that we have found z_n with residual of size

$$\|\Delta_n\| = O\left(\varepsilon^{n+1}\right) \quad \text{as} \quad \varepsilon \to 0\,.$$

Consider $F(z_{n+1})$ which, by definition, is just $F(z_n+\varepsilon^{n+1}u_{n+1})$. We wish to choose the term u_{n+1} in such a way that z_{n+1} has residual of size $\|\Delta_{n+1}\| = O(\varepsilon^{n+2})$ as $\varepsilon \to 0$. Using the Fréchet derivative of the residual of z_{n+1} at z_n, we see that

$$\Delta_{n+1} = F(z_n + \varepsilon^{n+1}u_{n+1}) = F(z_n) + F_1(z_n)\varepsilon^{n+1}u_{n+1} + O\left(\varepsilon^{2n+2}\right)\,. \tag{8}$$

By linearity of the Fréchet derivative, we also obtain $F_1(z_n) = F_1(z_0) + O(\varepsilon) = [\varepsilon^0]F_1(z_0) + O(\varepsilon)$. Here, $[\varepsilon^k]G$ refers to the coefficient of ε^k in the expansion of G. Let

$$\mathbf{A} = [\varepsilon^0]F_1(z_0)\,, \tag{9}$$

that is, the zeroth order term in $F_1(z_0)$. Thus, we reach the following expansion of Δ_{n+1}:

$$\Delta_{n+1} = F(z_n) + \mathbf{A}\varepsilon^{n+1}u_{n+1} + O\left(\varepsilon^{n+2}\right)\,. \tag{10}$$

Note that, in Eq. (8), one could keep $F_1(z_n)$, not simplifying to $\mathbf{A}$ and compute not just u_{n+1} but, just as in Newton's method, double the number of correct terms. However, this in practice is often too expensive [14, chap. 6], and so we will in general use this simplification. As noted, we only need $F_1(z_0)$ accurate to $O(\varepsilon)$, so in place of $F_1(z_0)$ in Eq. (10) we use $\mathbf{A}$.

As a result of the above expansion of Δ_{n+1}, we now see that to make $\Delta_{n+1} = O\left(\varepsilon^{n+2}\right)$, we must have $F(z_n) + \mathbf{A}\varepsilon^{n+1}u_{n+1} = O(\varepsilon^{n+2})$, in which case

$$\mathbf{A}u_{n+1} + \frac{F(z_n)}{\varepsilon^{n+1}} = \mathbf{A}u_{n+1} + \frac{\Delta_n}{\varepsilon^{n+1}} = O(\varepsilon)\,. \tag{11}$$

Since by hypothesis $\Delta_n = F(z_n) = O(\varepsilon^{n+1})$, we know that $\Delta_n/\varepsilon^{n+1} = O(1)$. In other words, to find u_{n+1} we solve the linear operator equation

$$\mathbf{A}u_{n+1} = -[\varepsilon^{n+1}]\Delta_n\,,$$

where, again, $[\varepsilon^{n+1}]$ is the coefficient of the $(n+1)$th power of ε in the series expansion of Δ. Note that by the inductive hypothesis the right hand side has norm $O(1)$ as $\varepsilon \to 0$. Then $\|\Delta_{n+1}\| = O(\varepsilon^{n+2})$ as desired, so u_{n+1} is indeed the coefficient we were seeking. We thus need $\mathbf{A} = [\varepsilon^0]F(z_0)$ to be invertible. If not, the

problem is singular, and essentially requires reformulation.[1] We shall see examples. If $\mathbf{A}$ is invertible, the problem is regular.

This general scheme can be compared to that of, say, [2]. Essential similarities can be seen. In Bellman's treatment, however, the residual is used implicitly, but not named or noted, and instead the equation defining u_{n+1} is derived by postulating an infinite expansion

$$u = u_0 + \varepsilon u_1 + \varepsilon^2 u_2 + \cdots . \tag{12}$$

By taking the coefficient of ε^{n+1} in the expansion of Δ_n we are implicitly doing the same work, but we will see advantages of this point of view. Also, note that in the frequent case of more general asymptotic sequences, namely Puiseux series or generalized approximations containing logarithmic terms, we can make the appropriate changes in a straightforward manner, as we will show below.

2.2 Conditioning and Sensitivity

We will not talk much about conditioning in this paper, although it is essential for mathematical modelling *even when you have the exact reference solution.* But note that in the abstract framework above, the norm of $\mathbf{A}^{-1}$ serves as an absolute condition number, giving a linear estimate of the forward error from a small perturbation to the problem. This corresponds of course, to comparing the solution as computed to one with just one more term in the expansion.

3 Algebraic Equations

We begin by applying the regular method from Sect. 2 to algebraic equations. We begin with a simple scalar equation and gradually increase the difficulty, thereby demonstrating the flexibility of the backward error point of view.

[1] We remark that it is a sufficient but not necessary condition for regular expansion to be able to find our initial point u_0 and to have invertible $\mathbf{A} = F_1(u_0; 0)$. A regular perturbation problem can be defined in many ways, not just in the way we have done, with invertible $\mathbf{A}$. For example, [3, Sec 7.2] essentially uses continuity in ε as $\varepsilon \to 0$ to characterize it. Another characterization is that for regular perturbation problems infinite perturbation series are convergent for some non-zero radius of convergence.

3.1 Regular Perturbation

In this section, after applying the method from Sect. 2 to a scalar equation, higher dimensional systems can be solved similarly. We give some computer algebra implementations (scripts that the reader may modify) of the basic method. Finally, in this section, we give an alternative method based on the Davidenko equation that is simpler to use in Maple.

3.1.1 Scalar Equations

Let us consider a simple example similar to many used in textbooks for classical perturbation analysis. Suppose we wish to find a real root of

$$x^5 - x - 1 = 0 \tag{13}$$

and, since the Abel-Ruffini theorem—which says that in general there are no solutions in radicals to equations of degree 5 or more—suggests it is unlikely that we can find an elementary expression for the solution of this *particular* equation of degree 5, we introduce a parameter which we call ε, and moreover which we suppose to be small. That is, we embed our problem in a parametrized family of similar problems. If we decide to introduce ε in the degree-1 term, so that

$$u^5 - \varepsilon u - 1 = 0\,, \tag{14}$$

we will see that we have a so-called regular perturbation problem.

To begin with, we wish to find a z_0 such that $\Delta_0 = F(z_0) = z_0^5 - \varepsilon z_0 - 1 = O(\varepsilon)$. Quite clearly, this can happen only if $z_0^5 - 1 = 0$. Ignoring the complex roots in this example, we take $z_0 = 1$. To continue the solution process, we now suppose that we have found

$$z_n = \sum_{k=0}^{n} u_k \varepsilon^k \tag{15}$$

such that $\Delta_n = F(z_n) = z_n^5 - \varepsilon z_n - 1 = O(\varepsilon^{n+1})$ and we wish to use our iterative procedure. We need the Fréchet derivative of F, which in this case is just

$$F_1(u) = 5u^4 - \varepsilon\,, \tag{16}$$

because

$$F(u) = u^5 - \varepsilon u - 1 = v^5 - \varepsilon v - 1 + F'(v)(u - v) + O(u - v)^2\,. \tag{17}$$

Hence, $\mathbf{A} = 5z_0^4 = 5$, which is invertible. As a result our iteration is $\Delta_n = F(z_n)$, i.e.,

$$5u_{n+1} = -[\varepsilon^{n+1}]\Delta_n \,. \tag{18}$$

Carrying out a few steps we have

$$\Delta_0 = F(z_0) = F(1) = 1 - \varepsilon - 1 = -\varepsilon \tag{19}$$

so

$$5 \cdot u_1 = -[\varepsilon]\Delta_0 = -[\varepsilon](-\varepsilon) = 1 \,. \tag{20}$$

Thus, $u_1 = 1/5$. Therefore, $z_1 = 1 + \varepsilon/5$ and

$$\Delta_1 = \left(1 + \frac{\varepsilon}{5}\right)^5 - \varepsilon\left(1 + \frac{\varepsilon}{5}\right) - 1 \tag{21}$$

$$= \left(1 + 5\frac{\varepsilon}{5} + 10\frac{\varepsilon^2}{25} + O\left(\varepsilon^3\right)\right) - \varepsilon - \frac{\varepsilon^2}{5} - 1 \tag{22}$$

$$= \left(\frac{2}{5} - \frac{1}{5}\right)\varepsilon^2 + O\left(\varepsilon^3\right) = \frac{1}{5}\varepsilon^2 + O\left(\varepsilon^3\right) \,. \tag{23}$$

Then we find that $Au_1 = -1/5$ and thus $u_1 = -1/25$. So, $u = 1 + \varepsilon/5 - \varepsilon^2/25 + O(\varepsilon^3)$. Finding more terms by this method is clearly possible although tedium might be expected at higher orders. Luckily nowadays computers and programs are widely available that can solve such problems without much human effort, but before we demonstrate that, let's compute the residual of our computed solution so far:

$$z_2 = 1 + \frac{1}{5}\varepsilon - \frac{1}{25}\varepsilon^2 \,.$$

Then $\Delta_2 = z_2^5 - \varepsilon z_2 - 1$ is

$$\begin{aligned}\Delta_2 &= \left(1 + \frac{1}{5}\varepsilon - \frac{1}{25}\varepsilon^2\right)^5 - \varepsilon\left(1 + \frac{1}{5}\varepsilon - \frac{1}{25}\varepsilon^2\right) - 1 \\ &= -\frac{1}{25}\varepsilon^3 - \frac{3}{125}\varepsilon^4 + \frac{11}{3125}\varepsilon^5 + \frac{3}{125}\varepsilon^6 - \frac{2}{15625}\varepsilon^7 \\ &\quad - \frac{1}{78125}\varepsilon^8 + \frac{1}{390625}\varepsilon^9 - \frac{1}{9765675}\varepsilon^{10} \,.\end{aligned} \tag{24}$$

We note the following. First, z_2 exactly solves the modified equation

$$x^5 - \varepsilon x - 1 \;\; + \frac{1}{25}\varepsilon^3 + \frac{3}{25}\varepsilon^4 - \ldots + \frac{1}{9765625}\varepsilon^{10} = 0 \tag{25}$$

which is $O(\varepsilon^3)$ different to the original. Second, the complete residual was computed rationally: there is no error in saying that $z_2 = 1 + \varepsilon/5 - \varepsilon^2/25$ solves Eq. (25) exactly. Third, if $\varepsilon = 1$ then $z_2 = 1 + 1/5 - 1/25 = 1.16$ exactly (or $14/25$ if you prefer), and the residual is then $(29/25)^5 - 29/25 - 1 \doteq -0.059658$, showing that 1.16 is the exact root of an equation about 6% different to the original.

Something simple but importantly different to the usual treatment of perturbation methods has happened here. We have assessed the quality of the solution in an explicit fashion without concern for convergence issues or for the exact solution to $x^5 - x - 1 = 0$, which we term the reference problem. We use this term because its solution will be the reference solution. We can't call it the "exact" solution because z_2 is *also* an "exact" solution, namely to Eq. (25).

Every numerical analyst and applied mathematician knows that this isn't the whole story—we need some evaluation or estimate of the effects of such perturbations of the problem. One effect is the difference between z_2 and x, the reference solution, and this is what people focus on. We believe this focus is sometimes excessive. The are other possible views. For instance, the modeller might care more that the backward error be *physically reasonable*. As an example, if $\varepsilon = 1$ and $z_2 = 1.16$ then z_2 exactly solves $y^5 - y - a = 0$ where $a \neq 1$ but rather $a \doteq 0.9403$. If the original equation was really $u^5 - u - \alpha = 0$ where $\alpha = 1 \pm 5\%$ we might be inclined to accept $z_2 = 1.16$ because, for all we know, we might have the true solution (even though we're outside the $\pm 5\%$ range, we're only just outside; and how confident are we in the $\pm 5\%$, after all?).

3.1.2 Simple Computer Algebra Solution

The following Maple script can be used to solve this or similar problems $f(u; \varepsilon) = 0$. Other computer algebra systems can also be used.

```
# Perturbation solution of F(u;epsilon) = 0
macro(e =varepsilon); #saves typing
F := z -> z^5-e*z-1;
# Zeroth order solution, by inspection, is
z := 1; #solve(eval(F(z),e=0),z);
A := coeff(series((D(F))(z), e, 1), e, 0);
# A must be nonzero for regularity
N := 3; #number of terms
Delta := F(z); #initial residual, must be O(e)
# Now, the iteration:
for k to N do
        u := -coeff(series(Delta, e, k+1), e, k);
        z := z + u*e^k/A;
        Delta := F(z);
end do;
z;
series(Delta, e, N+3);
```

That code is a straightforward implementation of the general scheme presented in Sect. 2. Its results, translated into LaTeX and cleaned up a bit, are that

$$z = 1 + \frac{1}{5}\varepsilon - \frac{1}{25}\varepsilon^2 + \frac{1}{125}\varepsilon^3 \tag{26}$$

and that the residual of this solution is

$$\Delta = \frac{21}{3125}\varepsilon^5 + O\left(\varepsilon^6\right) . \tag{27}$$

With $N = 3$, we get an extra order of accuracy as the next term in the series is zero, but this result is serendipitous.

3.1.3 Systems of Algebraic Equations

Regular perturbation for systems of equations using the framework from Sect. 2 is straightforward. We include an example to show some computer algebra and for completeness. Consider the following two equations in two unknowns:

$$f_1(v_1, v_2) = v_1^2 + v_2^2 - 1 - \varepsilon v_1 v_2 = 0 \tag{28}$$

$$f_2(v_1, v_2) = 25v_1 v_2 - 12 + 2\varepsilon v_1 = 0 \tag{29}$$

When $\varepsilon = 0$ these equations determine the intersections of a hyperbola with the unit circle. There are four such intersections: $(3/5, 4/5)$, $(4/5, 3/5)$, $(-3/5, -4/5)$ and $(-4/5, -3/5)$. The Jacobian matrix (which gives us the Fréchet derivative in the case of algebraic equations) is

$$F_1(v) = \begin{bmatrix} \frac{\partial f_1}{\partial v_1} & \frac{\partial f_1}{\partial v_2} \\ \frac{\partial f_2}{\partial v_1} & \frac{\partial f_2}{\partial v_2} \end{bmatrix} = \begin{bmatrix} 2v_1 & 2v_2 \\ 25v_2 & 25v_1 \end{bmatrix} + O(\varepsilon) . \tag{30}$$

Taking for instance $u_0 = [3/5, 4/5]^T$ we have

$$A = F_1(u_0) = \begin{bmatrix} 6/5 & 8/5 \\ 20 & 15 \end{bmatrix} . \tag{31}$$

Since $\det A = -14 \neq 0$, A is invertible and indeed

$$A^{-1} = \begin{bmatrix} -15/14 & 4/25 \\ 10/7 & -3/35 \end{bmatrix} . \tag{32}$$

The residual of the zeroth order solution is

$$\Delta_0 = F\left(\frac{3}{5}, \frac{4}{5}\right) = \begin{bmatrix} -12/25 \\ 6/5 \end{bmatrix}, \tag{33}$$

so $-[\varepsilon]\Delta_0 = [12/25, -6/5]^T$. Therefore

$$u_1 = \begin{bmatrix} u_{11} \\ u_{12} \end{bmatrix} = A^{-1} \begin{bmatrix} 12/25 \\ -6/25 \end{bmatrix} = \begin{bmatrix} -114/175 \\ 138/175 \end{bmatrix} \tag{34}$$

and $z_1 = u_0 + \varepsilon u_1$ is our improved solution:

$$z_1 = \begin{bmatrix} 3/5 \\ 4/5 \end{bmatrix} + \varepsilon \begin{bmatrix} -114/175 \\ 138/175 \end{bmatrix}. \tag{35}$$

To guard against slips, blunders, and bugs (some of those calculations were done by hand, and some were done in Sage on an Android phone) we compute

$$\Delta_1 = F(z_1) = \varepsilon^2 \begin{bmatrix} 6702/6125 \\ -17328/1225 \end{bmatrix} + O\left(\varepsilon^3\right). \tag{36}$$

That computation was done in Maple, completely independently. Initially it came out $O(\varepsilon)$ indicating that something was not right; tracking the error down we found a typo in the Maple data entry (183 was entered instead of 138). Correcting that typo we find $\Delta_1 = O(\varepsilon^2)$ as it should be. Here is the corrected Maple code:

```
# Residual computation for a system of two equations
macro(e = varepsilon); #saves typing
f1 := (v1,v2) -> v1^2 + v2^2 - 1 - e*v1*v2;
f2 := (v1,v2) -> 25*v1*v2 - 12 + 2*e*v1;
z11 := 3/5 + e*(-114/175);
z12 := 4/5 + e*138/175;
Delta11 := series( f1(z11,z12), e, 3);
Delta12 := series( f2(z11,z12), e, 3);
```

Just as for the scalar case, this process can be systematized and we give one way to do so in Maple, below. The code is not as pretty as the scalar case is, and one has to explicitly "map" the series function and the extraction of coefficients onto matrices and vectors, but this demonstrates feasibility.

```
# Residual computation for a system of two equations
macro(e = varepsilon); #saves typing
z := Vector(2,[3/5,4/5]); # z_0 = u_0
F := u -> Vector(2,
                 [ u[1]^2 + u[2]^2 - 1 - e*u[1]*u[2],
                   25*u[1]*u[2] - 12 + 2*e*u[1] ]);
```

```
A := VectorCalculus[Jacobian](
                  [ F([x,y])[1], F([x,y])[2]], [x,y] );
A := eval( A, [x=z[1], y=z[2], e=0] );
N := 3;
Delta := F(z);
for k to N do
        u := map(t -> -coeff( t, e, k ),
                  map( series, Delta, e, k+1 )
                 );
        z := z + LinearAlgebra[LinearSolve]( A, u )*e^k;
        Delta := F( z );
end do:
z;
map( series, Delta, e , N+2 );
```

This code computes z_3 correctly and gives a residual of $O(\varepsilon^4)$. From the backward error point of view, this code finds the intersection of curves that differ from the specified ones by terms of $O(\varepsilon^4)$. In the next section, we show a way to use a built-in feature of Maple to do the same thing with less human labour.

3.1.4 Solving Algebraic Systems by the Davidenko Equation

The general method outlined in Sect. 2 applies directly to systems of equations, as we just saw. Maple does not have a built-in facility to solve algebraic equations in series such as that one. Instead, Maple has a built-in facility for solving differential equations in series that (at the time of writing) is superior to its built-in facility for solving algebraic equations in series, because the latter can only handle scalar equations. This may change in the future, but it may not because there is the following simple workaround. To solve

$$F(u;\varepsilon) = 0 \tag{37}$$

for a function $u(\varepsilon)$ expressed as a series, simply differentiate to get

$$D_1(F)(u,\varepsilon)\frac{du}{d\varepsilon} + D_2(F)(u,\varepsilon) = 0\,. \tag{38}$$

Boyd [5] calls this the Davidenko equation. If we solve this in Taylor series with the initial condition $u(0) = u_0$, we have our perturbation series. Notice that what we were calling $\mathbf{A} = [\varepsilon^0]F_1(u_0)$ occurs here as $D_1(F)(u_0, 0)$ and this needs to be nonsingular to be solved as an ordinary differential equation; if $\text{rank}(D_1(F)(u_0, 0)) < n$ where n is the dimension of F, then this is in fact a nontrivial differential algebraic equation that Maple may still be able to solve using advanced techniques (see, e.g., [1]). The code below solves the same example as in the previous section.

```
# Residual computation for a system of two equations
macro(e = varepsilon); #saves typing
Order := 4;
z := Vector(2,[3/5,4/5]); # z_0 = u_0
F := u -> Vector(2,
                 [ u[1]^2 + u[2]^2 - 1 - e*u[1]*u[2],
                   25*u[1]*u[2] - 12 + 2*e*u[1] ]);
Zer := F( [ x(e), y(e) ] );
# That asks for F to be evaluated at functions x(e)
# and y(e) that are yet unspecified.
diffeqs := { diff( Zer[1], e ), diff( Zer[2], e ) };
# That creates a set of two differential equations,
# one from each component of F.
# Each equation will contain both dx/de and dy/de.
iniconds := { x(0) = z[1] , y(0) = z(2) };
sol := dsolve( diffeqs union iniconds,
               {x(e), y(e)}, type=series );
Delta := eval( F( [x(e), y(e)] ),
               map(convert, sol, polynom) ):
map( series, Delta, e, Order+2 );
```

This generates (to the specified value of the order, namely, `Order=4`) the solution

$$x(\varepsilon) = \frac{3}{5} - \frac{114}{175}\varepsilon + \frac{119577}{42875}\varepsilon^2 - \frac{43543632}{2100875}\varepsilon^3 \tag{39}$$

$$y(\varepsilon) = \frac{4}{5} + \frac{138}{175}\varepsilon - \frac{119004}{42875}\varepsilon^2 + \frac{43245168}{2100875}\varepsilon^3\,, \tag{40}$$

whose residual is $O(\varepsilon^4)$.

3.2 Puiseux Series

Puiseux series are simply Taylor series or Laurent series with fractional powers. A standard example is

$$\sin\sqrt{x} = x^{1/2} - \frac{1}{3!}x^{3/2} + \frac{1}{5!}x^{5/2} + \cdots \tag{41}$$

A simple change of variable (e.g. $t = \sqrt{x}$ so $x = t^2$) is enough to convert to Taylor series. Once the appropriate power n is known for $\varepsilon = \mu^n$, perturbation by Puiseux expansion reduces to computations similar to those we've seen already. For instance, had we chosen to embed $u^5 - u - 1$ in the family $u^5 - \varepsilon(u + 1)$ (which is somehow conjugate to the family of the last section), then because the equation

becomes $u^5 = 0$ when $\varepsilon = 0$ we see that we have a fivefold root to perturb, and we thus suspect we will need Puiseux series.

For scalar equations, there are built-in facilities in Maple for Puiseux series, which gives yet another way in Maple to solve scalar algebraic equations perturbatively. One can use the `RootOf` construct to do so as follows:

```
macro(e = varepsilon);
Order := 2;
alias(alpha = RootOf(z^5-1, z));
f := u -> u^5 - e*(u+1);
z := convert( series( RootOf(f(u),u), e), polynom);
Delta := series(f(z), e, Order+2):
map(simplify, Delta);
```

This yields

$$z = \alpha\varepsilon^{1/5} + \frac{1}{5}\alpha^2\varepsilon^{2/5} - \frac{1}{25}\alpha^3\varepsilon^{3/5} + \frac{1}{125}\alpha^4\varepsilon^{4/5} - \frac{21}{15626}\alpha\varepsilon^{6/5}\,. \tag{42}$$

This series describes all paths, accurately for small ε. Note that the command

```
alias(alpha = RootOf(u^5-1,u))
```

is a way to tell Maple that α represents a fixed fifth root of unity. Exactly which fixed root can be deferred till later. Working instead with the default value for the environment variable `Order`, namely `Order := 6`, gets us a longer series for z containing terms up to $\varepsilon^{29/5}$ but not $\varepsilon^{30/5} = \varepsilon^6$. Putting the resulting z_6 back into $f(u)$ we get a residual

$$\Delta_6 = f(z_6) = \frac{23927804441356816}{14551915228366851806640625}\varepsilon^7 + O(\varepsilon^8) \tag{43}$$

Thus we expect that for small ε the residual will be quite small, because the root is well-conditioned. For instance, with $\varepsilon = 1$ the exact residual is, for $\alpha = 1$, $\Delta_6 = 1.2 \cdot 10^{-9}$. This tells us that this approximation ought to get us quite accurate roots, and indeed we do.

3.3 Singular Perturbation

Suppose that instead of embedding $u^5 - u - 1 = 0$ in the regular family we used in the previous section, we had used $\varepsilon u^5 - u - 1 = 0$. If we run our previous Maple programs, we find that the zeroth order solution is unique, and $z_0 = -1$. The Fréchet derivative is -1 to $O(\varepsilon)$, and so $u_{n+1} = [\varepsilon^{n+1}]\Delta_n$ for all $n \geq 0$. We find, for instance,

$$z_7 = -1 - \varepsilon - 5\varepsilon^2 - 35\varepsilon^3 - 285\varepsilon^4 - 2530\varepsilon^5 - 23751\varepsilon^6 - 231880\varepsilon^7 \tag{44}$$

which has residual $\Delta_7 = O(\varepsilon^8)$ but with a larger integer as the constant hidden in that O symbol. For $\varepsilon = 0.2$, the value of z_7 becomes

$$z_7 \doteq -7.4337280 \tag{45}$$

while $\Delta_7 = -4533.64404$, which is not small at all. Thus we have no evidence this perturbation solution is any good: we have the exact solution to $u^5 - 0.2u - 1 = -4533.64404$ or $u^5 - 0.2u + 4532.64404 = 0$, probably not what was intended (and if it was, it would be a colossal fluke). Note that we do not need to know a reference value of a root of $u^5 - 0.2u - 1$ to determine this. Trying a smaller ε, we find that if $\varepsilon = 0.05$ we have $z_7 \doteq -1.07$ and $\Delta_7 \doteq -1.2 \cdot 10^{-4}$. This means z_7 is an exact root of $u^5 - 0.05u - 1.00012$; which may very well be what we want.

But this computation, valid as it is, only found one root out of five, and then only for sufficiently small ε. We now turn to the roots that go to infinity as $\varepsilon \to 0$. Preliminary investigation similar to that of Sect. 3.2 shows that it is convenient to replace ε by μ^4. Many singular perturbation problems including this one can be turned into regular ones by rescaling. Putting $u = y/\mu$, we get

$$\mu^4 \left(\frac{y}{\mu}\right)^5 - \frac{y}{\mu} - 1 = 0\,, \tag{46}$$

which reduces to

$$y^5 - y - \mu = 0\,. \tag{47}$$

This is now regular in μ. The zeroth order the equation is $y(y^4 - 1) = 0$ and the root $y = 0$ just recovers the regular series previously attained; so we let α be a root of $y^4 - 1$, i.e., $\alpha \in \{1, -1, i, -i\}$. A very similar Maple program (to either of the previous two) gives

$$y_5 = \alpha + \frac{1}{4}\mu - \frac{5}{32}\alpha^3\mu^2 + \frac{5}{32}\alpha^2\mu^3 - \frac{385}{2048}\alpha\mu^4 + \frac{1}{4}\mu^5 \tag{48}$$

so our appoximate solution is y_5/μ or

$$z_5 = \frac{\alpha}{\mu} + \frac{1}{4} - \frac{5}{32}\alpha^3\mu^2 - \frac{385}{2048}\alpha\mu^3 + \frac{1}{4}\mu^4 \tag{49}$$

which has residual *in the original equation*

$$\Delta_5 = \mu^4 z^5 - z - 1 = \frac{23205}{16384}\alpha^3\mu^5 - \frac{21255}{65536}\alpha^2\mu^6 + O(\mu^7)\,. \tag{50}$$

That is, z_5 exactly solves $\mu^4 u^5 - u - 1 - {}^{23205}/_{16384}\,\alpha^2\mu^5 = O(\mu^6)$ instead of the one we had wanted to solve. This differs from the original by $O(|\varepsilon|^{5/4})$, and for small enough ε this may suffice.

3.4 Optimal Backward Error

Interestingly enough, we can do better. The residual is only one kind of backward error. Taking the lead from the Oettli-Prager theorem [8, chap. 6], we look for equations of the form

$$\left(\mu^4 + \sum_{j=10}^{15} a_j \mu^j\right) u^5 - u - 1 \tag{51}$$

for which z_5 is a better solution yet. Simply equating coefficients of the residual

$$\tilde{\Delta}_5 = \left(\mu^4 + \sum_{j=10}^{15} a_j \mu^j\right) z_5^5 - z_5 - 1 \tag{52}$$

to zero, we find

$$(\mu^4 - \frac{23205}{16384}\alpha^2\mu^{10} + \frac{2145}{1024}\alpha\mu^{11})z_5^5 - z_5 - 1 = \frac{12165535425}{1073741824}\alpha\mu^{11} + O(\mu^{12}) \tag{53}$$

and thus z_5 solves an equation that is $O(\mu^{10/4}) = O(\varepsilon^{5/2})$ close to the original, not just an Eq. (50) that is $O(\mu^6) = O(|\varepsilon|^{5/4})$. This is a superior explanation of the quality of z_5. This was obtained with the following Maple code:

```
# Perturbation solution of F(u;epsilon) = 0
macro( e=varepsilon );
e := mu^4;
Forig := z -> e*z^5 - z - 1;
F := y -> y^5 - y - mu;
# Zeroth order solution, by inspection:
alias(alpha = RootOf(Z^4-1, Z));
y := alpha;
A := coeff( series( (D(F))(y), mu, 1), mu, 0);
A := simplify(A);
N := 5;
Delta := simplify( F(y) );
for k to N do
```

```
        u := -coeff( series(Delta, mu, k+1), mu, k);
        y := y+u*mu^k/A;
        Delta := simplify( F(y) );
end do:
y;
series(Delta, mu, N+3);
M := 5+2*N;
modified := u -> (mu^4+add(a[j]*mu^j, j=5+N..M))*u^5-u-1;
z := map( simplify, series(y/mu, mu, N+1) );
zer := series(modified(z), mu, M+1):
eqs := [seq(simplify(coeff(zer, mu, k)), k = N .. M-5)];
sol := solve(eqs, [seq(a[j], j = 5+N .. M)]);
perteq := eval(modified(U), sol[1]):
newresid := eval(perteq, U = z):
map(simplify, series(newresid, mu, M+2));
```

Computing to higher orders (see the worksheet) gives e.g. that z_8 is the exact solution to an equation that differs by $O(\mu^{13})$ from the original, or better than $O(\varepsilon^3)$. This in spite of the fact that the basic residual $\Delta_8 = O(\varepsilon^{9/4})$, only slightly better than $O(\varepsilon^2)$.

We will see other examples of improved backward error over residual for singularly-perturbed problems. In retrospect it's not so surprising, or shouldn't have been: singular problems are sensitive to changes in the leading term, and so it takes less effort to match a given solution.

3.5 A Hyperasymptotic Example

In [5, sect. 15.3, pp. 285–288], Boyd takes up the perturbation series expansion of the root near -1 of

$$f(x,\varepsilon) = 1 + x + \varepsilon \operatorname{sech}\left(\frac{x}{\varepsilon}\right) = 0\,, \tag{54}$$

a problem he took from [17, p. 22]. After computing the desired expansion using a two-variable technique, Boyd then sketches an alternative approach suggested by one of us (based on [9]), namely to use the Lambert W function. Unfortunately, there are a number of sign errors in Boyd's equation (15.28). We take the opportunity here to offer a correction, together with a residual-based analysis that confirms the validity of the correction. First, the erroneous formula: Boyd has

$$z_0 = \frac{W(-2e^{1/\varepsilon})\varepsilon - 1}{\varepsilon} \tag{55}$$

and $x_0 = -\varepsilon z_0$, so allegedly $x_0 = 1 - \varepsilon W(-2\varepsilon^{1/\varepsilon})$. This can't be right: as $\varepsilon \to 0^+$, $e^{1/\varepsilon} \to \infty$ and the argument to W is negative and large; but W is real only if its argument is between $-e^{-1}$ and 0, if it's negative at all. We claim that the correct formula is

$$x_0 = -1 - \varepsilon W(2e^{-1/\varepsilon}) \tag{56}$$

which shows that the errors in Boyd's equation (15.28) are explainable as trivial. Indeed, Boyd's derivation is correct up to the last step; rather than fill in the algebraic details of the derivation of formula (56), we here verify that it works by computing the residual:

$$\Delta_0 = 1 + x_0 + \varepsilon \operatorname{sech}\left(\frac{x_0}{\varepsilon}\right). \tag{57}$$

For notational simplicity, we will omit the argument to the Lambert W function and just write W for $W(2e^{-1/\varepsilon})$. Then, note that $\operatorname{sech}(x_0/\varepsilon) = \operatorname{sech}(1+\varepsilon W/\varepsilon)$ since each sech is even, and that

$$\operatorname{sech}\left(\frac{x_0}{\varepsilon}\right) = \frac{2}{e^{x_0/\varepsilon} + e^{-x_0/\varepsilon}} = \frac{1}{e^{(1/\varepsilon)+W} + e^{-1/\varepsilon - W}}. \tag{58}$$

Now, by definition,

$$We^W = 2e^{-1/\varepsilon} \tag{59}$$

and thus we obtain

$$e^W = \frac{2e^{-1/\varepsilon}}{W} \qquad \text{and} \qquad e^{-W} = \frac{We^{1/\varepsilon}}{2}. \tag{60}$$

It follows that

$$\operatorname{sech}\left(\frac{x_0}{\varepsilon}\right) = \frac{2}{2/W + W/2} = \frac{W}{1 + W^2/4}, \tag{61}$$

and hence the residual is

$$\begin{aligned} \Delta_0 &= 1 + (-1 - \varepsilon W) + \varepsilon \frac{W}{1 + W^2/4} = \frac{-\varepsilon W(1 + W^2/4) + \varepsilon W}{1 + W^2/4} \\ &= \frac{-\varepsilon W^3/4}{1 + W^2/4} = \frac{-\varepsilon W^3}{4 + W^2}. \end{aligned} \tag{62}$$

Now $W = W(2e^{-1/\varepsilon})$ and as $\varepsilon \to 0^+$, $2e^{-1/\varepsilon} \to 0$ rapidly; since the Taylor series for $W(z)$ starts as $W(z) = z - z^2 + \frac{3}{2}z^3 + \ldots$, we have that $W(2e^{-1/\varepsilon}) \sim 2e^{-1/\varepsilon}$ and therefore

$$\Delta_0 = -\varepsilon 2e^{-3/\varepsilon} + O(e^{-5/\varepsilon}) \,. \tag{63}$$

We see that this residual is very small indeed. But we can say even more. Boyd leaves us the exercise of computing higher order terms; here is our solution to the exercise. A Newton correction would give us

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \tag{64}$$

and we have already computed $f(x_0) = \Delta_0$. What is $f'(x_0)$? Since $f(x) = 1 + x + \varepsilon \operatorname{sech}(x/\varepsilon)$, this derivative is

$$f'(x) = 1 - \operatorname{sech}\left(\frac{x}{\varepsilon}\right) \tanh\left(\frac{x}{\varepsilon}\right) \,. \tag{65}$$

Simplifying similarly to Eq. (61), we obtain

$$\tanh\left(\frac{x_0}{\varepsilon}\right) = \frac{e^{1/\varepsilon+W} - e^{-1/\varepsilon-W}}{e^{1/\varepsilon+W} + e^{-1/\varepsilon+W}} = \frac{\frac{2}{W} - \frac{W}{2}}{\frac{2}{W} + \frac{W}{2}} = \frac{4 - W^2}{4 + W^2} \,. \tag{66}$$

Thus

$$f'(x_0) = 1 - \operatorname{sech}\left(\frac{x_0}{\varepsilon}\right) \tanh\left(\frac{x_0}{\varepsilon}\right) = 1 - \frac{W(1 - W^2/4)}{(1 + W^2/4)^2} \,. \tag{67}$$

It follows that

$$x_1 = x_0 - \frac{\Delta_0}{f'(x_0)} = -1 - \varepsilon W + \frac{\varepsilon W^3/4 + W^2}{1 - \dfrac{W(1 - W^2/4)}{(1 + W^2/4)^2}} \tag{68}$$

$$= -1 - \varepsilon W + \frac{\varepsilon W^3(4 + W^2)}{16 - 16W + 8W^2 + 4W^3 + W^4} \tag{69}$$

$$= -1 - \varepsilon W + \frac{\varepsilon}{4} W^3 + \frac{\varepsilon}{4} W^4 + \frac{3}{16} \varepsilon W^5 - \frac{11}{64} \varepsilon W^6 + O(W^7) \tag{70}$$

Finally, the residual of x_1 is

$$\Delta_1 = 4\varepsilon e^{-7/\varepsilon} + O(\varepsilon e^{-8/\varepsilon}) \,. \tag{71}$$

We thus see an example of the use of $f'(x_0)$ instead of just A, as discussed in Sect. 2, to approximately double the number of correct terms in the approximation.

This analysis can be implemented in Maple as follows:

```
with(MultiSeries);
macro(e = varepsilon);
```

```
alias(W = LambertW);
f := x -> 1 + x + e*sech(x/e);
df := D(f);
x[0] := -1 - e*W(2*exp(-1/e));
Delta[0] := f( x[0] );
series(Delta[0], e, 3);
x[1] := x[0] - Delta[0]/df(x[0]);
Delta[1] := f(x[1]);
s := multiseries(x[1],e=0);
scale := SeriesInfo[Scale](s);
multiseries(x[1],scale,3);
multiseries(Delta[1],scale,5);
# In what follows we have substituted expressions in W
# for sech and tanh since Maple couldn't simplify
# the expression well.
x[1] := -1-e*W+e*W^3/((4+W^2)*(1-W*(1-(1/4)*W^2)
                                  /(1+(1/4)*W^2)^2));
change := factor(x[1]+1+e*W);
series(change,W=0,8);
```

Note that we had to use the MultiSeries package [28] to expand the series in Eq. (71), for understanding how accurate z_2 was. z_2 is slightly more lacunary than the two-variable expansion in [5], because we have a zero coefficient for W^2.

4 Divergent Asymptotic Series

Before we begin, a note about the section title: some authors give the impression that the word "asymptotic" is used *only* for divergent series, and so the title might seem redundant. But the proper definition of an asymptotic series can include convergent series (see, e.g., [10]), as it means that the relevant limit is not as the number of terms N goes to infinity, but rather as the variable in question (be it ε, or x, or whatever) approaches a distinguished point (be it 0, or infinity, or whatever). In this sense, an asymptotic series might diverge as N goes to infinity, or it might converge, but typically we don't care. We concentrate in this section on divergent asymptotic series.

Beginning students are often confused when they learn the usual "rule of thumb" for optimal accuracy when using divergent asymptotic series, namely to truncate the series *before* adding in the smallest (magnitude) term. This rule is usually motivated by an analogy with *convergent* alternating series, where the error is less than the magnitude of the first term neglected. But why should this work (if it does) for divergent series?

The answer we present in this section isn't as clear-cut as we would like, but nonetheless we find it explanatory. The basis for the answer is that one can measure

the residual Δ that arises on truncating the series at, say, M terms, and choose M to minimize the residual. Since the forward error is bounded by the condition number times the size of the residual, by minimizing $\|\Delta\|$ one minimizes a bound on the forward error. It often turns out that this method gives the same M as the rule of thumb, though not always.

An example may clarify this. We use the large-x asymptotics of $J_0(x)$, the zeroth-order Bessel function of the first kind. In [23, section 10.17(i)], we find the following asymptotic series, which is attributed to Hankel:

$$J_0(x) = \left(\frac{2}{\pi x}\right)^{1/2} \left(A(x)\cos\left(x - \frac{\pi}{4}\right) - B(x)\sin\left(x - \frac{\pi}{4}\right)\right) \tag{72}$$

where

$$A(x) = \sum_{k\geq 0} \frac{a_{2k}}{x^{2k}} \qquad \text{and} \qquad B(x) = \sum_{k\geq 0} \frac{a_{2k+1}}{x^{2k+1}} \tag{73}$$

and where

$$\begin{aligned} a_0 &= 1 \\ a_k &= \frac{(-1)^k}{k!8^k} \prod_{j=1}^{k} (2j-1)^2 \,. \end{aligned} \tag{74}$$

For the first few a_ks, we get

$$a_0 = 1, a_1 = -\frac{1}{8}, a_2 = -\frac{9}{128}, a_3 = \frac{75}{1024}\,, \tag{75}$$

and so on. The ratio test immediately shows the two series (73) diverge for all finite x.

Luckily, we always have to truncate anyway, and if we do, the forward errors get arbitrarily small so long as we take x arbitrarily large. Because the Bessel functions are so well-studied, we have alternative methods for computation, for instance

$$J_0(x) = \frac{1}{\pi}\int_0^{\pi} \cos(x\sin\theta)d\theta \tag{76}$$

which, given x, can be evaluated numerically (although it's ill-conditioned in a relative sense near any zero of $J_0(x)$). So we can directly compute the forward error. But let's pretend that we can't. We have the asymptotic series, and not much more. Or course we have to have a defining equation—Bessel's differential equation

$$x^2y'' + xy' + x^2y = 0 \tag{77}$$

with the appropriate normalizations at ∞. We look at

$$y_{N,M} = \left(\frac{2}{\pi x}\right)^{1/2} A_N(x)\cos\left(x-\frac{\pi}{4}\right) - \frac{2}{\pi x} B_M(x)\cos\left(x-\frac{\pi}{4}\right) \tag{78}$$

where

$$A_N(x) = \sum_{k=0}^{N} \frac{a_{2k}}{x^{2k}} \qquad \text{and} \qquad B_M(x) = \sum_{k=0}^{M} \frac{a_{2k+1}}{x^{2k+1}}\,. \tag{79}$$

Inspection shows that there are only two cases that matter: when we end on an even term a_{2k} or on an odd term a_{2k+1}. The first terms omitted will be odd and even. A little work shows that the residual

$$\Delta = x^2 y''_{N,M} + x y'_{N,M} + x^2 y_{N,M} \tag{80}$$

is just

$$\frac{(k+1/2)^2 a_k}{x^{k+1/2}} \cdot \left\{ \begin{array}{c} \cos(x-\pi/4) \\ \sin(x-\pi/4) \end{array} \right\} \tag{81}$$

if the final term *kept*, odd or even, is a_k. If even, then multiply by $\cos(x-\pi/4)$; if odd, then $\sin(x-\pi/4)$.

Let's pause a moment. The algebra to show this is a bit finicky but not hard (the equation is, after all, linear). This end result is an extremely simple (and exact!) formula for Δ. The finite series $y_{N,M}$ is then the exact solution to

$$x^2 y'' + x y' + x y = \Delta \tag{82}$$

$$= \frac{(k+1/2)^2 a_k}{x^{k+1/2}} \cdot \left\{ \begin{array}{c} \cos(x-\frac{\pi}{4}) \\ \sin(x-\frac{\pi}{4}) \end{array} \right\} \tag{83}$$

and, provided x is large enough, this is only a small perturbation of Bessel's equation. In many modelling situations, such a small perturbation may be of direct physical significance, and we'd be done. Here, though, Bessel's equation typically arises as an intermediate step, after separation of variables, say. Hence one might be interested in the forward error. By the theory of Green's functions, we may express this as

$$J_0(x) - y_{N,M}(x) = \int_x^\infty K(x,\xi)\Delta(\xi)d\xi \tag{84}$$

for a suitable kernel $K(x,\xi)$. The obvious conclusion is that if Δ is small then so will $J_0(x) - y_{N,M}(x)$; but $K(x,\xi)$ will have some effect, possibly amplifying the effects of Δ, or perhaps even damping its effects. Hence, the connection is indirect.

To have an error in Δ of at most ε, we must have

$$\left(k+\frac{1}{2}\right)^2 \frac{|a_k|}{x^{k+1/2}} \leq \varepsilon \tag{85}$$

(remember, $x > 0$). This will happen only if

$$x \geq \left(\left(k+\frac{1}{2}\right)^2 \frac{|a_k|}{\varepsilon}\right)^{2/(2k+1)} \tag{86}$$

and this, for fixed k, goes to ∞ as $\varepsilon \to 0$. Alternatively, we may ask which k, for a fixed x, minimizes

$$\left(k+\frac{1}{2}\right)^2 \frac{|a_k|}{x^{k+1/2}} \tag{87}$$

and this answers the truncation question in a rational way. In this particular case, minimizing $\|\Delta\|$ doesn't necessarily minimize the forward error (although, it's close). For $x = 2.3$, for instance, the sequence $(k+1/2)^2|a_k|x^{-k-1/2}$ is (no $\sqrt{2/\pi}$)

$$\begin{array}{cccccccc} k & 0 & 1 & 2 & 3 & 4 & 5 \\ A_k & 0.165 & 0.081 & 0.055 & 0.049 & 0.054 & 0.070 \end{array} \tag{88}$$

The clear winner seems to be $k = 3$. This suggests that for $x = 2.3$, the best series to take is

$$y_3 = \left(\frac{2}{\pi x}\right)^{1/2}\left(\left(1-\frac{9}{128x^2}\right)\cos\left(x-\frac{\pi}{4}\right)+\left(\frac{1}{8x}-\frac{75}{1024x^3}\right)\sin\left(x-\frac{\pi}{4}\right)\right). \tag{89}$$

This gives $5.454 \cdot 10^{-2}$ for $x = 2.3$. But the cosine versus sine plays a role, here: $\cos(2.3-\pi/4) \doteq 0.056$ while $\sin(2.3-\pi/4) \doteq 0.998$, so we should have included this. When we do, the estimates for Δ_0, Δ_2 and Δ_4 are all significantly reduced—and this changes our selection, and makes $k = 4$ the right choice; $\Delta_6 > \Delta_4$ as well (either way). But the influence of the integral is mollifying. Comparing to a better answer (computers via the integral formula) 0.0555398, we see that the error is about $8.8 \cdot 10^{-4}$ whereas $((4+1/2)^2 a_4/2.3^{4+1/2})\cos(2.3-\pi/4)$ is $3.06 \cdot 10^{-3}$; hence the residual overestimates the error slightly.

How does the rule of thumb do? The first term that is neglected here is $(1/x)^{1/2}a_5x^{-5}\sin(x-\pi/4)$ which is $\sim 2.3 \cdot 10^{-3}$ apart from the $(2/\pi)^{1/2} = 0.797$ factor, so about $1.86 \cdot 10^{-3}$. The *next* term is, however, $(2/\pi x)^{1/2}a_6x^{-6}\cos(x-\pi/4) \doteq -1.14 \cdot 10^{-4}$ which is smaller yet, suggesting that we should keep the a_5 term. But we shouldn't. Stopping with a_4 gives a better answer, just as the residual suggests that it should.

We emphasize that this is only a slightly more rational rule of thumb, because minimizing $\|\Delta\|$ only minimizes a bound on the forward error, not the forward error itself. Still, we have not seen this discussed in the literature before. A final comment is that the defining equation and its scale, define also the scale for what's a "small" residual.

So, a justification for the "rule of thumb" would be as follows. In our general scheme,

$$Au_{n+1} = -[\varepsilon^{n+1}]\Delta_n \tag{90}$$

and thus, loosely speaking,

$$u_{n+1} \sim -A^{-1}\Delta_n + O(\varepsilon^{n+1})\,. \tag{91}$$

Thus, if we stop when u_{n+1} is smallest, this would tend to happen at the same integer n that Δ_n was smallest.

5 Initial-Value Problems

BEA has successfully been applied to the *numerical* solution of differential equations for a long time, now. Examples include the works of Enright since the 1980s, e.g., [12, 13], and indeed the Lanczos τ-method is yet older [19]. It was pointed out in [7] and [6] that BEA could be used for perturbation and other series solutions of differential equations, also. We here display several examples illustrating this fact. We use regular expansion, matched asymptotic expansions, the renormalization group method, and the method of multiple scales.

5.1 Duffing's Equation

This proposed way of interpreting solutions obtained by perturbation methods has interesting advantages for the analysis of series solutions to differential equations. Consider for example an unforced weakly nonlinear Duffing oscillator, which we take from [3]:

$$y'' + y + \varepsilon y^3 = 0 \tag{92}$$

with initial conditions $y(0) = 1$ and $y'(0) = 0$. As usual, we assume that $0 < \varepsilon \ll 1$. Our discussion of this example does not provide a new method of solving this problem, but instead it improves the interpretation of the quality of solutions obtained by various methods.

5.1.1 Regular Expansion

The classical perturbation analysis supposes that the solution to this equation can be written as the power series

$$y(t) = y_0(t) + y_1(t)\varepsilon + y_2(t)\varepsilon^2 + y_3(t)\varepsilon^3 + \cdots . \tag{93}$$

Substituting this series in Eq. (92) and solving the equations obtained by equating to zero the coefficients of powers of ε in the residual, we find $y_0(t)$ and $y_1(t)$ and we thus have the solution

$$z_1(t) = \cos(t) + \varepsilon\left(\frac{1}{32}\cos(3t) - \frac{1}{32}\cos(t) - \frac{3}{8}t\sin(t)\right) . \tag{94}$$

The difficulty with this solution is typically characterized in one of two ways. Physically, the secular term $t \sin t$ shows that our simple perturbative method has failed since the energy conservation prohibits unbounded solutions. Mathematically, the secular term $t \sin t$ shows that our method has failed since the periodicity of the solution contradicts the existence of secular terms.

Both these characterizations are correct, but require foreknowledge of what is physically meaningful or of whether the solutions are bounded. In contrast, interpreting (94) from the backward error viewpoint is much simpler. To compute the residual, we simply substitute z_2 in Eq. (92), that is, the residual is defined by

$$\Delta_1(t) = z_1'' + z_1 + \varepsilon z_1^3 . \tag{95}$$

For the first-order solution of Eq. (94), the residual is

$$\Delta_1(t) = \left(-\tfrac{3}{64}\cos(t) + \tfrac{3}{128}\cos(5t) + \tfrac{3}{128}\cos(3t) - \tfrac{9}{32}t\sin(t) - \tfrac{9}{32}t\sin(3t)\right)\varepsilon^2 + O(\varepsilon^3). \tag{96}$$

$\Delta_1(t)$ is exactly computable. We don't print it all here because it's too ugly, but in Fig. 1, we see that the complete residual grows rapidly. This is due to the secular term $-\frac{9}{32}t(\sin(t) - \sin(3t))$ of Eq. (96). Thus we come to the conclusion that the secular term contained in the first-order solution obtained in Eq. (94) invalidate it, but this time we do not need to know in advance what to physically expect or to prove that the solution is bounded. This is a slight but sometimes useful gain in simplicity.[2]

A simple Maple code makes it possible to easily obtain higher-order solutions:

```
# Regular Expansion for Duffing's Equation
# We choose initial conditions y(0)=1 and y'(0)=0 so
```

[2]In addition, this method makes it easy to find mistakes of various kinds. For instance, we uncovered a typo in the 1978 edition of [3] by computing the residual. That typo does not seem to be in the later editions, so it's likely that the authors found and fixed it themselves, as well.

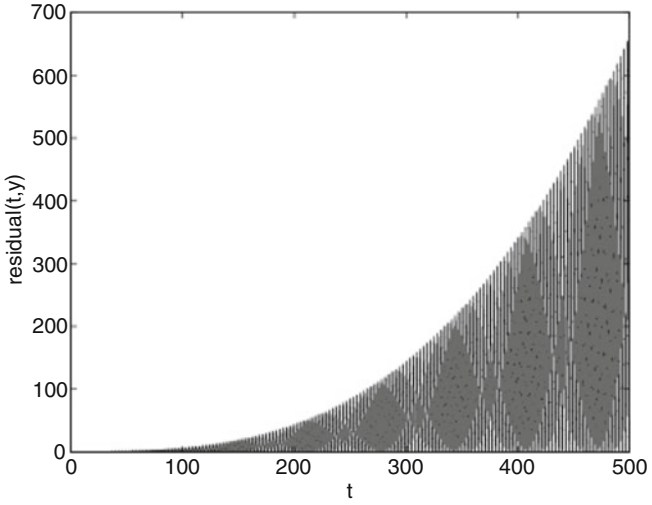


Fig. 1 Absolute Residual for the first-order classical perturbative solution of the unforced weakly damped Duffing equation with $\varepsilon = 0.1$

```
# that y(t)=cos(t) to O(e).
macro(e=varepsilon);
N := 3;
Order := N+1;
z := add(y[k](t)*e^k, k = 0 .. N);
DE := y -> diff(y, t, t)+y+e*y^3;
des := series( DE(z), e);
dos := dsolve({coeff(des, e, 0), y[0](0) = 1,
                (D(y[0]))(0) = 0}, y[0](t) );
assign(dos);
for k to N do
  tmp:=dsolve({coeff(des,e,k),y[k](0)=0,
                (D(y[k]))(0)=0},y[k](t));
  assign(tmp);
end do:
Delta := DE(z):
ResidualSeries := map(combine,
                      series(Delta, e, Order+3), trig);
```

Experiments with this code suggests the conjecture that $\Delta_n = O(t^n \varepsilon^{n+1})$. For this to be small, we must have $\varepsilon t = o(1)$ or $t < O(1/\varepsilon)$.

5.1.2 Lindstedt's Method

The failure to obtain an accurate solution on unbounded time intervals by means of the classical perturbation method suggests that another method that eliminates the secular terms will be preferable. A natural choice is Lindstedt's method, which rescales the time variable t in order to cancel the secular terms. The idea is that if we use a rescaling $\tau = \omega t$ of the time variable and chose ω wisely the secular terms from the classical perturbation method will cancel each other out.[3] Applying this transformation, Eq. (92) becomes

$$\omega^2 y''(\tau) + y(\tau) + \varepsilon y^3(\tau) \qquad y(0) = 1, \quad y'(0) = 0\,. \tag{97}$$

In addition to writing the solution as a truncated series

$$z_1(\tau) = y_0(\tau) + y_1(\tau)\varepsilon \tag{98}$$

we expand the scaling factor as a truncated power series in ε:

$$\omega = 1 + \omega_1 \varepsilon\,. \tag{99}$$

Substituting (98) and (99) back in Eq. (97) to obtain the residual and setting the terms of the residual to zero in sequence, we find the equations

$$y_0'' + y_0 = 0\,, \tag{100}$$

so that $y_0 = \cos(\tau)$, and

$$y_1'' + y_1 = -y_0^3 - 2\omega_1 y_0'' \tag{101}$$

subject to the same initial conditions, $y_0(0) = 1$, $y_0'(0) = 0$, $y_1(0) = 0$, and $y_1'(0) = 0$. By solving this last equation, we find

$$y_1(\tau) = \frac{31}{32}\cos(\tau) + \frac{1}{32}\cos(3\tau) - \frac{3}{8}\tau\sin(\tau) + \omega_1 \tau \sin(\tau)\,. \tag{102}$$

So, we only need to choose $\omega_1 = 3/8$ to cancel out the secular terms containing $\tau\sin(\tau)$. Finally, we simply write the solution $y(t)$ by taking the first two terms of $y(\tau)$ and plug in $\tau = (1 + 3\varepsilon/8)t$:

$$z_1(t) = \cos\tau + \varepsilon\left(\frac{31}{32}\cos\tau + \frac{1}{32}\cos\tau\right) \tag{103}$$

[3] Interpret this as: we choose ω to keep the residual small over as long a time-interval as possible.

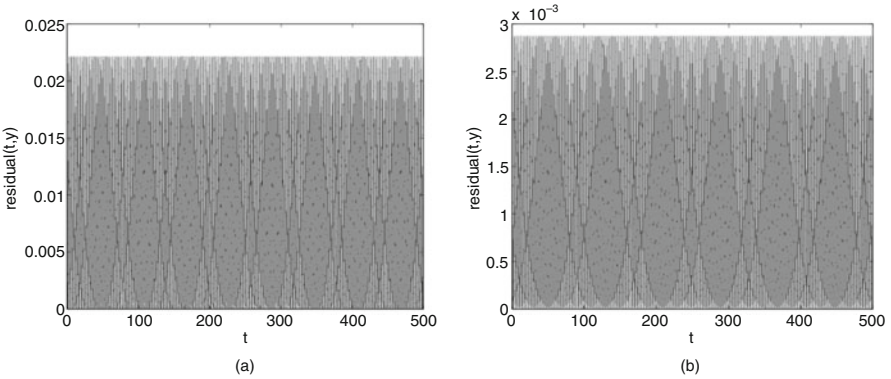


Fig. 2 Absolute Residual for the Lindstedt solutions of the unforced weakly damped Duffing equation with $\varepsilon = 0.1$. (**a**) First-Order. (**b**) Second-Order

This truncated power series can be substituted back in the left-hand side of Eq. (92) to obtain an expression for the residual:

$$\Delta_1(t) = \left(\frac{171}{128}\cos(t) + \frac{3}{128}\cos(5t) + \frac{9}{16}\cos(3t)\right)\varepsilon^2 + O\left(\varepsilon^3\right) \tag{104}$$

See Fig. 2a. We then do the same with the second term ω_2. The following Maple code has been tested up to order 12:

```
# Elimination of Secular Terms in the Solution of the
# Duffing Equation with the Poincare-Lindstedt method.
restart;
macro(e=varepsilon);
N := 4;
Order := N+1;
z := add(y[k](tau)*e^k, k = 0..N);
omega := 1+add(a[k]*e^k, k = 1..N);
DE := y -> omega^2*(diff(y, tau, tau))+y+e*y^3;
des := series(DE(z), e);
dos := dsolve({coeff(des,e,0),y[0](0)=1,
                (D(y[0]))(0)=0},y[0](tau));
assign(dos);
for k to N do
      tmp := convert(combine(coeff(des,e,k),trig),exp);
      UZ := eval(tmp, [exp(I*tau)=Z,exp(-I*tau)=1/Z]);
      ah := coeff(UZ, Z, 1);
      antisecular := solve(ah = 0, a[k]);
      if {antisecular} <> {} then
            a[k] := antisecular;
```

```
        end if;
        tmp := dsolve({evalc(tmp),y[k](0)=0,
                                (D(y[k]))(0)=0},y[k](tau));
        assign(tmp);
end do;
Delta := DE(z);
Sdelta := map(simplify, series(Delta, e, Order+4));
map(combine, Sdelta, trig);
```

The significance of this is as follows: The normal presentation of the method first requires a proof (an independent proof) that the reference solution is bounded and therefore the secular term $\varepsilon t \sin t$ in the classical solution is spurious. *But* the residual analysis needs no such proof. It says directly that the classical solution solves neither

$$f(t, y, y', y'') = 0 \tag{105}$$

nor $f + \Delta f = 0$ for uniformly small Δ but rather that the residual *departs* from 0 and is *not* uniformly small whereas the residual for the Lindstedt solution *is* uniformly small.

5.2 Morrison's Counterexample

In [24, pp. 192–193], we find a discussion of the equation

$$y'' + y + \varepsilon(y')^3 + 3\varepsilon^2(y') = 0\,. \tag{106}$$

O'Malley attributed the equation to [21]. The equation is one that is supposed to illustrate a difficulty with the (very popular and effective) method of multiple scales. We give a relatively full treatment here because a residual-based approach shows that the method of multiple scales, applied somewhat artfully, can be quite successful and moreover we can demonstrate a posteriori that the method was successful. The solution sketched in [24] uses the complex exponential format, which one of us used to good effect in his PhD, but in this case the real trigonometric form leads to slightly simpler formulæ. We are very much indebted to our colleague, Professor Pei Yu at Western, for his careful solution, which we follow and analyze here.[4]

The first thing to note is that we will use three time scales, $T_0 = t$, $T_1 = \varepsilon t$, and $T_2 = \varepsilon^2 t$ because the DE contains an ε^2 term, which will prove to be important. Then the multiple scales formalism gives

[4]We had asked him to solve this problem using one of his many computer algebra programs; instead, he presented us with an elegant handwritten solution.

$$\frac{d}{dt} = \frac{\partial}{\partial T_0} + \varepsilon \frac{\partial}{\partial T_1} + \varepsilon^2 \frac{\partial}{\partial T_2} \tag{107}$$

This formalism gives most students some pause, at first: replace an ordinary derivative by a sum of partial derivatives using the chain rule? What could this mean? But soon the student, emboldened by success on simple problems, gets used to the idea and eventually the conceptual headaches are forgotten.[5] But sometimes they return, as with this example.

To proceed, we take

$$y = y_0 + \varepsilon y_1 + \varepsilon^2 y_2 + O(\varepsilon^3) \tag{108}$$

and equate to zero like powers of ε in the residual. The expansion of d^2y/dt^2 is straightforward:

$$\left(\frac{\partial}{\partial T_0} + \varepsilon \frac{\partial}{\partial T_1} + \varepsilon^2 \frac{\partial}{\partial T_2}\right)^2 (y_0 + \varepsilon y_1 + \varepsilon^2 y_2) =$$
$$\frac{\partial^2 y_0}{\partial T_0^2} + \varepsilon \left(\frac{\partial^2 y_1}{\partial T_0^2} + 2\frac{\partial^2 y_0}{\partial T_0 \partial T_1}\right) + \varepsilon^2 \left(\frac{\partial^2 y_2}{\partial T_0^2} + 2\frac{\partial^2 y_1}{\partial T_0 \partial T_1} + \frac{\partial^2 y_0}{\partial T_1^2} + 2\frac{\partial^2 y_0}{\partial T_0 \partial T_1}\right) \tag{109}$$

For completeness we include the other necessary terms, even though this construction may be familiar to the reader. We have

$$\varepsilon \left(\frac{dy}{dt}\right)^3 = \varepsilon \left(\left(\frac{\partial}{\partial T_0} + \varepsilon \frac{\partial}{\partial T_1}\right)(y_0 + \varepsilon y_1)\right)^3 \tag{110}$$
$$= \varepsilon \left(\frac{\partial y_0}{\partial T_0}\right)^3 + 3\varepsilon^2 \left(\frac{\partial y_0}{\partial T_0}\right)^2 \left(\frac{\partial y_0}{\partial T_1} + \frac{\partial y_1}{\partial T_0}\right) + \cdots, \tag{111}$$

and $y = y_0 + \varepsilon y_1 + \varepsilon^2 y_2$ is straightforward, and also

$$3\varepsilon^2 \left(\left(\frac{\partial}{\partial T_0} + \cdots\right)(y_0 + \cdots)\right) = 3\varepsilon^2 \frac{\partial y_0}{\partial T_0} + \cdots \tag{112}$$

[5] This can be made to make sense, after the fact. We imagine $F(T_1, T_2, T_3)$ describing the problem, and $d/dt = \partial F/\partial T_1 \partial T_1/\partial t + \partial F/\partial T_2 \partial T_2/\partial t + \partial F/\partial T_3 \partial T_3/\partial t$ which gives $d/dt = \partial F/\partial T_1 + \varepsilon \partial F/\partial T_2 + \varepsilon^2 \partial F/\partial T_3$ if $T_1 = t$, $T_2 = \varepsilon t$ and $T_3 = \varepsilon^2 t$.

is at this order likewise straightforward. At $O(\varepsilon^0)$ the residual is

$$\frac{\partial^2 y_0}{\partial T_0^2} + y_0 = 0 \tag{113}$$

and without loss of generality we take as solution

$$y_0 = a(T_1, T_2)\cos(T_0 + \varphi(T_1, T_2)) \tag{114}$$

by shifting the origin to a local maximum when $T_0 = 0$. For notational simplicity put $\theta = T_0 + \varphi(T_1, T_2)$. At $O(\varepsilon^1)$ the equation is

$$\frac{\partial^2 y_1}{\partial T_0^2} + y_1 = -\left(\frac{\partial y_0}{\partial T_0}\right)^3 - 2\frac{\partial^2 y_0}{\partial T_0 \partial T_1} \tag{115}$$

where the first term on the right comes from the $\varepsilon \dot{y}^3$ term whilst the second comes from the multiple scales formalism. Using $\sin^3\theta = {}^3\!/\!_4 \sin\theta - {}^1\!/\!_4 \sin 3\theta$, this gives

$$\frac{\partial^2 y_1}{\partial T_0^2} + y_1 = \left(2\frac{\partial a}{\partial T_1} + \frac{3}{4}a^3\right)\sin\theta + 2a\frac{\partial \varphi}{\partial T_1}\cos\theta - \frac{a^3}{4}\sin 3\theta \tag{116}$$

and to suppress the resonance that would generate secular terms we put

$$\frac{\partial a}{\partial T_1} = -\frac{3}{8}a^3 \quad \text{and} \qquad \frac{\partial \varphi}{\partial T_1} = 0\,. \tag{117}$$

Then $y_1 = \frac{a^3}{32}\sin 3\theta$ solves this equation and has $y_1(0) = 0$, which does not disturb the initial condition $y_0(0) = a_0$, although since ${}^{dy_1}\!/\!_{dT_0} = {}^{3a^2}\!/\!_{32}\cos 3\theta$ the derivative of $y_0 + \varepsilon y_1$ will differ by $O(\varepsilon)$ from zero at $T_0 = 0$. This does not matter and we may adjust this by choice of initial conditions for φ, later.

The $O(\varepsilon^2)$ term is somewhat finicky, being

$$\frac{\partial^2 y_2}{\partial T_0^2} + y_2 = -2\frac{\partial^2 y_0}{\partial T_0 \partial T_2} - 2\frac{\partial^2 y_1}{\partial T_0 \partial T_1} - 3\left(\frac{\partial y_0}{\partial T_0}\right)^2\left(\frac{\partial y_0}{\partial T_1} + \frac{\partial y_1}{\partial T_0}\right) - \frac{\partial^2 y_0}{\partial T_1^2} - 3\frac{\partial y_0}{\partial T_0} \tag{118}$$

where the last term came from $3(\dot{y})\varepsilon^2$. Proceeding as before, and using $\partial\varphi/\partial T_1 = 0$ and ${}^{\partial a}\!/\!_{\partial T_1} = -{}^3\!/\!_8\, a^3$ as well as some other trigonometric identities, we find the right-hand side can be written as

$$\left(2\frac{\partial a}{\partial T_2} + 3a\right)\sin\theta + \left(2a\frac{\partial \varphi}{\partial T_2} - \frac{9}{128}a^5\right)\cos\theta - \frac{27}{1024}a^5\cos 3\theta + \frac{9}{128}a^5\cos 5\theta\,. \tag{119}$$

Again setting the coefficients of $\sin\theta$ and $\cos\theta$ to zero to prevent resonance we have

$$\frac{\partial a}{\partial T_2} = -\frac{3}{2}a \tag{120}$$

and

$$\frac{\partial \varphi}{\partial T_2} = \frac{9}{256}a^4 \qquad (a \neq 0). \tag{121}$$

This leaves

$$y_2 = \frac{27}{1024}a^5 \cos 3\theta - \frac{3a^5}{1024}\cos 5\theta \tag{122}$$

again setting the homogeneous part to zero.

Now comes a bit of multiple scales magic: instead of solving Eqs. (117) and (120) in sequence, as would be usual, we write

$$\begin{aligned}\frac{da}{dt} &= \frac{\partial a}{\partial T_0} + \varepsilon\frac{\partial a}{\partial T_1} + \varepsilon^2\frac{\partial a}{\partial T_2} = 0 + \varepsilon\left(-\frac{3}{8}a^3\right) + \varepsilon^2\left(-\frac{3}{2}a\right) \\ &= -\frac{3}{8}\varepsilon a(a^2 + 4\varepsilon)\,.\end{aligned} \tag{123}$$

Using $a = 2R$ this is equation (6.50) in [24]. Similarly

$$\frac{d\varphi}{dt} = \varepsilon\frac{\partial \varphi}{\partial T_1} + \varepsilon^2\frac{\partial \varphi}{\partial T_2} = 0 + \varepsilon^2\frac{9}{256}a^4 \tag{124}$$

and once a has been identified, φ can be found by quadrature. Solving (123) and (124) by Maple,

$$a = \frac{\sqrt{\varepsilon}a_0}{\sqrt{\varepsilon e^{3\varepsilon^2 t} + \dfrac{a_0^2}{4}(e^{3\varepsilon^2 t} - 1)}} = 2\frac{\sqrt{\varepsilon}a_0}{\sqrt{u}} \tag{125}$$

and

$$\varphi = -\frac{3}{16}\varepsilon^2 \ln u + \frac{9}{16}\varepsilon^4 t - \frac{3}{16}\frac{\varepsilon^2 a_0^2}{u} \tag{126}$$

where $u = 4\varepsilon e^{3\varepsilon^2 t} + a_0^2(e^{3\varepsilon^2 t} - 1)$. The residual is (again by Maple)

$$\varepsilon^3\left(\frac{9}{16}a_0^3\cos 3t + a_0^7\left(-\frac{351}{4096}\sin t - \frac{9}{512}\sin 7t + \frac{333}{4096}\sin 3t + \frac{459}{4096}\sin 5t\right)\right) + O(\varepsilon^4) \tag{127}$$

and there is no secularity visible in this term.

It is important to note that the construction of Eq. (123) for $a(t)$ required both $\partial a/\partial T_1$ and $\partial a/\partial T_2$. Either one alone gives misleading or inconsistent answers. While it may be obvious to an expert that both terms must be used at once, the situation is somewhat unusual and a novice or casual user of perturbation methods may well wish reassurance. (We did!) Computing (and plotting) the residual $\Delta = \ddot{z} + z + \varepsilon(\dot{z})^3 + 3\varepsilon^2\dot{z}$ does just that (see Fig. 3). It is simple to verify that, say, for $\varepsilon = 1/100$, $|\Delta| < \varepsilon^3 a$ on $0 < t < 10^5\pi$. Notice that $a \sim O(e^{-3/2\,\varepsilon^2 t})$ and $e^{-3/2\cdot 10^{-4}\cdot 10^5\cdot\pi} = e^{-15\pi} \doteq 10^{-15}$ by the end of this range. The method of multiple scales has thus produced z, the exact solution of an equation uniformly and relatively near to the original equation. In trigonometric form,

$$\begin{aligned} z = \; & a\cos(t+\varphi) + \varepsilon\frac{a^3}{32}\cos(3(t+\varphi)) \\ & + \varepsilon^2\left(\frac{27}{1024}a^5\cos(3(t+\varphi)) - \frac{3}{1024}a^5\cos^5((5(t+\varphi))\right) \end{aligned} \tag{128}$$

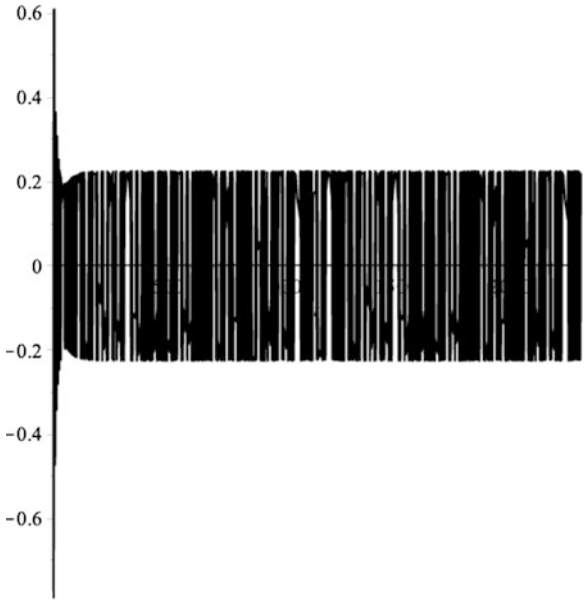


Fig. 3 The residual $|\Delta_3|$ divided by $\varepsilon^3 a$, with $\varepsilon = 0.1$, where $a = O(e^{-3/2\,\varepsilon^2 t})$, on $0 \le t \le {}^{10\ln(10)}/_{\varepsilon^2}$ (at which point $a = 10^{-15}$). We see that $|{}^{\Delta_3}/_{\varepsilon^3 a}| < 1$ on this entire interval

and a and φ are as in Eqs. (123) and (124). Note that φ asymptotically approaches zero. Note that the trigonometric solution we have demonstrated here to be correct, which was derived for us by our colleague Pei Yu, appears to differ from that given in [24], which is

$$y = Ae^{it} + \varepsilon Be^{3it} + \varepsilon^2 Ce^{5it} + \cdots \tag{129}$$

where (with $\tau = \varepsilon t$)

$$C \sim \frac{3}{64}A^5 + \cdots \qquad \text{and} \qquad B \sim -\frac{A^3}{8}(i + \frac{45}{8}\varepsilon|A|^2 + \cdots) \tag{130}$$

and, if $A = Re^{i\varphi}$,

$$\frac{dR}{d\tau} = -\frac{3}{2}(R^3 + \varepsilon R + \cdots) \qquad \text{and} \qquad \frac{d\varphi}{d\tau} = -\frac{3}{2}R^2(1 + \frac{3\varepsilon}{8}R^2 + \cdots) \tag{131}$$

Of course with the trigonometric form $y = a\cos(t + \varphi)$, the equivalent complex form is

$$y = a\left(\frac{e^{it+i\varphi} + e^{-it-i\varphi}}{2}\right) = \frac{a}{2}e^{i\varphi}e^{it} + c.c. \tag{132}$$

and so $R = a/2$. As expected, equation (6.50) in [24] becomes

$$\frac{da}{d\tau}\left(\frac{a}{2}\right) = -\frac{3}{2}\frac{a}{2}\left(\frac{a^2}{4} + \varepsilon\right) \tag{133}$$

or, alternatively,

$$\frac{da}{d\tau} = -\frac{3}{8}\varepsilon a(a^2 + 4\varepsilon) \tag{134}$$

which agrees with that computed for us by Pei Yu. However, O'Malley's equation (6.48) gives

$$C \cdot e^{i\cdot 5t} = \frac{3}{64}A^5 e^{i5t} = \frac{3}{64}R^5 e^{i5\theta} = \frac{3}{2048}a^5 e^{i5\theta}\,, \tag{135}$$

so that

$$Ce^{i5t} + c.c = \frac{3}{1024}a^5\cos 5\theta\,, \tag{136}$$

whereas Pei Yu has $-3/1024$. As demonstrated by the residual in Fig. 3, Pei Yu is correct. Well, sign errors are trivial enough.

More differences occur for B, however. The $-A^3/8\ ie^{3it}$ term becomes $a^3/32 \cos 3\theta$, as expected, but $-45/64A^3 \cdot |A|^2 e^{3it} + c.c.$ becomes $-45/32a^5/32 \cos 3\theta = -45/1024a^5 \cos 3\theta$, not $27/1024a^5 \cos 3\theta$. Thus we believe there has been an arithmetic error in [24]. This is also present in [25]. Similarly, we believe the $d\varphi/dt$ equation there is wrong.

Arithmetic errors in perturbation solutions are, obviously, a constant hazard even for experts. We do not point out this error (or the other errors highlighted in this paper) in a spirit of glee—goodness knows we've made our own share. No, the reason we do so is to emphasize the value of a separate, independent check using the residual. Because we have done so here, we are certain that Eq. (128) is correct: it produces a residual that is uniformly $O(\varepsilon^3)$ for bounded time, and which is $O(\varepsilon^{9/2} e^{-3/2\,\varepsilon^2 t})$ as $t \to \infty$. (We do not know why there is extra accuracy for large times).

Finally, we remark that the difficulty this example presents for the method of multiple scales is that Eq. (123) cannot be solved itself by perturbation methods (or, at least, we couldn't do it). One has to use all three terms at once; the fact that this works is amply demonstrated afterwards. Indeed the whole multiple scales procedure based on Eq. (107) is really very strange when you think about it, but it can be justified afterwards. It really doesn't matter how we find Eq. (128). Once we have done so, verifying that it is the exact solution of a small perturbation of the original equation is quite straightforward. The implementation is described in the following Maple code:

```
macro(e=varepsilon);
r := e;
de := u -> (diff(u, t, t)+u
      +r*(diff(u, t))^3+3*r^2*(diff(u, t)));
U := -a0^2+4*exp(3*e^2*t)*e+exp(3*e^2*t)*a0^2;
a := 2*sqrt(r)*a0/sqrt(U);
phi := -(3/16)*e^2*ln(U)+(9/16)*e^4*t-(3/16)*e^2*a0^2/U;
z := a*cos(t+phi)+(1/32)*r*a^3*sin(3*t+3*phi)
     +r^2*((27/1024)*a^5*cos(3*(t+phi))
     -(3/1024)*a^5*cos(5*(t+phi))):
resid := de(z):
zer := MultiSeries[series](resid, r, 4):
map(combine, zer, trig);
eps := 1/10;
plot(eval(resid/(a*r^3), [a0 = 1.0, r = eps]),
     t = 0 .. 10*ln(10)/eps^2, colour = BLACK);
```

5.3 The Lengthening Pendulum

As an interesting example with a genuine secular term, [4] discusses the lengthening pendulum. There, Boas solves the linearized equation exactly in terms of Bessel functions. We use the model here as an example of a perturbation solution in a physical context. The original Lagrangian leads to

$$\frac{d}{dt}\left(m\ell^2\frac{d\theta}{dt}\right) + mg\ell\sin\theta = 0 \tag{137}$$

(having already neglected any system damping). The length of the pendulum at time t is modelled as $\ell = \ell_0 + vt$, and implicitly v is small compared to the oscillatory speed $d\theta/dt$ (else why would it be a pendulum at all?). The presence of $\sin\theta$ makes this a nonlinear problem; when $v = 0$ there is an analytic solution using elliptic functions [20, chap. 4].

We *could* do a perturbation solution about that analytic solution; indeed there is computer algebra code to do so automatically [26]. For the purpose of this illustration, however, we make the same small-amplitude linerization that Boas did and replace $\sin\theta$ by θ. Dividing the resulting equation by ℓ_0, putting $\varepsilon = v/\ell_0\omega$ with $\omega = \sqrt{g/\ell_0}$ and rescaling time to $\tau = \omega t$, we get

$$(1+\varepsilon\tau)\frac{d^2\theta}{d\tau^2} + 2\varepsilon\frac{d\theta}{d\tau} + \theta = 0\,. \tag{138}$$

This supposes, of course, that the pin holding the base of the pendulum is held perfectly still (and is frictionless besides).

Computing a regular perturbation approximation

$$z_{\text{reg}} = \sum_{k=0}^{N}\theta_k(\tau)\varepsilon^k \tag{139}$$

is straightforward, for any reasonable N, by using computer algebra. For instance, with $N = 1$ we have

$$z_{\text{reg}} = \cos\tau + \varepsilon\left(\frac{3}{4}\sin\tau + \frac{\tau^2}{4}\sin\tau - \frac{3}{4}\tau\cos\tau\right)\,. \tag{140}$$

This has residual

$$\Delta_{\text{reg}} = (1+\varepsilon\tau)z''_{\text{reg}} + 2\varepsilon z'_{\text{reg}} + z_{\text{reg}} \tag{141}$$

$$= -\frac{\varepsilon^2}{4}\left(\tau^3\sin\tau - 9\tau^2\cos\tau - 15\tau\sin\tau\right) \tag{142}$$

also computed straightforwardly with computer algebra. By experiment with various N we find that the residuals are always of $O(\varepsilon^{N+1})$ but contain powers of τ, as high as τ^{2N+1}. This naturally raises the question of just when this can be considered "small." We thus have the *exact* solution of

$$(1+\varepsilon\tau)\frac{d^2\theta}{d\tau^2}+2\varepsilon\frac{d\theta}{d\tau}+\theta=\Delta_{\text{reg}}(\tau)=P(\varepsilon^{N+1}\tau^{2N+1}) \tag{143}$$

and it seems clear that if $\varepsilon^{N+1}\tau^{2N+1}$ is to be considered small it should at least be smaller than $\varepsilon\tau$, which appear on the left hand side of the equation. [$d^2/d\tau^2$ is $-\cos\tau$ to leading order, so this is periodically $O(1)$.] This means $\varepsilon^N\tau^{2N}$ should be smaller than 1, which forces $\tau\le T$ where $T=O(\varepsilon^{-q})$ with $q<\frac{1}{2}$. That is, this regular perturbation solution is valid only on a limited range of τ, namely, $\tau=O(\varepsilon^{-1/2})$.

Of course, the original equation contains a term $\varepsilon\tau$, and this itself is small only if $\tau\le T_{\max}$ with $T_{\max}=O(\varepsilon^{-1+\delta})$ for $\delta>0$. Notice that we have discovered this limitation of the regular perturbation solution without reference to the 'exact' Bessel function solution of this linearized equation. Notice also that Δ_{reg} can be interpreted as a small forcing term; a vibration of the pin holding the pendulum, say. Knowing that, say, such physical vibrations, perhaps caused by trucks driving past the laboratory holding the pendulum, are bounded in size by a certain amount, can help to decide what N to take, and over which τ-interval the resulting solution is valid.

Of course, one might be interested in the forward error $\theta-z_{\text{reg}}$; but then one should be interested in the forward errors caused by neglecting physical vibrations (e.g. of trucks passing by) and the same theory—what a numerical analyst calls a condition number—can be used for both.

But before we pursue that farther, let us first try to improve the perturbation solution. The method of multiple scales, or equivalent but easier in this case the renormalization group method [18] which consists for a linear problem of taking the regular perturbation solution and replacing $\cos\tau$ by $(e^{i\tau}+e^{-i\tau})/2$ and $\sin\tau$ by $(e^{i\tau}-e^{-i\tau})/2i$, gathering up the result and writing it as $1/2\,A(\tau;e)e^{i\tau}+1/2\,\bar{A}(\tau;\varepsilon)e^{-i\tau}$. One then writes $A(\tau;\varepsilon)=e^{L(\tau;\varepsilon)}+O(\varepsilon^{N+1})$ (that is, taking the logarithm of the ε-series for $A(\tau;\varepsilon)=A_0(\tau)+\varepsilon A_1(\tau)+\cdots+\varepsilon^N A_N(\tau)+O(\varepsilon^{N+1})$, a straightforward exercise (especially in a computer algebra system) and then (if one likes) rewriting $1/2\;e^{L(\tau;\varepsilon)+i\tau}+$ c.c. in real trigonometric form again, gives an excellent result. If $N=1$, we get

$$\tilde{z}_{\text{renorm}}=e^{-3/4\,\varepsilon\tau}\cos\left(\frac{3}{4}\varepsilon+\tau-\varepsilon\frac{\tau^2}{4}\right) \tag{144}$$

which contains an irrelevant phase change $\frac{3}{4}\varepsilon$ which we remove here as a distraction to get

$$z_{\text{renorm}}=e^{-3/4\,\varepsilon\tau}\cos\left(\tau-\varepsilon\frac{\tau^2}{4}\right). \tag{145}$$

This has residual:

$$\begin{aligned}\Delta_{\text{renorm}} &= (1+\varepsilon\tau)\frac{d^2 z_{\text{renorm}}}{d\tau^2} + 2\varepsilon\frac{dz_{\text{renorm}}}{d\tau} + z_{\text{renorm}} \\ &= \varepsilon^2 e^{-\frac{3}{4}\varepsilon\tau}\left((\frac{3}{4}\tau^2 - \frac{15}{16})\cos(\tau - \varepsilon\frac{\tau^2}{4}) - \frac{9}{4}\tau\sin(\tau - \varepsilon\frac{t^2}{4})\right) \\ &\quad + O(\varepsilon^3\tau^3 e^{-\frac{3}{4}\varepsilon\tau})\,. \end{aligned} \tag{146}$$

By inspection, we see that this is superior in several ways to the residual from the regular perturbation method. First, it contains the damping term $e^{-3/4\,\varepsilon\tau}$ just as the computed solution does; this residual will be small compared even to the decaying solution. Second, at order N it contains only τ^{N+1} as its highest power of ε, not τ^{2N+1}. This will be small compared to $\varepsilon\tau$ for times $\tau < T$ with $T = O(\varepsilon^{-1+\delta})$ for *any* $\delta > 0$; that is, this perturbation solution will provide a good solution so long as its fundamental assumption, that the $\varepsilon\tau$ term in the original equation, can be considered 'small', is good.

Note that again the quality of this perturbation solution has been judged without reference to the exact solution, and quite independently of whatever assumptions are usually made to argue for multiple scales solutions (such as boundedness of θ) or the renormalization group method. Thus, we conclude that the renormalization group method gives a superior solution in this case, and this judgement was made possible by computing the residual. We have used the following Maple implementation:

```
macro(e = varepsilon);
de := y -> (1+e*t)*(diff(y, t, t))+2*e*(diff(y, t))+y;
z := cos(t);
N := 1;
Order := N+1;
for i to N do
        zt := z+e^i*y[i](t);
        res := series(de(zt), e, i+1);
        eqs := coeff(res, e, i);
        yi := dsolve({eqs, y[i](0) = 0,
                (D(y[i]))(0) = 0}, y[i](t));
        z := eval(zt, yi);
end do:
res := de(z);
expform := convert(z, exp);
expform := collect(expform,[exp(I*t),exp(-I*t)],factor);
zp := coeff(expform, exp(I*t));
lg := convert(series(ln(series(zp+O(e^Order), e)), e),
            polynom);
lg := collect(lg, e, factor);
zrg := exp(lg)*exp(I*t);
```

```
zrg := zrg+evalc(conjugate(zrg));
zrg := combine(evalc(zrg), trig);
zrg := simplify(zrg);
zrg := exp(-(3/4)*e*t)*cos(t-(1/4)*e*t^2);
resrg := collect(de(zrg), e,
                 t -> combine(simplify(t), trig));
tiny := 1/1000;
plot(eval([z, zrg], e = tiny), t = 0 .. 1/tiny^(3/4),
     colour = BLACK, linestyle = [2, 1] );
plot(eval(res, e = tiny), t = 1 .. 2500,
     colour = BLACK, linestyle = 2);
plot(eval(resrg/(tiny*t), e = tiny), t = 1 .. 2500,
     colour = BLACK, style = POINT, numpoints=2016,
     symbolsize=1 );
```

See Fig. 4.

Note that this renormalized residual contains terms of the form $(\varepsilon\tau)^k e^{-3/4\,\varepsilon\tau}$. No matter what order we compute to, these have maxima $O(1)$ when $\tau = O(1/\varepsilon)$, but as noted previously the fundamental assumption of perturbation has been violated by that large a τ.

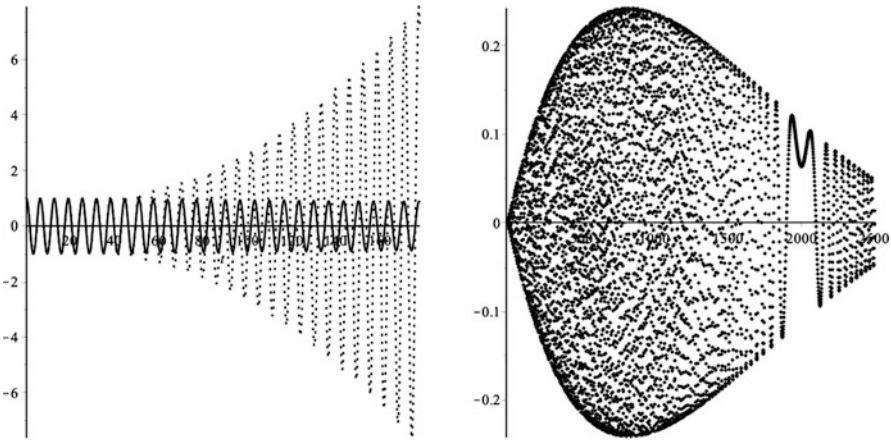


Fig. 4 On the left, solutions to the lengthening pendulum equation (the renormalized solution is the solid line). On the right, residual of the renormalized solution, which is orders of magnitudes smaller than that of the regular expansion

5.4 Optimal Backward Error Again

Now, one further refinement is possible. We may look for an $O(\varepsilon^2)$ perturbation of the lengthening of the pendulum, that explains part of this computed residual! That is, we look for $p(t)$, say, so that

$$\Delta_2 := (1 + \varepsilon\tau + \varepsilon p(\tau))z''_{\text{renorm}} + 2(\varepsilon + \varepsilon^2 p'(\tau))z'_{\text{renorm}} + z_{\text{renorm}} \tag{147}$$

has only *smaller* terms in it than Δ_{renorm}. Note the correlated changes, $\varepsilon^2 p(\tau)$ and $\varepsilon^2 p'(\tau)$.

At this point, we don't know if this is possible or useful, but it's a good thing to try. In numerical analysis terms, we are trying to find a structured backward error for this computed solution.

The procedure for identifying $p(\tau)$ in Eq. (147) is straightforward. We put $p(\tau) = a_0 + a_1\tau + a_2\tau^2$ with unknown coefficients, compute Δ_2, and try to choose a_0, a_1, and a_2 in order to make as many coefficients of powers of ε in Δ_2 to be zero as we can. When we do this, we find that

$$p = -\frac{15}{16} + \frac{3}{4}\tau^2 \tag{148}$$

makes

$$\Delta_{\text{mod}} = \left(1+\varepsilon\tau+\varepsilon^2\left(\frac{3}{4}\tau^2-\frac{15}{16}\right)\right)z''_{\text{renorm}} + 2\left(\varepsilon + \varepsilon^2\left(\frac{3}{2}\tau\right)\right)z'_{\text{renorm}} + z_{\text{renorm}} \tag{149}$$

$$= \varepsilon^2 e^{-3/4\,\varepsilon\tau}\left(-\frac{3}{4}\tau\sin\left(\tau - 1/4\,\varepsilon\tau^2\right)\right) + O(\varepsilon^3\tau^3 e^{-3\varepsilon\tau/4})\,. \tag{150}$$

This is $O(\varepsilon^2\tau e^{-3\varepsilon\tau/4})$ instead of $O(\varepsilon^2\tau^2 e^{-3\varepsilon\tau/4})$, and therefore smaller. This *interprets* the largest term of the original residual, the $O(\varepsilon^2\tau^2)$ term, as a perturbation in the lengthening of the pendulum. The gain is one of interpretation; the solution is the same, but the equation it solves exactly is slightly different. For $O(\varepsilon^N\tau^N)$ solutions the modifications will probably be similar. Now, if $z \doteq \cos\tau$ then $z' \doteq -\sin\tau$; so if we include a damping term

$$\left(+\varepsilon^2 \cdot \frac{3}{8} \cdot \tau\theta'\right) \tag{151}$$

in the model, we have

$$\left(1+\varepsilon\tau+\varepsilon^2\left(\frac{3}{4}\tau^2-\frac{15}{16}\right)\right)z''_{\text{renorm}}+2\left(\varepsilon-\varepsilon^2\left(\frac{3}{2}\tau\right)+\varepsilon^2\frac{3}{8}\tau\right)z'_{\text{renorm}}+z_{\text{renorm}} = O\left(\varepsilon^3\tau^3 e^{-3/4\,\varepsilon\tau}\right) \tag{152}$$

and *all* of the leading terms of the residual have been "explained" in the physical context. If the damping term had been negative, we might have rejected it; having it increase with time also isn't very physical (although one might imagine heating effects or some such).

5.5 Vanishing Lag Delay DE

For another example we consider an expansion that "everybody knows" can be problematic. We take the DDE

$$\dot{y}(t)+ay(t-\varepsilon)+by(t)=0 \tag{153}$$

from [2, p. 52] as a simple instance. Expanding $y(t-\varepsilon)=y(t)-\dot{y}(t)\varepsilon+O(\varepsilon^2)$ we get

$$(1-a\varepsilon)\dot{y}(t)+(b+a)y(t)=0 \tag{154}$$

by ignoring $O(\varepsilon^2)$ terms, with solution

$$z(t)=\exp(-\frac{b+a}{1-a\varepsilon}t)u_0 \tag{155}$$

if a simple initial condition $u(0)=u_0$ is given. Direct computation of the residual shows

$$\Delta=\dot{z}+az(t-\varepsilon)+bz(t) \tag{156}$$

$$=O(\varepsilon^2)z(t) \tag{157}$$

uniformly for all t; in other words, our computed solution $z(t)$ exactly solves

$$\dot{y}+ay(t-\varepsilon)+(b+O(\varepsilon^2))y(t)=0 \tag{158}$$

which is an equation of the same type as the original, with only $O(\varepsilon^2)$ perturbed coefficients. The initial history for the DDE should be prescribed on $-\varepsilon\le t<0$

as well as the initial condition, and that's an issue, but often that history is an issue anyway. So, in this case, contrary to the usual vague folklore that Taylor series expansion in the vanishing lag "can lead to difficulties", we have a successful solution and we know that it's successful.

We now need to assess the sensitivity of the problem to small changes in b, but we all know that has to be done anyway, even if we often ignore it.

Another example of Bellman's on the same page, $\ddot{y}(t) + ay(t - \varepsilon) = 0$, can be treated in the same manner. Bellman cautions there that seemingly similar approaches can lead to singular perturbation problems, which can indeed lead to difficulties, but even there a residual/backward error analysis can help to navigate those difficulties.

6 Concluding Remarks

Decades ago, van Dyke had already made the point that, in perturbation theory, "[t]he possibilities are too diverse to be subject to rules" [29, p. 31]. Van Dyke was talking about the useful freedom to choose expansion variables artfully, but the same might be said for perturbation methods generally. This paper has attempted (in the face of that observation) to lift a known technique, namely the residual as a backward error, out of numerical analysis and apply it to perturbation theory. The approach is surprisingly useful and clarifies several issues, namely

- BEA allows one to directly use approximations taken from divergent series in an optimal fashion without appealing to "rules of thumb" such as stopping before including the smallest term.
- BEA allows the justification of removing spurious secular terms, even when true secular terms are present.
- Not least, residual computation and a posteriori BEA makes detection of slips, blunders, and bugs all but certain, as illustrated in our examples.
- Finally BEA interprets the computed solution solution z as the exact solution to just as good a model.

In this paper we have used BEA to demonstrate the validity of solutions obtained by the iterative method, by Lindstedt's method, by the method of multiple scales, by the renormalization group method, and by matched asymptotic expansions. We have also successfully used the residual and BEA in many problems not shown here: eigenvalue problems from [22]; an example from [29] using the method of strained coordinates; and many more. The purpose is for independent assurance that the output of the computation is faithful to the original model being solved. It is a kind of "reproducibility check," a topic of growing importance.

The examples here have largely been for algebraic equations and for ODEs, but the method was used to good effect in [34] for a PDE system describing heat transfer between concentric cylinders, with a high-order perturbation series in Rayleigh number. Aside from the amount of computational work required, there is

no theoretical obstacle to using the technique for other PDE; indeed the residual of a computed solution z (perturbation solution, in this paper) to an operator equation $\varphi(y; x) = 0$ is usually computable: $\Delta = \varphi(z; x)$ and its size (in our case, leading term in the expansion in the gauge functions) easily assessed.

It's remarkable to us that the notion, while present here and there in the literature, is not used more to justify the validity of the perturbation series. The work of Roberts [27] is an exceptional example, as discussed in the introduction.

We end with a caution. Of course, BEA is not a panacea. There are problems for which it is not possible. For instance, there may be hidden constraints, something like solvability conditions, that play a crucial role and where the residual tells you nothing. A residual can even be zero and if there are multiple solutions, one needs a way to get the right one. There are other things that can go wrong with this backward error approach. First, the final residual computation might not be independent enough from the computation of z, and repeat the same error. An example is if one correctly solves

$$\ddot{y} + y + \varepsilon \dot{y}^3 + 3\varepsilon^2 \dot{y} = 0 \tag{159}$$

and verifies that the residual is small, while *intending* to solve

$$\ddot{y} + y + \varepsilon \dot{y}^3 - 3\varepsilon^2 \dot{y} = 0\,, \tag{160}$$

i.e., getting the wrong sign on the $\dot{y}$ term, both times. Another thing that can go wrong is to have an error in your independent check but not your solution. The discrepancy alerts us that there *was* a problem, so this at least is noticeable. A third thing that can go wrong is that you verify the residual is small but forget to check the boundary conditions. A fourth thing that can go wrong is that the residual may be small in an absolute sense but still larger than important terms in the equation—the residual may need to be smaller than you expect, in order to get good qualitative results. A fifth thing is that the residual may be small but of the 'wrong character', i.e., be unphysical. Perhaps the method has introduced the equivalent of negative damping, for instance. This point can be very subtle.

A final point is that a good solution needs not just a small backward error, but also information about the sensitivity (or robustness) of the model to physical perturbations. We have not discussed computation of sensitivity, but we emphasize that even if $\Delta \equiv 0$, you still have to do it, because real situations have real perturbations. Nonetheless, we hope that we have convinced you that BEA can be helpful.

Acknowledgements We would like to thank Pei Yu, Robert H.C. Moir, and Julia Jankowski for their various contributions to this paper. We are also indebted to NSERC, Western University, and Galima Hassan for the key logistic support they provided.

References

1. Avrachenkov KE, Filar JA, Howlett PG (2013) Analytic perturbation theory and its applications. SIAM, Philadelphia
2. Bellman RE (1972) Perturbation techniques in mathematics, physics, and engineering. Dover Publications, Mineola
3. Bender CM, Orszag SA (1978) Advanced mathematical methods for scientists and engineers: Asymptotic methods and perturbation theory, vol 1. Springer, New York
4. Boas ML (1996) Mathematical methods in the physical sciences. Wiley, New York
5. Boyd JP (2014) Solving transcendental equations. SIAM, Philadelphia
6. Corless RM (1993) What is a solution of an ODE? ACM SIGSAM Bull 27(4):15–19
7. Corless RM, Corliss GF (1992) Rationale for guaranteed ODE defect control. In: Atanassova L, Herzberger J (eds) Computer arithmetic and enclosure methods, pp 3–12. North-Holland, Amsterdam
8. Corless RM, Fillion N (2013) A graduate introduction to numerical methods, from the viewpoint of backward error analysis. Springer, New York, 868 pp.
9. Corless RM, Gonnet GH, Hare DEG, Jeffrey DJ, Knuth DE (1996) On the Lambert *W* function. Adv Comput Math 5(1):329–359
10. De Bruijn NG (1981) Asymptotic methods in analysis, vol 4. Dover Publications, Mineola
11. Deuflhard P, Hohmann A (2003) Numerical analysis in modern scientific computing: an introduction, vol 43. Springer, New York
12. Enright WH (1989a) A new error-control for initial value solvers. Appl Math Comput 31:288–301
13. Enright WH (1989b) Analysis of error control strategies for continuous Runge-Kutta methods. SIAM J Numer Anal 26(3):588–599
14. Geddes KO, Czapor SR, Labahn G (1992) Algorithms for computer algebra. Kluwer Academic, Boston
15. Grcar JF (2011) John von Neumann's analysis of Gaussian elimination and the origins of modern numerical analysis. SIAM Rev 53(4):607–682
16. Higham NJ (1996) Accuracy and stability of numerical algorithms, 2nd edn. SIAM, Philadelphia
17. Holmes MH (1995) Introduction to perturbation methods. Springer, New York
18. Kirkinis E (2012) The renormalization group: a perturbation method for the graduate curriculum. SIAM Rev 54(2):374–388
19. Lanczos C (1988) Applied analysis. Dover Publications, Mineola
20. Lawden DF (2013) Elliptic functions and applications, vol 80. Springer Science & Business Media, Heidelberg
21. Morrison JA (1966) Comparison of the modified method of averaging and the two variable expansion procedure. SIAM Rev 8(1):66–85
22. Nayfeh AH (2011) Introduction to perturbation techniques. Wiley, New York
23. Olver FWJ, Lozier DW, Boisvert RF, Clark CW (2010) NIST handbook of mathematical functions. Cambridge University Press, Cambridge
24. O'Malley RE (2014) Historical developments in singular pertubations. Springer, New York
25. O'Malley RE, Kirkinis E (2010) A combined renormalization group-multiple scale method for singularly perturbed problems. Stud Appl Math 124(4):383–410
26. Rand R, Armbruster D (2012) Perturbation methods, bifurcation theory and computer algebra, vol 65. Springer Science & Business Media, New York
27. Roberts AJ (2014) Model emergent dynamics in complex systems. In: SIAM
28. Salvy B, Shackell J (2010) Measured limits and multiseries. J Lond Math Soc 82(3):747–762
29. Van Dyke M (1964) Perturbation methods in fluid mechanics. Academic, New York
30. Wilkinson JH (1963) Rounding errors in algebraic processes. Prentice-Hall series in automatic computation. Prentice-Hall, Englewood Cliffs
31. Wilkinson JH (1965) The algebraic eigenvalue problem. Oxford University Press, New York

32. Wilkinson JH (1971) Modern error analysis. SIAM Rev 13(4):548–568
33. Wilkinson JH (1984) The perfidious polynomial, vol 24. Mathematical Association of America, Washington
34. Zhang Y, Corless RM (2014) High-accuracy series solution for two-dimensional convection in a horizontal concentric cylinder. SIAM J Appl Math 74(3):599–619

Proof Verification Technology and Elementary Physics

Ernest Davis

Abstract Software technology that can be used to validate the logical correctness of mathematical proofs has attained a high degree of power and sophistication; extremely difficult and complex mathematical theorems have been verified. This paper discusses the prospects of doing something comparable for elementary physics: what it would mean, the challenges that would have to be overcome; and the potential impact, both practical and theoretical.

1 Memories of Jonathan Borwein

I knew Jon Borwein only briefly and slightly, but my few interactions were extremely memorable.

I first encountered Jon in connection with a collection of essays on the ontology of mathematics that my late father, Philip Davis, and I were putting together. Jeremy Avigad recommended him to me as a contributor, writing that "he has a lot to say about lots of things", which was certainly true. Jon and David Bailey agreed to write a chapter, and contributed a marvelous essay [3], spanning the world of experimental mathematics from computations of the partition function, to reciprocal series for π, to Ising integrals, to protein structure, to chimera states in oscillator arrays. Being a rather fussy editor, I asked for revisions, and then for more revisions, until on the third go-around Jon informed me, politely but firmly, that this was the final version.

Some time later, Jon generously invited me to present a talk at the 2016 meeting of ACMES. I didn't know how I fit in, since I barely do mathematics at all, and certainly don't do experimental mathematics, but Jon was very encouraging, and I ended up giving a talk which was an early version of the paper below.

The highlight of my visit to ACMES was certainly my dinner with Jon, Judi, and friends that evening. Jon, as his friends know much better than I, was in person

E. Davis (✉)
Department of Computer Science, New York University, New York, NY, USA
e-mail: davise@cs.nyu.edu

© Springer Science+Business Media, LLC, part of Springer Nature 2019

N. Fillion et al. (eds.), *Algorithms and Complexity in Mathematics, Epistemology, and Science*, Fields Institute Communications 82,
https://doi.org/10.1007/978-1-4939-9051-1_4

an ebullient, larger-than-life character and a wonderful raconteur; the conversation wandered from tales of mathematicians to the cleverness of octopi. It was worth going out to London, Ontario just for that evening.

In the months following, I had a couple of pleasant email exchanges with Jon: one about whether mathematicians worked through the proofs of the theorems they use, one about a historical point—a supposed medieval invention of a random number generator. (It proved to be fictitious.) I very much looked forward, then, to further interactions with him. I wish that I had the chance to know him much better and much longer.

2 Mathematical Proof Verification Software

One of the major accomplishments of late nineteenth and early twentieth century mathematics was the determination that essentially every rigorous mathematical proof can in principle be fully formalized as symbolic logical inference over set theory. To be precise, there are three statements here:

1. Practically[1] every mathematical concept can be defined in set-theoretic terms; and therefore every mathematical proposition can be formulated as a proposition in set theory.
2. Practically every mathematical proposition that has been rigorously proved, when cast into set theory, can be proved from standard axiomatizations of set theory using first-order logic.
3. Proofs in first-order logic can be characterized purely in terms of rules for manipulating strings of symbols; no understanding of the symbols, or mathematical intuition, or anything of the kind, is required.

The central landmark in establishing these facts was Whitehead and Russell's *Principia Mathematica*, though many other mathematicians, logicians, and philosophers both before and after were involved. The validity of a proof expressed in this symbolic form can be checked by a simple computer program that verifies that the sequence of assertions in the proof conforms to a set of rules for manipulating symbols. The verification program need understand nothing about the content of the proof, and the identical verification program will work for proofs in virtually every subfield of mathematics.

The software instantiation of this logical theory has been the development of *mathematical proof verification systems*. Over the past 50 years, software environment such as Isabelle/HOL [70], and others have been developed, which

[1] I do not know to what extent the experts agree on which, if any, kinds of theorems lie outside generalizations (1) and (2). As far as I know, there is essentially universal agreement that (1) and (2) are valid across most subfields in mathematics. Whether set theory is the *best* foundation for mathematics, or whether it is *important* for mathematics to have foundations at all, are separate questions.

allow a user to formulate symbolic encodings of proofs of mathematical theorems, which the software can then check for correctness. Substantial libraries of basic theorems and lemmas to draw on have been created, and some number of advanced, difficult proofs of major theorems have been formally verified, including:

- The prime number theorem, both using the analytical proof based on the zeta-function [38] and the "elementary" proof due to Selberg and Erdős [2].
- The Feit-Thompson theorem that every simple group of odd order is cyclic [30].
- The Kepler optimal packing theorem [34].

More or less, it seems safe to claim that;

- Any proof that is standardly taught in undergraduate math courses either already has been verified with this technology or could be a fairly small amount of work.
- Practically any theorem in the mathematical literature that has been proved could be verified with this technology; however, any given theorem might well require very substantial amounts of expert labor. This obviously does not apply to exceptionally complex proofs, such as the categorization of finite simple groups, which presumably would require truly impossible amounts of expert labor, or the proof of Mochizuki's ABC theorem, which, as of the time of writing, is not fully understood by anyone other than Mochizuki himself.

The question I wish to explore in this paper is this:

> Can a software technology comparable to mathematical proof technology be constructed that would allow the expression and validation of arguments in elementary physics, particularly those that connect theory and observation?

It will be convenient, for purpose of reference, to give this hypothetical project a name; I will dub it PAVEL.

Disclaimer This paper is exploratory and discursive; it neither presents established results nor constructs a tight argument. Moreover, my own limitations for carrying out this kind of investigation will soon become all too obvious to the reader; I do not know as much philosophy of science as I should for this purpose, and my knowledge of physics is altogether inadequate. The reason that the discussion in this paper is limited to elementary physics is that that's all the physics I know. (I will briefly discuss more advanced physics in Sect. 4.8.1, relying entirely for my information on [56].) However, to paraphrase Donald Rumsfeld, at 61 years old, one largely does analysis with the knowledge and abilities that one has, and not those that one would like to have.

The paper will proceed as follows. Section 3 will further discuss aspects of formal mathematical proof and of proof verification software further, since those are our primary comparanda and starting points. Section 4, which is the bulk of this paper, discusses the PAVEL project: What it would look like, and what it might accomplish. As part of this discussion, we will set up a straw man as a proposed architecture for PAVEL; the process of knocking down that straw man will help clarify what PAVEL should look like. Section 5 present a formalization of a simple word problem of the kind that might be used in PAVEL. Section 6 review the history of related ideas

and proposals. Section 7 discusses possible impact of a successful implementation of PAVEL on the philosophy of science. Section 8 will summarize and will discuss directions forward.

3 Formal Proof and Proof Technology in Mathematics

To begin with, let us consider the case of mathematics in more depth. We will discuss briefly the value of the logic-based theory of mathematical proof and of proof-verification technology and their limitations; this will be useful as a point of comparison for discussing the potential value and limitations of pursuing these in the context of physics.

Logic-based analysis of mathematical concepts and proofs provides a normative model for rigorous argumentation in mathematics, which is perfectly well-defined, and which applies to practically every proof throughout the discipline. We will note some limits on the significance of this below; however, those limits do not make this finding any less significant or astonishing.

Moreover, logic-based analysis of mathematics led to the development of mathematical logic, a field that is of enormous inherent interest; provides results important for other areas of mathematics, e.g. the unsolvability of Diophantine equations; and is central to computation theory. The practical consequences throughout computer technology are incalculable.

It is certainly important to keep in mind the limits of logical analysis as a characterization of mathematics. It is presumably of little or no value in developing a *cognitive* theory of mathematical understanding and reasoning; that is, a psychological theory of how professional mathematicians, lay people, children, or animals understand mathematical concepts and arguments [24]. In *historical* studies, the twentieth-century logical analysis is treacherous to use as a framework; it can lead one to a "Whig history" point of view in which, let us say, Newton's conception of a point at infinity or Euler's conception of a function is viewed as a defective version of our own perfect understanding. Even as regards contemporary mathematics, it has been argued that the logical sense of proof does not encompass all that we mean by proof, and that the formulation of mathematical concepts in set-theoretic terms does not encompass all that we mean by those concepts [15]. The formal viewpoint omits the social role of proofs; proofs are one form of communication among mathematicians. But, again, these limitations do not negate the enormous importance of this kind of analysis.

Moreover, thus far the impact of either the theory or the technology on the daily labors of the mass of professional mathematicians, working in, say, partial differential equations, or homology theory, or ideal theory, has been much less than the notoriety of mathematical logic and of theorems such as those of Gödel's in popular mathematics and among philosophers of mathematics might suggest. Few, if any, undergraduate math majors at American universities require a course in mathematical logic; and more than one well-regarded math department

does not offer *any* regular course in mathematical logic. As for proof verification software, most mathematicians are probably only dimly aware that it exists at all.

The impact of the *technology* of proof verification systems has been enormously less than the *theory* of mathematical logic. Still, it has had a significant impact in certain areas, and may well have greater impact in the future. Perhaps its greatest impact to date is as part of a wide range of activities in implementing logical reasoning on computer systems. This body of work in general has had many practical applications, including logic-based programming languages, automated software and hardware verification, knowledge-based artificial intelligence (AI) reasoners and expert systems. Broadly speaking, these kinds of systems lie along a spectrum, with different trade-offs of the expressivity and depth of the representation, on the one hand, versus efficiency of inference, on the other. Mathematical proof verification lies on the extreme end of favoring expressivity at the expense of efficiency; nonetheless, technical developments here have impact on similar project with more directly practical applications.

In particular, proof verification is closely related to logic-based software and hardware verification. Much more work has been invested in software and hardware verification than in mathematical proof verification because of its direct practical significance. The goal of these kinds of verification system can range from limited verification, determining that the software is free from specific kinds of bugs, to complete verification that the program works correctly in all respects. Bug-checking verification is currently a very powerful technology which can be applied to enormous, complex programs such as operating systems, and complex hardware architectures, such as state-of-the-art CPUs.

Complete verification of software correctness is much more difficult. A major obstacle is that it is extremely hard even to state complete specifications for what a complex program should do; the specification statement ends up being almost as long, and much less intelligible, than the program. Therefore, verification of a formal specification works best for functionalities where the logical specification of the desired functionality is much simpler than its implementation, such as mathematically-oriented software. For example, Harrison [37] carried out the formal verification of library functions that do floating-point computation of trigonometric functions; the verification raised some interesting subtle issues of correctness beyond what is usually considered in numerical analysis.

In the long term, we can hope to see other kinds of impact on mathematical practice:

- Confidence in highly complex proofs can be increased.
- The development of representation might be a step toward content-based search for theorems in the mathematical literature. Currently, it is often easier to reprove a lemma than to find it in the literature.
- Ultimately, this is a step toward a “general AI mathematician”; an AI that carry out all, or many, of the activities of a research mathematician, either by itself or in partnership with a human.

3.1 What Hasn't Been Done for Math

A number of limitations of the technology should be noted.

Obviously, we do not have AI programs that can *generate* proofs of a general kind in advanced math or even in college-level math. The technology for symbolic manipulation, in systems like MAPLE and MATLAB has become extraordinarily sophisticated [3] this will suffice for most proofs in high-school and some fraction of proofs in some areas of math. Beyond that, a handful of interesting original proofs have been generated by computers, either using general theorem-proving technology (e.g. the Robbins conjecture [64]) or using programs specifically written for a particular case (e.g. the four-color theorem [1].) But we are far from having a program that can generate the kinds of proofs required of undergraduate math majors.

We are nowhere near having an AI program that can read the mathematical literature and "understand" it, in the sense of translating it to a formal representation, or even a program that can do most of this with occasional assistance from a "human in the loop". There has been some work on the much more limited task of translating word problems stated in English into a representation and then solving the equations. For instance Kushman et al. [53] report a program that achieves an overall accuracy of 68.7% on textbook problems that translate into two equations in two unknowns.

A more immediate issue is user-unfriendliness. By all accounts, the learning curve for this technology is extremely challenging and the user interface uninviting. Consequently, when a new theorem is verified it is much more likely that an expert on verification has learned the math involved in the theorem than a mathematician who is an expert in the area of the theorem has learned to use the verification technology. Verifying the Feit-Thompson theorem involved a 6-year collaborative effort by a team of fifteen mathematicians[2] (Gonthier, 2013). If the technology were easy to use, then one could imagine the "mathematician in the street" taking the trouble to master in order to check that their proofs are correct; but currently that seems far off.

3.2 Word Problems

Another part of math, particularly elementary math, is word problems.

Let us pass over the large problems of natural language processing and of knowledge base construction and focus on the representational problem: How can the content of a word problem plausibly be expressed in a logical representation that describes the real world situation and that suffices for the solution of the problem, when combined with the relevant mathematical theory? The problem formulation

[2]This does not, of course, imply that it required ninety man-years of work.

should as far as possible be a direct expression of the *meaning* of the natural language formulation of the problem. That is, we want as much as possible of the reasoning needed to find the solution to be made explicit in the proof structure built on the formulation, and as little reasoning as possible done implicitly in the process of translating the natural language expression into the formal problem specification.

Tables 1 and 2 illustrate what I have in mind, for one well-known brain teaser.

Some comments about the formalization in Tables 1 and 2. The representation uses a sorted, first-order logic with theories of time, dimensioned quantities and vectors, and Euclidean geometry, that I have developed for representing physical theories [20]. The semantics is straightforward, and the intended meaning is hopefully self-evident. There is a partial account in [23]. `Typewriter font` is used for object-level symbols; *Italics* are used for sortal symbols. Non-logical symbols have an initial upper-case letters, object-level variables have an initial lower-case letter, and sortal variables use Greek letters. Sorts of symbols are declared in a form modeled on declarations in typed programming languages such as Java. Thus, for example, the declaration

`VectorFrom(x,y:`*`Point`*`)` $\rightarrow$ *`Vector[Distance]`*

means that `VectorFrom` is a function symbol, taking two arguments, `x` and `y`, both of which are *Point*s, and returning a value which is a vector of dimension *Distance*.

The problem formulation in Tables 1 and 2 combined with suitable basic axioms and definitions of the dimensions involved, time, and Euclidean space will support a proof of the conclusion `ArcLength(Z,T0,TC) = 150 * Mile.`

The complexity of Tables 1 and 2 together with the domain axiomatization not shown here, as compared to the simplicity, both of the natural language expression, and of the mathematical forms that a human reasoner might write down or think through in solving this problem, might be taken as a sign that we are seriously on the wrong track here. In particular the gap between the phrase "the bird flies back and forth between the two trains" and the complex axioms 6 and 7 is concerning. Certainly any human being would find it much easier to solve the problem directly from the natural language formulation than to translate the natural language into the formulas in these tables. (I myself spent some hours getting them right, and I have 30 years' practice in writing these kinds of formalisms.) More than that, one might well worry that it would be easier to write a program that could solve these kinds of problem than to write one that could generate these axiomatizations.

There are a number of partial answers, at different levels. First, the gap from "flies back and forth" to axioms 6 and 7 can be bridged by positing an intermediate form,[3] such as `Until(t,`ϕ`,`ψ`)`, meaning "Starting at time `t`, ϕ remains true at least until ψ becomes true." Axiom 6 can then be worded,

[3]Technically speaking, the operator `Until` here can be viewed as "syntactic sugar", or as a temporal modal operator, or, if one performs some "representational tinkering" on the arguments, as a first-order predicate.

Table 1 Formalization of a word problem: sorts and symbols

Problem: Two trains 100 miles apart are speeding toward one another. One is going 75 mph, the other is going 25 mph. A bird flies back and forth between them at 150 mph. How far does the bird travel before the trains collide?
Sorts: *Object, Time, Duration, Point, Distance, Speed, Real*
Sortal Functions:
Fluent[α]—Function from *Time* to sort α.
Vector[α]—If α is a real-valued dimension, then a vector of dimension α.
For example, *Vector[Speed]* is the sort of velocities.
Vector[Real] is the sort of dimensionless vectors.
$\alpha \otimes \beta$—Infix operator: Dimension α times dimension β.
For example, *Duration* $\otimes$ *Speed* = *Distance*.
$\alpha \oslash \beta$—Dimension α divided by dimension β.
For example, *Distance* $\oslash$ *Duration* = *Speed*
Constant Symbols:
`TrA`→*Object*—the first train.
`TrB`→*Object*—the second train.
`B`→*Object*—the bird.
`T0`→*Time*—the initial time.
`TC`→*Time*—the time the two trains collide.
`Mile`→*Distance*—a mile
`Hour`→*Duration*—an hour
Standard numerals → *Real*.
Function Symbols:
`Place(x: Object)` → *Fluent[Point]*. The function tracking the position of object `x` over time.
`Velocity(x: Object)` → *Fluent[Speed]*. The function tracking the velocity of object `x` over time.
`Magnitude(v:Vector[α])` → α. Magnitude of vector `v`. $\lvert\vec{v}\rvert$.
`Direction(v:Vector[α])` → *Vector[Real]*. Direction of `v`. $\vec{v}/\lvert\vec{v}\rvert$.
`V(t:Time, q:Fluent[α])` → α. Value of fluent `q` at time `t`.
`VectorFrom(x,y:Point)` → *Vector[Distance]*. The vector `y` − `x`.
`Vec*(s:α, v:Vector[β])` → *Vector*[$\alpha \otimes \beta$].
Scalar `s` of dimension α times vector `v` of dimension β.
`x:α * y:β` → $\alpha \otimes \beta$.
Infix operator `x` $*$ `y` where `x` has dimension α and `y` has dimension β.
`x:α / y:β` → $\alpha \oslash \beta$.
Infix operator `x`/`y` where `x` has dimension α and `y` has dimension β.

Table 2 Formalization of a word problem: problem formulation

Problem Statement:

1. `Magnitude(VectorFrom(V(T0,Place(TrA)),V(T0,Place(TrB)))) = 100*Mile.`
   The two trains are initially 100 miles apart.

2. `V(TC,Place(TrA)) = V(TC,Place(TrB))`
   The two trains collide at time `TC`.

3. $\forall_{\texttt{t}}$ `T0` $<$ `t` $<$ `TC` $\Longrightarrow$
   `V(t,Velocity(TrA)) =`
   `Vec*(25 * Mile/Hour,Direction(VectorFrom(V(t,Place(TrA)),`
   `V(t,Place(TrB)))).`

   Between `T0` and the collision, train `TrA` moves at 25 mph toward train `TrB`.

4. $\forall_{\texttt{t}}$ T0 $<$ t $<$ TC $\Longrightarrow$
   `V(t,Velocity(TrB)) =`
   `Vec*(75 * Mile/Hour,Direction(VectorFrom(V(t,Place(TrB)),`
   `V(t,Place(TrA)))).`

   Between `T0` and the collision, train `TrB` moves at 75 mph toward train `TrA`.

5. `V(T0,Place(B)) = V(T0,Place(TrA)).`
   The bird starts at train `TrA`.

6. $\forall_{\texttt{ta,tb}}$ `T0` $<$ `ta` $<$ `TC` $\wedge$ `V(ta,Place(B))` $=$ `V(ta,Place(TrA))` $\wedge$ `ta` $<$ `tb` $<$ `TC` $\wedge$
   $[\forall_{\texttt{tx}}$ `ta` $<$ `tx` $\leq$ `tb` $\Longrightarrow$ `V(tx,Place(B))` $\neq$ `V(tx,Place(TrB))`$]$ $\Longrightarrow$
   `V(tb,Velocity(B)) =`
   `Vec*(150 * Mile/Hour,`
   `Direction(VectorFrom(V(tb,Place(B)), V(tb,Place(TrB)))).`
   If the bird is at train `TrA` at time `ta`, and it does not reach train `TrB` any time between `ta` and `tb` inclusive, then at time `tb` it is moving toward `TrB` at 150 mph.

7. $\forall_{\texttt{ta,tb}}$ `T0` $<$ `ta` $<$ `TC` $\wedge$ `V(ta,Place(B))` $=$ `V(ta,Place(TrB))` $\wedge$ `ta` $<$ `tb` $<$ `TC` $\wedge$
   $[\forall_{\texttt{tx}}$ `ta` $<$ `tx` $\leq$ `tb` $\Longrightarrow$ `V(tx,Place(B))` $\neq$ `V(tx,Place(TrA))`$]$ $\Longrightarrow$
   `V(tb,Velocity(B)) =`
   `Vec*(150 * Mile/Hour,`
   `Direction(VectorFrom(V(tb,Place(B)), V(tb,Place(TrA)))).`
   If the bird is at train `TrB` at time `ta`, and it does not reach train `TrA` any time between `ta` and `tb` inclusive, then at time `tb` it is moving toward `TrA` at 150 mph.

Evaluate: `ArcLength(T0,TC,Place(Z)).`

```
6′ ∀ta T0 < ta < TC ∧ V(ta,Place(B)) = V(ta,Place(TrA)) ⟹
    Until(ta,Place(B) = Place(TrB),
                Velocity(B) =
                Vec*(150 * Mile/Hour,Direction
                  (VectorFrom(Place(B), Place(TrB))))).
```

and axiom 7′ would be analogous.

Getting to these intermediate representation 6′ and 7′ from "flies back and forth", seems considerably more doable, though certainly not a solved problem; and the process of getting from the intermediate forms 6′ and 7′ to axioms 6 and 7 can easily be completely specified.

Second, while a shallower semantic analysis might *often* suffice to build a computer program that solves word problems, in the same way that human students sometimes learn to solve math problems by pattern matching against problems that they have seen before, I would argue that solving these problems *robustly* will require a semantic representation of the depth of Tables 1 and 2. For instance, to answer the particular question "How far will the bird fly?", a computer does not actually have to understand what is meant by "back and forth" at all; it suffices to understand that the bird is flying at 150 mph. However, that will not suffice if you change the problem statement or the question:

- How many times is the bird exactly 10 miles from one or the other train?"
- Is there any time at which the distance from the bird to the first train and the distance to the second train are both simultaneously decreasing?
- Suppose that whenever the bird reaches a train, it rests for a minute. How far does it fly in that case?

For any of these, you will need a level of understanding comparable to Tables 1 and 2.

The objection that people find it easier to solve the problem than to work through the notation of Tables 1 and 2, though often raised as a derisive dismissal of logic-based notations, really has no weight at all. Working through any description of how a cognitive task is carried out is almost always more difficult than performing the task. I can guarantee that if somebody builds a system based on machine learning that solves the bird problem, that will also be harder to understand than solving the bird problem.

In general, what is the state of the art in representing math word problems in this way? I don't know of any systematic study; it would be interesting to carry one out. But my guess would be that problems in high school level or freshman college level math—that is, elementary problems in Euclidean geometry and trigonometry, basic algebra, differential and integral calculus through the first three college courses, and combinatorics—would rarely if ever present difficulties.

Probability theory might often be challenging. The Kolmogorov formulation of probability theory suffices for all formal mathematical theorems in probability theory (as far as I know); if you want to prove the central limit theorem, say, or the existence of limiting distribution for a Markov chain, you can state it and prove

it within the Kolmogorov formulation. Likewise, if a word problem can be easily cast in terms of a sample space, then it can be represented and solved. For instance, if we wish to answer the question, "What is the probability that a five-card hand is a flush (including straight flush)?", then it is straightforward to axiomatize the combinatorics and prove that $\#\texttt{Hand} = C(52,5)$, $\#\texttt{Flush} = 4 \cdot C(13,5)$ and therefore $|\texttt{Flush}| = 4 \cdot C(13,5)/C(52,5) = 0.00198$

However, in many cases the derivation is much more problematic. Consider the following well-known puzzle[4]:

A. John has two children and at least one of them is a boy. What is the probability that he has two sons? **Answer:** 1/3.
B. John has two children; the older is a boy. What is the probability that John has two sons? **Answer:** 1/2.
C. John has two children; at least one is a boy born on Tuesday. What is the probability that John has two sons? **Answer:** 13/27.

I'm ignoring here the slight correlation in days of birth due to twins, the even slighter correlation in sex due to identical twins, and the fact that male births are not exactly 50% of all births.

(Peter Winkler (email to the author, 12/28/17) has pointed out that almost any real world situation where you know that John has 2 children and one is a boy—for instance, if you are told that he has two children, and then you run into him with one child, who is a boy—conforms to the analysis in (A) or (C) rather than the one in (B). However, he reports running into one real-world exception: A friend of his was pregnant with fraternal twins, and had some kind of genetic test that gives positive results if either fetus has a Y chromosome. In that case, the analysis in (A) held; there was a 1/3 chance that she was bearing two boys.)

If you consider $\texttt{Prob}(\phi|\psi)$ to be a sentential operator then the probabilities to be evaluated are easily expressed:

A. $\texttt{Prob}(\#\{\texttt{x}|\texttt{Child}(\texttt{x},\texttt{John}) \wedge \texttt{Sex}(\texttt{x}) = \texttt{Male}\} = 2 \mid \exists_{\texttt{y},\texttt{z}}\{\texttt{x}|\texttt{Child}(\texttt{x},\texttt{John})\} = \{\texttt{y},\texttt{z}\} \wedge \texttt{y} \neq \texttt{z} \wedge \texttt{Male}(\texttt{y}))$
B. $\texttt{Prob}(\#\{\texttt{x}|\texttt{Child}(\texttt{x},\texttt{John}) \wedge \texttt{Sex}(\texttt{x}) = \texttt{Male}\} = 2 \mid \exists_{\texttt{y},\texttt{z}}\{\texttt{x}|\texttt{Child}(\texttt{x},\texttt{John})\} = \{\texttt{y},\texttt{z}\} \wedge \texttt{y} \neq \texttt{z} \wedge \texttt{Male}(\texttt{y}) \wedge \texttt{Older}(\texttt{y},\texttt{z}))$.
C. $\texttt{Prob}(\#\{\texttt{x}|\texttt{Child}(\texttt{x},\texttt{John}) \wedge \texttt{Sex}(\texttt{x}) = \texttt{Male}\} = 2 \mid \exists_{\texttt{y},\texttt{z}}\{\texttt{x}|\texttt{Child}(\texttt{x},\texttt{John})\} = \{\texttt{y},\texttt{z}\} \wedge \texttt{y} \neq \texttt{z} \wedge \texttt{Male}(\texttt{y}) \wedge \texttt{Born}(\texttt{y},\texttt{Tuesday}))$

But I don't know of any logical formalization which will allow one to go from forms like the above to stochastic models in which the specified probabilities can be calculated.

Furthermore, stochastic models whose complexity seems quite moderate when presented in an applied probability textbook, such as the k-gram model of language

[4]The "Monty Hall" problem is even trickier, and has tripped up professional mathematicians.

production, end up being much more intricate when written out in full in a logical notation. The elegant mathematical formulas used to describe such models in the research literature often turn out, on careful analysis, to be a morass of implicit quantifiers of implicit scope and ambiguous variable symbols, superscripts, and subscripts, meaningful only to someone who reads the accompanying text and understands what is intended.

Mathematically, statistics is largely a subfield of probability, but it seems to gravitate toward that class of probability problems that are particularly difficult to formulate logically. I suspect that many word problems in statistics would be extremely difficult to represent in a reasonable way that supports the statistical inference.

4 Physics

With the example of mathematics in mind, as inspiration and point of comparison, we can now enter on the main topic of the question. Vaguely put, can we carry out this same kind of project for physics? More specifically, can we achieve the following:

- Represent some significant part of the content of physics, including both foundational theories and the experimental and observational results that they rest on, in a formal language?
- Characterize some significant part of reasoning and argumentation in physics, particularly the reasoning that connects foundational theories to "real world" situations, in a formal theory of reasoning?
- Implement the representation and reasoning mechanisms in a technology for argument verification for physics?

4.1 The Potential Value of This Undertaking

If PAVEL can be built, then it seems to me that both the finished product and the work involved in developing the product are likely to have significant payoffs, in a number of different directions.

First, the work involved in PAVEL might shed some light on issues in the philosophy of science. That will be easier to discuss after we have looked at specific issues, so I am deferring it to Sect. 7.

Second, work on PAVEL would be a step toward in developing AI that can do flexible, powerful commonsense physical reasoning. Gary Marcus and I have argued at length elsewhere [21, 22] that approaches to physical reasoning based on simulation, which currently entirely dominate AI physical reasoning, are insufficient for many of the kinds of problems that a general purpose AI will confront, besides being implausible as general cognitive models. It is certainly the case that physicists,

in reasoning about physical situations, use a wide variety of reasoning techniques beyond simulation. It seems likely, therefore, that analyzing the kinds of reasoning needed to do physics may open up the space of automated reasoning techniques available to AI reasoners.

Third, it may be possible to integrate the reasoning in PAVEL with program verification technology, and thus to formally verify the validity of programs that control physical devices, in safety-critical ways: airplanes, robots, nuclear reactors and so on. A major accomplishment in program verification, some years ago [80] was the verification of the control software for the Airbus airplane. However, that verification only proved that the *program* won't crash; it didn't prove that the *airplane* won't crash. In work closer to PAVEL, Jeannin et al. [48] formally verified a hybrid system for the avoidance of aircraft collision. Their domain axiomatization is similar in flavor to the axiomatizations we develop in this paper, but are quite specialized to the problem under discussion.

Finally, PAVEL would be a step toward the "super-AI-scientist" fantasized by the many "AI as messiah" enthusiasts; an AI that can achieve an integrated, total, understanding of *all* of science and thus can solve those of our problems that can be solved that way. In fact, it seems to me that solving the issues involved in PAVEL is a *necessary* step; the super-AI-scientist *must* have the kind of general understanding that is encoded in PAVEL.

Paleo [71] similarly argue in favor of expressing arguments in physics in proof-theoretic terms, arguing that this will clarify existing debates in the philosophy of science and "open new conceptual bridges between the disciplines of Physics and Computer Science."

4.2 The Bayesian Formulation

In thinking about PAVEL, I find it helpful to keep in mind the Bayesian approach to scientific hypothesis and data [46, 47, 72, 81], partly as a framework to make things concrete, partly as a foil to work against.

The basic Bayesian formulation of scientific theorizing is straightforward. There is a space Φ of possible scientific theories; that is, each hypothesis $h \in \Phi$ is a complete theory of physics. There is a space Δ of possible total data collections; that is, each element $D \in \Delta$ is a combined record of all the outcomes of all the experiments and observations ever performed. We are given one particular collection of data $D \in \Delta$. We are looking for the most likely theory given the data; that is, $\text{argmax}_{h\in\Phi} P(h|D)$. So now, as always, we use Bayes' Law:

$$\text{argmax}_{h\in\Phi} P(h|D) = \text{argmax}_{h\in\Phi} P(D|h)P(h)$$

All that's left is to set the priors $P(h)$, to compute the conditional probabilities $P(D|h)$, and to find the maximum of the expression. Within reason, the exact values of the priors don't matter much anyway, since their contribution is soon swamped

by the data. That is, you think of each imaginable theory of physics as a generative stochastic process that outputs data, and thus defines a probability distribution $P(\cdot|h)$ over Δ. You imagine a prior distribution over all such processes. Then you match the observed data to the predicted data.

One thing that's appealing about this is that it completely eliminates the need for scientific induction as a separate mode of reasoning. There is no need to address the difficult question of what it means for data to support a hypothesis; Bayes' law allows you to turn that into the much more straightforward question of whether a hypothesis predicts data.

The hypotheses in Φ must all be mutually exclusive or the method doesn't work. They cannot be theories in the logical sense, organized in a lattice of generality, because the probability of a more general theory is necessarily less than a narrow theory. Given any premise or data, the conditional probability of $\forall_x B(x) \implies A(x)$ is cannot be greater than the conditional probability of $\forall_x B(x) \wedge C(x) \implies A(x)$, because the first sentence implies the second. If, therefore, Φ included more and less general theories, the maximum would never land on the most general theories; those are always the least probable. In Bayesian models, therefore, all the hypotheses are maximally specific. For example, in the Hierarchical Bayesian Models theory [85], all the theories in Φ are generative stochastic models that generate data. The choice therefore, is not between "$\forall_x B(x) \implies A(x)$" and "$\forall_x B(x) \wedge C(x) \implies A(x)$". Rather the choice is between

> H1(p): $\forall_x B(x) \implies A(x)$ and A occurs randomly with probability p among entities that are not B;
>
> vs.
>
> H2(p): $\forall_x B(x) \wedge C(x) \implies A(x)$ and A occurs randomly with probability p among entities that are not both B and C;

Here p is a parameter that will be optimized (viz. set to the measured frequency of A in the two referent sets). Since H1 no longer implies H2, there is now nothing to prevent us from assigning a higher prior probability to H1 than to H2.

As is well known, Bayesian theories are equivalent to minimum description length theories under the information-theoretic correspondence $I(\phi) = -\log_2 P(\phi)$. That is: you choose an optimal encoding for hypotheses based on their prior probabilities, or, conversely, you set the prior probability to be exponential in the length of the theory: $P(h) = 2^{-I(h)}$ where $I(h)$ is the number of bits needed to express h. For each hypothesis h, you choose an optimal encoding for possible data outputs where $I(D|h) = -\log_2 P(D|h)$. So overall you have attained an expression of length $I(D) = I(D|h) + I(h)$. Choosing the most probably value of h given D is then equivalent to using Occam's razor to choose the shortest expression of the data; that is, we find the simplest, most elegant theory that explains the data.

In some ways, this seems enormously appealing, almost inevitable; in other ways it seems completely far-fetched (Sober [79] is a sharp critique.) The idea that there exists a space Φ of fully formed physical theories prior to making any observations and the idea that there is a space Δ of possible data collections that exists independent of the physical theory—all the theory does is to change the

conditional probability distribution over Δ—do not correspond to our experience of how science actually progresses. What we see, rather, is a tight mutual dependence between theory and data. On the one hand, in the development of science, the data collected thus far affects, not just the choice of theories, but even the language that the theories are expressed in. On the other hand, the choice of which experiments to carry out or observations to make depends on what is known about the physics. As we will discuss further in Sect. 4.5, an experimental device or design and the interpretation of its behavior as data depend critically on knowledge of physics; if the physics of world were otherwise, then the experiment would be not merely inconclusive, it would be meaningless or impossible.

A Bayesian might justify the spaces Φ and Δ with the following *Gedanken* experiment. Let us fix the scientific investigator under discussion: perhaps a new born baby [31], perhaps a scientific community over millennia. Imagine now the collection Ω of all epistemically possible physical worlds; or, at least, all those consistent with the existence of a baby/scientific community. (This is somewhat similar to Tegmark's [84] Level IV multiverse.) We insert a clone of the investigator into each possible world. The investigator's task in each world is to find out which world he is in; or at least to get some information about that. We now, from the outside, observe all these investigators in all these worlds. At a certain point, we stop him; we find out what data he have seen and we ask him what physical theory he now believes, or what set of alternative theories he has under consideration. For each world $w \in \Omega$, let D_w be the collection of data that the investigator has compiled in w and let Φ_w be the set of alternative theories that the investigator in w reports. Then $\Delta = \{D_w | w \in \Omega\}$ and $\Phi = \bigcup_{w\in\Omega} \Phi_w$.

The fact that, in different worlds, the investigator will perform different experiments and make different observations is merely the standard scenario in decision theory in which the space of possible actions may depend on prior observations. It slightly complicates Bayesian inference, but does not fundamentally alter it.

The reason that this view seems alien (the Bayesian can continue) is that, due to our own cognitive limitations, we are not used to taking such a large view; we are used to looking at the development of science through a much narrower window. However, fundamentally, behind the scenes, this is what is going on. In fact the ultimate AI scientist will be able to take exactly this view of things; it will take into account *all* of the scientific data D that has been collected and chose the best among *all* possible scientific theories $h \in \Phi$, up to limits of computational power.

The transformation of a theory of physics—that is, a collection of physical laws—into a stochastic model elicits starkly varying reactions from different people. To a Bayesian, this is natural, indeed inevitable; trying to do inference without a distribution is like trying to bake a cake without an oven. To a logicist, burdening an elegant, well-motivated logical theory with an ugly, arbitrary probability distribution is adding an unnecessary excrescence; it is like trying to bake a cake with a blowtorch. As we will discuss in Sect. 4.5, the relation between theory and a scientific theory, in general will carve out a strangely shaped, lower-dimensional

manifold[5] in the space Δ of all data collections; and defining a natural distribution over such a manifold is a problematic and ill-defined undertaking. It is hard enough to characterize the sense in which the observations of the tides, for example, can be explained in terms of Newton's law of gravity. The question, "What is the probability distribution over observations of the tides, given Newton's law?" seems a truly strange one.

We will not pursue this argumentation back and forth further here; however, in the course of our discussion, we will refer back to this as a possible frame of reference. An implementation of this approach by Kemp and Tenenbaum [49] will be discussed in Sect. 6.4.2.

4.3 Straw Man: The Tee-Shirt Model of PAVEL

At this point I want to put up a straw man proposal for an approach to building PAVEL; the process of knocking it down will serve as an effective frame for making the points I want to make.

The straw man is this: We express the famous laws of physics in a formal logic. These are the axioms of our system. Everything else is proved from those axioms. In our Bayesian formulation, this collection of axioms is the hypothesis h.

I call this "the tee-shirt model", because tee-shirts printed with a few elegant equations are popular among the geekier part of the population. Full disclosure: As an undergraduate I owned and wore a Maxwell's equations sweatshirt. Less snarkily, I will also call this approach "the foundational approach" when that is more appropriate.

Now, the tee-shirt model is *exactly* the equivalent of what is done in mathematical proof verification systems. The basic axioms given are the ZFC axioms of set theory (or some other similar foundational set); everything else in math is defined in terms of sets and all proofs can ultimately be traced back to the foundational axioms. At the other extreme, it is hard to imagine that anyone would propose anything like the tee-shirt model for chemistry or biology, let alone for the cognitive or social sciences, with the possible exception of economics. But physics occupies a middle ground here, and it seems as though the tee-shirt model should be more or less attainable. I will argue that, at least in our current state of understanding, the tee-shirt model is nowhere close to right, for quite a number of reasons.

[5] You may argue that because of noise, the theory does not correspond to a lower-dimensional manifold, it corresponds to a probability distribution centered on the manifold. That hardly helps, because now the probability distribution of the projection onto the manifold depends strongly on largely arbitrary assumptions about the noise.

4.4 The Equations Are More Complicated than Their Tee-Shirt Version

To begin with a rather minor point: the actual equations of physics are often more complicated than they appear on tee shirts.[6]

To take a simple example: On the tee-shirts Newton's theory of universal gravitation might well be given in two equations;

$$F = G\frac{m_i m_j}{r^2} \qquad \text{Universal law of gravitation}$$

$$F = m\frac{d^2x}{dt^2} \qquad \text{Newton's 2nd law}$$

But actually, a force is a vector with a direction, and Newton's second law applies to the vector sum of all the forces incident on a particle. Forces and positions are functions of time. We need to exclude forces by a particle on itself. So for point particles, the equations become

$$i \neq j \implies \vec{F}_{i,j}(t) = G\frac{m_i m_j \cdot \hat{\theta}(\vec{x}_j(t) - \vec{x}_i(t))}{|\vec{x}_j(t) - \vec{x}_i(t)|^2}$$

$$m_i \frac{d^2\vec{x}_i(t)}{dt^2} = \sum_{j \neq i} \vec{F}_{i,j}(t)$$

The indices i, j range over particles. We use $\hat{\theta}(\vec{v})$ to mean the direction of vector $\vec{v}$: $\hat{\theta}(\vec{v}) = \vec{v}/|\vec{v}|$.

If we want to have extended objects, then things become still more complicated. We can develop a theory of eternal extended objects constructed from particle by introducing a predicate $c(p_i, p_j)$, meaning "particle p_i is connected to particle p_j." The object is then the set of particles within the transitive closure of the relation c.

For a rigid object, ignoring contact forces between the objects—that is, allowing objects to freely interpenetrate—we get the following rules:

$$c(p_i, p_j) \implies c(p_j, p_i)$$

$$c(p_i, p_j) \implies |x_j(t) - x_i(t)| = d_{i,j}$$

$$\vec{F}_{i,j}(t) = -\vec{F}_{j,i}(t)$$

[6]The Lagrangian for the Standard Model, given in full in [33], is 36 lines long and has something like 170 terms and 1000 symbols. However, Gutierrez does claim that he has printed tee shirts with the whole thing.

$$\neg c(p_i, p_j) \implies \vec{F}_{i,j}(t) = G\frac{m_i m_j \cdot \hat{\theta}(\vec{x}_j(t) - \vec{x}_i(t))}{|\vec{x}_j(t) - \vec{x}_i(t)|^2}$$

$$m_i \frac{d^2\vec{x}_i(t)}{dt^2} = \sum_{j \neq i} \vec{F}_{i,j}(t)$$

The second equation above expresses the rigidity constraints by requiring the distance between connected particles to be constant. The third equation is Newton's third law.

For elastic objects, the second equation above, characterizing the constraint between connected particles, is replaced by Hooke's law:

$$c(p_i, p_j) \implies \vec{F}_{i,j}(t) = k_{i,j}(|\vec{x}_j(t) - \vec{x}_i(t)| - d_{i,j}) \cdot \hat{\theta}(\vec{x}_j(t) - \vec{x}_i(t))$$

The formulation for continuum mechanics is similar, but replaces the force by force density, the relation between connected particles by the corresponding partial differential equations, and the summation by an integral.

These don't have quite the same panache on a tee shirt. This observation does not refute the possibility of using a foundational model to build PAVEL, but it does suggest that formulating the foundational equations correctly may take more care than one might suppose.

4.5 The Grounding of Physics in Observation and Experiment

The most serious objection to the tee-shirt model is it ignores the problem of expressing the connection between the terms in the equations and the ways that these are manifested in the world that a physicist interacts with.

A hypothetical student who merely knows the above equations and has worked through their mathematical consequences can hardly be said to have an adequate understanding of gravity. She additionally needs to understand the consequences of these equations in the observable world; *how* they explain falling objects in the everyday setting; the weight of objects, as perceived and as measured on scales of various designs; the motion of planets in the solar system; the tides; and so on.

None of these observations in itself validates the entire Newtonian theory of universal gravity; each corresponds to part of the theory, with some degree of indirectness. Measuring the time that an object takes to fall various distances gives indirect information about the acceleration, but none about the forces, the masses, or the distance to the center of the earth. Feeling the weight of an object being held gives fairly direct but very imprecise information about the force of gravity on the object (what you are directly experiencing is the normal force of the object on your hand). Using a spring scale gives indirect information about the weight of the object, in the form of the height of an indicator, mediated by the compression of a

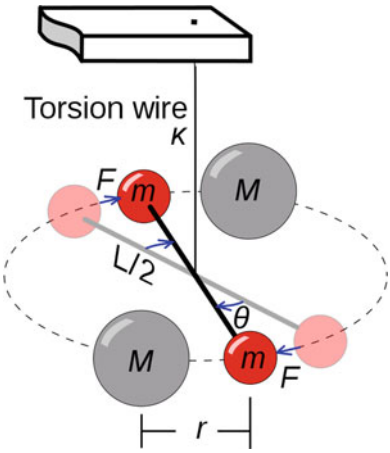


Fig. 1 Cavendish's experiment. Drawing by Chris Burks. From the Wikipedia article, "Cavendish Experiment"

spring. Using a balance scale gives information about the weight of the object being weighed as compared to standard weights mediated by the law of the lever. For the observations of the planets, which were the major source for the theory of gravity, the data was a record, over time, of the direction from the earth to the planets, the earth itself, of course, being a platform with a complex movement. It took the combined genius of Kepler and Newton to show how these measurements related to the equations of gravity, and even so, the astronomical observations did not give any information about the absolute distance of the planets. Until the observations became precise enough for the effect of one planet on another to be measurable, they likewise gave no information about the relative masses of the planets. The tides, correctly explained, are an effect of the spatial derivative of the gravitational force of the moon, as reflected, though a complex mechanism, in a twice-daily rising and falling of sea level at every sea-coast location.

As experiments and theories become more complex, the relation between the observations and the theory generally become more indirect, at least in some respects. Cavendish's experiment (Fig. 1) to determine the gravitational constant (from his point of view, to determine the mass of the earth), for the first time succeeded in creating a setting in which the masses and the distances could all be directly measured. But the measurement of the minuscule gravitational force created ($1.74 \cdot 10^{-7}$ N) is quite indirect: The torsion coefficient for the wire is determined by timing the oscillation period for the small balls twisting back and forth on the wire; the force needed to twist the wire is then calculated from the small angular deflection created.

The deeper the science, the more indirect the experimental evidence. The relation between Schödinger's equation and the experiments that support it are very indirect. You need to know a lot of physics to understand how gravitational wave detectors work or how the Higgs boson was detected.

At the other end of the spectrum, we have been speaking of measuring distances and time as though those were atomic percepts. But measurements rely

on measuring devices; measuring small distances requires an accurately calibrated ruler, and measuring durations of time requires a clock. Designing high quality rulers (see for instance Berger [5, pp. 116–120]) or clocks requires some physics and some engineering. (The foundations of theories of measurement is analyzed in [83].)

Moreover, measurements are taken separately, and the experimenter assumes that they remain [close to] constant from one stage of the experiment to the next. In the Cavendish experiment, you first measures the torsion coefficient using oscillation; and then you assume that same coefficient is valid when you are measuring the gravitational force between the balls. You first weigh the balls on a scales and then place them in the apparatus. We are thus drawing on a basic commonsense understanding of world in reasoning about the experiment, but we also know that that the commonsense view is insufficient.

Therefore, in PAVEL's encoding of the relation of Cavendish's experiment to the law of universal gravitation, the statement of the law of gravity is only a small part of the physics knowledge that you need, and the final actual measurements—the masses of the objects, the length of the rod, the oscillation period, and the displacement of the balls—are only a very small part of the description of the situation. Most of the knowledge of physics—the relevant part of h, in our Bayesian formulation—has to do with the properties of parts of the apparatus: most obviously, that the wire will exert a force against twisting proportional to the angle of twist, but also that the rod remains (reasonably) straight, that the masses of the balls remain (close to) constant in between being weighed and being placed in the apparatus. Almost all of the data—the relevant part of D in the Bayesian formulation—is a description of the design of the apparatus and of the procedure followed. The representation of the procedure must, at least implicitly, characterize all the things you *didn't* do in the course of the experiment: you didn't cut the rod shorter after measuring it or chop a chunk out of the balls after weighing them.

Moreover, a *full* description of the experiment should in principle include a description of the measurement apparatus and how it is used. The oscillation period was 20 min; but what kind of clock did you use? The small balls weigh 1.61 pounds; but what kind of scales did you use? Life being finite, the regress here cannot be infinite; and it would seem to bottom out, partly in systems of circular support (e.g. two independent rulers or clocks confirm one another), partly in direct perception (e.g. the ticks on the ruler look equally spaced), partly in some physical assumptions in the reasoning system that are made and not justified (e.g. that the masses do not change between being weighed and being put in the experimental set-up); and, at the individual level, in trust in the scientific community.

This last issue of trust is a major epistemic difference between mathematics and physics. In principle, a mathematician can check the proof of every theorem she is using; in practice, mathematicians do work through the proofs of many of the basic results in their area, and, even in our time, some mathematicians are known for their care in checking the proofs of the theorems they use [62]. By contrast, a physicist must trust both that the suppliers of scientific equipment are not sabotaging her lab by sending her defective instruments and equipment, and

that other physicists are accurately reporting their experimental results. Even in principle, a scientist cannot rerun all the experiments that underlie her theory. Some require unique equipment (the Hubble telescope, the CERN accelerator); others, such as astronomical observations, can only be made from particular locations at particular times (or must be made at multiple locations simultaneously). In mathematics, the communal aspect is important [61]; in physics and the other sciences, it is inescapable.[7]

To calculate the mass of the earth, Cavendish additionally needed to know the radius of the earth, which, at least in Cavendish's time, in turn was based on all kinds of *geographic* knowledge—knowing the north-south distance between two cities and comparing the angle of shadows at noon on the same day, and such. (The radius of the earth is also one important starting point for much of the knowledge of astronomical distances.) One doesn't necessarily think of pacing out the distance from Cyrene to Alexandria as a physics experiment, but these measurements certainly have implications for physics, and they are all part of the data D in our Bayesian formulation.

From the standpoint of the foundational approach, all this information consists of rules for translating human-scale realities into boundary conditions. That seems like an strange characterization, but, in the foundational approach, there is nothing else that it can be, as far as I can see. There are the differential equations, which are the foundational dynamics laws, and then there are the boundary conditions, and there is no room for anything else to enter in.

4.6 Is the Complexity of Grounding Different in Physics than Math?

I have argued that, in our reasoning system, the fundamental laws can only be a small part of the content. One might respond that an analogous situation holds with mathematics. Only a very small part of the mathematical knowledge of a mathematical proof verifier consists of the base ZFC axioms of set theory; most of the content is the definition of more complex mathematical concepts—the real numbers, the Gamma function, the regular dodecahedron, the class of NP-complete languages, Lie algebras, and so on—as set-theoretic constructions. Similarly, we could start with the foundational elements of physics, define things like the Cavendish experiment as a construction over the foundational elements, and then prove the behavior of the experiment from the foundational laws.

In principle, this is presumably possible; in fact, as we will discuss in Sect. 4.7, it is an important principle of physics that in principle this is possible. In practice,

[7] Large levels of trust are needed in any such enterprise. That is why conspiracy theorists, who are willing to distrust any evidence that runs against their theory, are so crazy and so unanswerable; and why any violation of trust—by scientists, by technologists, by the media—is so damaging, not just to the specific instance, but to the entire scientific/technological enterprise.

however, it is so far from being possible as to be not worth discussing. In mathematics the reduction to set theory is reasonably straightforward; any mathematician could work out the set-theoretic definition of the Gamma function and the rest of them, perhaps occasionally looking up some forgotten definition in Wikipedia or MathWorld. By contrast, characterizing the internal structure of the wire in Cavendish's experiment in terms of the atomic structure of its material, and proving that when twisted it exerts a restoring force proportional to the angle (rather than, for example, breaking, deforming, disintegrating, exerting a negative force, or exerting a force that is non-linear in the angle, within the angle range under discussion) are extremely difficult. We will discuss this issue of argumentation further in Sect. 4.8.

In mathematics, one sometimes gets out of these difficulties by positing the properties that you want; define a "Cavendish wire" to be one that, on twisting, exerts a restorative force proportional to the angle of twist, and then define the Cavendish experiment as using a Cavendish wire. But in this context, that doesn't help; we now have to prove that there exist Cavendish wires, and that the wire that was actually used in the experiment is a Cavendish wire.

4.7 Claims to Universality

A distinguishing feature of physics, as compared to other disciplines, is that it makes claims to universality of a certain kind. Specifically, physics makes one very general universal claim, which I will get to, but it also makes a number of more limited, but still very broad claims. Let me discuss a few, in increasing order of generality.

Historically, perhaps the first important finding of this kind was Laplace's successful explanation of all the motions of the planets then known in terms of Newton's law, which he published in his five-volume opus *Méchanique Céleste* (1799–1825). (The precession of the perihelion of Mercury, which requires general relativity, was reported by Le Verrier in 1859.)

Second: In chapter 1 of his *Lectures on Physics*, Feynman [28] wrote,

> If, in some cataclysm, all of scientific knowledge were to be destroyed and only one sentence passed on to the next generation, what statement would contain the most information in the fewest words? I believe that it is the *atomic hypothesis* . . . that *all things are made of atoms* — little particles that move around in perpetual motion, attracting each other when they are a little distance apart, but repelling upon being squeezed into one another.

As further confirmation of the centrality of the atomic hypothesis, we may note that the reality of atoms was a matter of fierce debate in the late nineteenth century and the first two decades of the twentieth, with Mach and others vehemently arguing that they were just a theoretical construct. The establishment of the physical reality of atoms, by Jean Perrin, Einstein, and others, was one of the major accomplishments of the early part of the twentieth century (less well known than relativity or quantum theory, because it was the consolidation of an established doctrine rather than a revolutionary new one).

Though fundamental, atoms are not on the tee-shirt; you will not get rich selling tee-shirts reading "All things are made of atoms". They are also not foundational, in the current view of things; an atom is the lowest energy state solution to the quantum electrodynamical equation describing a system with k electrons orbiting a nucleus with k protons. Atoms are not universal; there are no atoms in neutron stars. What Feynman's rather vague "all things" means is, presumably, "all matter within the terrestrial setting".

The fact that the atoms are fundamental but not foundational is not, in itself, an argument against grounding our reasoning system in foundational theories. One might say the same of the construction of real numbers from set theory. Real numbers predate infinite sets, certainly historically, almost certainly cognitively; and I know many mathematicians who, faced with Feynman's hypothetical cataclysm, would much prefer that mankind remember the reals rather than remember ZFC.

A third universalizing statement seems to me important, though difficult to state precisely. (This is discussed, in a somewhat more limited form, in [56].) The claim is more or less this: Taking the influx of radiation from outside earth to be an exogenous boundary condition, practically all physical events and physical properties of things that people encounter on earth are consequences of the earth's gravity together with non-relativistic quantum mechanics (Schrödinger's equation) applied to the electromagnetic interactions of atomic nuclei and electrons. There are some number of exceptions—the tides, the occasional meteor, radioactive decay, the things that happen inside sophisticated physics experiments—but those are largely known, and otherwise it is a very reliable rule. That is, if you make some physical observation or encounter a physical phenomenon, whether in meteorology, earth science, biology, chemistry, material science, or whatever, then it is overwhelmingly likely that this is a consequence of these two theories. The presumption is that it would not be necessary to invoke quantum chromodynamics, or the weak force, let alone to posit physical processes or entities previously unknown to physics. Moreover, these theories are mathematically simple: the equation of terrestrial gravity is extremely simple, and the necessary quantum mechanics, "can be written down simply and is completely specified by a handful of known quantities: the charge and mass of the electron, the charges and masses of the atomic nuclei, and Planck's constant" [56].

The final statement is completely universal. The claim is that anything in the universe that happens, happens by virtue of physical changes to physical substances, governed by universal physical law:

> Schematically, physicalism can be thought of as the claim that the physical facts determine all the facts. . . . In developing a claim of this sort, we need to do two things: first provide some dependence relation that explicates the thought that one set of facts "determines" another; second, decide what kinds of facts are to count as physical. Physicalist positions have been articulated in terms of a variety of dependence relations, including supervenience (there can be no change without physical change), realization (non-physical properties are s second-order, properties of physical properties), and token identity (everything (concrete) that instantiates a non-physical property also instantiates a physical property, to name but a few. . . . [T]he causal level must be "causally closed" with respect to the higher level; there is no "downward causation" from the higher level to the lower level [41].

Laplace's finding is easily expressed in a logical system; one simply states that Newtonian gravitation exactly characterizes the motion of the planets. I can see, more or less, how to represent Feynman's atomic hypothesis; one can state that all solids, bodies of liquids, bodies of gas, and unions of these are a set of atoms; or that the mass of all the matter within a given spatial region at a time is equal to the sum of the masses of the atoms. However, I have no idea how to formalize the latter two statements. Indeed, their logical status is not clear to me; I don't know whether they are statements *in* physics or meta-level statements about physics or heuristics for carrying out research in physics.

However, it does seem that there can be experimental *evidence* for these claims. For example, Rubner's 1894 demonstration that conservation of energy holds within a dog is an important experiment for *physics*, because it demonstrates that the principle holds for living creatures, which is not obvious on the face of it. More generally, the justification for these two claims rest on an enormous body of experimental evidence showing the profound regularities in chemical behavior, material behavior, biochemical behavior, and biological behavior; and the theoretical analysis and experimental evidence that demonstrates, as far as it goes, that chemical and material behavior can be explained in terms of physics, that biochemistry can be explained in terms of chemistry, and that biology can be explained in terms of biochemistry. Conversely, any phenomenon that is puzzling and not explained, such as the reversal of the earth's magnetic field, is necessarily to some extent evidence against the claims. (In a Bayesian theory, if a positive outcome is evidence for a claim, then necessarily a negative outcome is evidence against it.) All of these are, in principle, part of the data D to be considered.

Also, it seems to me, these claims indicate that Occam's razor, as used by physicists, involves something more than just the minimum description length principle. When you make a new experimental finding, then the MDL principle gives you brownie points (so to speak) if you can explain it in terms of known laws of physics, because you can use that to compress the description of the data. That in turn translates back into a increased probability for those laws and hence into predictive power. But I don't see any justification for the MDL principle giving you brownie points as a reward for speculating that the new findings ought to somehow be explicable using known laws of physics.

4.8 Argumentation in Physics

From the AI perspective, the difficulties discusses above are mostly problems of *representation*. Even greater are the difficulties of *reasoning*—how one can characterize an argument and implement the validation of arguments in a computer program.

Rigorous mathematical proof consists entirely on deductive reasoning: The conclusion is a logically necessary consequence of the assumptions. In actual mathematical discourse, there are certainly informal arguments, but, as discussed

at the start of this article, the great discovery that powers verification technology is that, in the vast body of math that is considered rigorously proved, it is possible to eliminate all informal, "hand-waving" arguments and to fill in all logical gaps.

However, such an undertaking does not seem to be close to possible in physics. Unlike math, it is not possible to ground the reasoning about physical systems on the human scale in deductive inference from the foundational theories; the complexities are simply too large.

4.8.1 Deduction from the Absolute Foundations

One extreme form of the tee-shirt approach to PAVEL is to start from a minimal set of absolutely fundamental concepts and laws, and do everything deductively from there. This idea is demolished in [56]; I really cannot do better than to quote from them at a little length, and then I have nothing to add.

> We know that [the Schrödinger equation for electrodynamics] is correct ... But it cannot be solved accurately when the number of particles exceeds about 10. No computer existing, or that will ever exist, can break this barrier because it is a catastrophe of dimension. If the amount of computer memory required to represent the quantum wavefunction of one particle is N, then the amount required to represent the wavefunction of k particles is N^k. It is possible to perform approximate calculations for larger systems, and it is through such calculation that we have learned why atoms have the size they do, why chemical bonds have the length and strength they do, why solid matter has the elastic properties it does, why some things are transparent while others reflect or absorb light .... With a little more experimental input for guidance, it is even possible to predict atomic conformations of small molecules, simple chemical reaction rates, structural phase transitions, ferromagnetism, and sometimes even superconducting transition temperatures .... *But the schemes for approximating are not first-principles deductions but are rather art keyed to experiment* [emphasis added] and thus tend to be the least reliable precisely when reliability is most needed, i.e. when experimental information is scarce, the physical behavior has no precedent, and the key questions have not yet been identified. There are many notorious failures of alleged *ab initio* computation methods, including the phase diagram of liquid ^{3}He and the entire phenomenology of high-temperature superconductors .... Predicting protein functionality or the behavior of the human brain from these equations is patently absurd.

This is from 2000; certainly we can now compute much more than we could 18 years ago, and for all I know, some of the specific examples that Laughlin and Pines mentioned may be outdated.[8] Moreover, these kinds of calculations may be a good fit for quantum computing, when that technology becomes practical. But as far as I can determine, the general point still holds, and will continue to hold for the foreseeable future.

[8] Hendry [41] similarly argues that molecular structures cannot be calculated from Schrödinger's equation. Rather, given the structure, it is possible to use quantum mechanics to calculate various physical values.

It would certainly be immensely desirable to include in PAVEL the kinds of arguments based on "art keyed to experiment" that connect quantum theory to the many phenomena mentioned by Laughlin and Pines. However, I do not have the knowledge to discuss the logical structure of these arguments or what would be involved in incorporating them into PAVEL.

4.8.2 Argumentation in Elementary Physics

Let me return to the level of physics that I understand. In arguments that use elementary physics to analyze real-world situations, we can characterize a variety of non-deductive forms of reasoning:

- **The closed world assumption.** It is assumed that everything that will affect the outcome of the experiment has been accounted for.
- **Ignoring irrelevant issues.** A description of Cavendish's experiment need not specify the geographic location where the experiment was performed. (By contrast, the latitude is critical in a description of Foucault's pendulum.)
- **Ignoring small quantities.** In some cases, the value of some small quantity is known, or can be bounded, and it is assumed without proof that, because it is small, its impact on the analysis is small. In other cases, the value of a quantity is not known with any precision, but it is assumed to be small and further assumed to have a small impact on the analysis.
- **Approximation.** "Assume a spherical cow" as the old joke says. Surfaces are taken to be flat, densities are taken to be uniform, resistances are taken to be linear, and so on.

  Certainly approximation, and order-of-magnitude reasoning which is similar, can sometimes be carried out deductively. If an upper bound on the inaccuracy of the approximation is known, it may be possible to answer Boolean questions with certainty or to give an upper bound on the inaccuracy of numerical calculation. In a probabilistic setting, if an upper bound on the variance is known, then it may be possible to compute a lower bound on the certainty of the answer to a Boolean question or an upper bound on the variance of a numerical answer.
- **Idealization and abstraction.** Almost every analysis of a physical situation idealizes the entities involved and abstracts the relations between them. One reasons about a physical electronic circuit in terms of a circuit diagram. In a mechanics problem, a string is taken to be massless and one-dimensional. Continuum mechanics is an abstraction of the actual particles structure of matter.

  Moreover, a single argument may use multiple idealizations of the same thing. Analyses of chemical reactions, for example, will often combine an molecular model of substances, to describe the reaction, with a continuous model, or multiple continuous models, to describe the fluid mechanics and thermodynamics. An analysis of the tides caused by a planet's moons might well first calculate the moon's orbit approximating the planet as a point mass, and then use the planet's extent and material composition in calculating the tidal effects.

The same physical object and even the same physical situation may have many different possible models, depending on what is the range of behaviors under consideration, the accuracy desired, and the measurements being made. Consider a pendulum on a string. You have the following choices, among others [4].

- The setting can be two-dimensional or three-dimensional. It can even be one-dimensional, if you simply set up the problem in terms of the Lagrangian $\mathcal{L}(\theta) = m(r\dot{\theta})^2/2 + mgr\sin(\theta)$, where θ is the angle from vertical downward.
- The bob can be a point mass, a circle or sphere, or a more complex shape.
- There are many different options for the string:
  - It can be considered like a rod, holding the weight at a fixed distance from the attachment point; or a hard constraint maintaining an upper bound on the distance from the bob to the attachment point; or a soft constraint, exerting an elastic force when stretched beyond a fully extended position. In the Lagrangian formulation mentioned above, the string is completely abstracted away, into the formulation of the energy function.
  - It can be one dimensional or three dimensional.
  - It can be massless or massed.
  - It can bend along its length or twist along its axis or both.
  - It can be immutable, or it can snap, or it can be cut.
- Dissipative forces can include air resistance or friction at the attachment point or both; various kinds of approximations can be used.
- The frame within which the pendulum is set up can be fixed, or it can be attached to a rotating earth. This option would hardly cross one's mind, except that it is critical in Foucault's pendulum.
- Gravity can be a uniform field, or a Newtonian field, or follow general relativity

Different circumstances call for different idealizations. A problem in a freshman course would probably use a two-dimensional setting, a point mass, and a string of fixed length. A problem in an advanced mechanics class might simplify the analysis to a one-dimensional Lagrangian formulation or might complicate it by positing an extended mass or a three-dimensional setting. Cavendish's experiment requires a three-dimensional setting, an extended bob, (the two weights on the rod) and a cord that twists along its axis. Foucault's pendulum requires a three-dimensional setting, a cord of fixed length, and a frame attached to the rotating earth. Smith et al. [78] describe a psychological experiment in which subjects were ask to predict the trajectory of a pendulum if its cord is cut in mid flight; this requires a fixed length string that can be cut. Reasoning about a yo-yo requires an extended object and a flexible one-dimensional string. Reasoning about a cord swinging freely requires a one-dimensional string with constant density. The pendulum in a grandfather clock is connected to a mechanism that adds energy at every swing. In the Poe story, "The Pit and the Pendulum", the cord is a brass rod, the bob is "a crescent of glittering steel, about a foot in length from horn to horn; the horns upward, and the under edge evidently as keen as that of a razor;" and the frame gradually descends.

It is tempting to propose that one should always use the most detailed possible model. But this is hardly feasible; not only does the complexity of calculations go up rapidly, but, more seriously, so does the kind of information needed. If you approximate a cord in a pendulum as a distance constraint, all you need to know is its length; if you want a detailed model you need to know additionally its radius and its material characteristics. In a given situation, these may be unspecified or hard to determine. (Again, of course, the Bayesians will tell you blithely that, if you don't know them, you should use a probability distribution over the range of values.)

4.9 Reasoning About Things That Are Partially Understood

Physical reasoning can be applied to phenomena that are only partially understood, such as plate tectonics, the planetary magnetic fields, the million-degree temperature in the sun's corona, and lightning [27]. Feynman [28] book-ends his volume-long textbook on electromagnetism as follows:

> [End of chapter 1] Let us end this chapter by pointing out that among the many phenomena studied by the Greeks, there were two very strange ones: that if you rubbed a piece of amber, you could lift up little pieces of papyrus, and that there was a strange rock from the island of Magnesia which attracted iron. It is amazing to think that these were the only phenomena known to the It is amazing to think that these were the only phenomena known to the Greeks in which the effects of electricity or magnetism were apparent. [Feynman seems to have forgotten lightning.] (Feynman [28, end of chapter 1])

> [End of chapter 37] We now close our study of electricity and magnetism. In the first chapter we spoke of the great strides that have been made since the early Greek observations of the strange behavior of amber and of lodestone. Yet in all our long and involved discussion, we have never explained why it is that when we rub a piece of amber we get a charge on it nor have we explained why a lodestone is magnetized . . . So you see this physics of ours is a lot of fakery — we start out with the phenomena of lodestone and amber, and we end up not understanding either of them very well [end of chapter 37]

Almost 60 years later, these phenomena are certainly *better* understood, but none are perfectly understood; in particular the triboelectric effect, in which rubbing one material with another creates an electric charge "is not very predictable" (Wikipedia, triboelectric effect). Nonetheless, a lot of physical reasoning about these *is* possible, through a combination of fundamental principles, experimental evidence, approximations, and speculative reconstruction of structure and mechanisms. Such explanations typically fail to match observed reality in some respects or fail to distinguish the circumstances where the phenomenon occurs from those where it doesn't. Nonetheless, these explanations are considered valid as far as they go; no one seriously proposes that these phenomena are evidence of a fundamental physical process that lies outside of the known fundamental theories.

5 An Example Word Problem

To illustrate what would be involved in formalizing simple physics reasoning in PAVEL, we present a formalization of the following simple word problem:

> **Problem:** A 1 kg pendulum bob on a 1 meter inelastic string is dropped from the point 0.5 meters directly to the right of the attachment point. How long will it take to reach a point directly below the attachment point? What force will the string be exerting on the bob at that point? (Fig. 2).

Table 3 shows the sorts and the sortal operators. Table 4 shows the language of geometry and kinematics used. Table 5 shows the theory associated with string. Table 6 shows the dynamic laws of physics. Finally Table 7 shows the formulation of the problem. What is missing here is the purely mathematical theory (the theory of the reals, vector algebra, and vector calculus); the axioms governing the relation between dimensions; and the purely kinematic theory.

The formalization in Tables 3, 4, 5, 6, and 7 should be largely self-explanatory, but a few points require explanation.

A "pseudo-object" (introduced in Davis [13]) is a geometrical feature that moves around with an object: The center of a spherical object, the surface of an object, the apex or base of a cone, the hole in a donut, and so on. In this case, we mark the two ends of the string as pseudo-objects.

The long-winded and unappealing symbols that we have used for vector and function operators—`PointPlusVec(x,v)` instead of simply $x+\vec{v}$, and so on—are there in order to keep our system of sorts simple. Standard mathematical notation, and many math-oriented programming languages such as MATLAB,

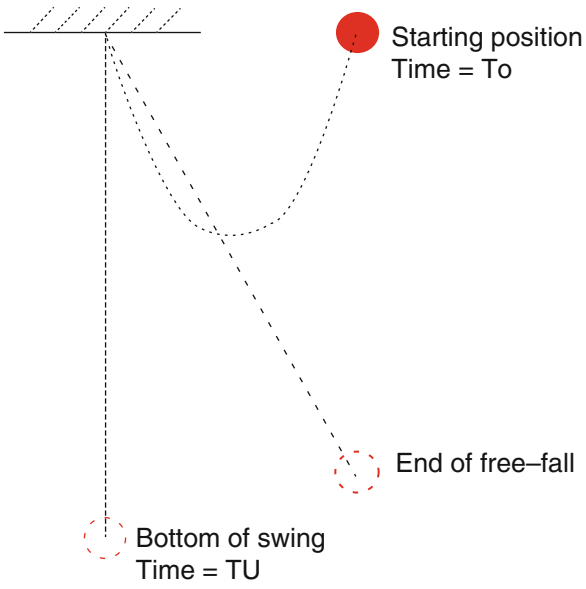


Fig. 2 Dropping a pendulum on a string

Table 3 Physics word problem: sorts

Sorts: *Object, String, PseudoObject, Time, Point, Real, Duration, Distance, Speed, Acceleration, Mass, Force*
Sortal Functions:
Fluent[σ]—Function from *Time* to sort α.
Vector[α]—If α is a real-valued dimension, then a vector of dimension α.
$\alpha \otimes \beta$—Infix operator: Dimension α times dimension β.
$\alpha \oslash \beta$—Infix operator: Dimension α divided by dimension β.

Notes: *Acceleration, Force,* and *Momentum* are here scalar dimensions, not vectors
Here and in the tables below, sortal variables σ and τ range over all sorts; variables α and β range over the additive dimensions (i.e. real-valued dimensions with a natural sense of zero and of addition) *Real, Duration, Distance, Speed, Acceleration, Mass,* and *Force*

enormously overload standard symbols such as '+' and '·'. In a practical implementation of PAVEL, this might end up being worthwhile; for this simple example, it seemed better to keep the sorting system simple and burden ourselves with separate symbols.

Since the velocity of the bob is discontinuous at the moment when the end of the string is reached, we define the velocity of an object **o** before time `t` to be the limit of the derivative of its position at time $\mathtt{t}'$ as $\mathtt{t}' \rightarrow \mathtt{t}^-$ and the velocity after `t` analogously. (The definition would be included in the kinematic axioms, not enumerated here.)

In Sect. 4.5 we raised the issue of the assumption that masses and so on remain constant from one stage of an experiment to another. In our formalization here, we have unabashedly cheated on all such concerns by using time-independent symbols for every quantity or relation that does not change over time in this particular problem. For instance `MassOf(o)` is presumed to be a time-invariant property of an object `o`; `Attached(o,q)` is assumed to be a time-invariant relation between object `o` and pseudo-object `q`; and so on.

We use a simple theory of non-elastic, one-dimensional, massless strings, governed by the following rules, enumerated in Tables 5 and 6

- A string has two ends (axiom S.1) which cannot be more than a fixed distance apart (the length of the string) (axiom S.2).
- The end of a string may be *attached* to a single point object, or it may be *fixed* in space (presumably actually attached to some fixed frame, but we did not include the frame in our formulation here) (axiom S.3). If it is attached to an object, then the object and the end of the string are always at the same point (axiom S.4). If the end of the string is fixed, then it is always at the same point (axiom S.5).
- A string is *taut* if both ends are either attached or fixed, and if it is fully extended; that is, the distance between the two ends is equal to the length of the string (axiom S.6)
- An inelastic event involving the string, called a *yanking* occurs when the string is taut, and the difference in velocities between the two ends has a positive component in the direction from one end to the other (axiom S.7). Note that if

Table 4 Physics word problem: geometric and kinematic primitives

Constant Symbols:
`Meter` → *Distance*.
`Second` → *Duration*.
`X` → *Vector[Real]*. Horizontal dimensionless unit vector.
`Z` → *Vector[Real]*. Vertical dimensionless unit vector.
Function Symbols:
`Dist(pa:Point, pb:Point)` → *Distance*.
Distance between *Point*s `pa` and `pb`.
`Magnitude(v:Vector[α])` → α. The magnitude $\|\vec{v}\|$.
`PointPlusVec(p:Point, v:Vector[Distance])` → *Point*
The sum $\mathbf{p} + \vec{v}$ of point $\mathbf{p}$ plus vector $\vec{v}$.
`VecFrom(pa:Point,pb:Point)` → *Vector[Distance]*.
Vector `pb` − `pa` where `pa` and `pb` are *Point*s.
`VecMinus(u:Vector[α], v:Vector[α])` → `Vector[α]`. Vector $\vec{v} - \vec{u}$.
`ScalarTimesVec(x:α, v:Vector[α])` → *Vector*$[\alpha \otimes \beta]$. The scalar product $x \cdot \vec{v}$.
`DotProd(u: Vector[α], v:Vector[β])` → $\alpha \otimes \beta$. Dot product $\vec{u} \cdot \vec{v}$.
`Direction(v:Vector[α])` → *Vector[Real]*.
Dimensionless direction of $\vec{v}$. $\vec{v}/\|\vec{v}\|$.
`V(t:Time, q:Fluent[α])` → α. Value of fluent `q` at time `t`.
`VelocBefore(p:Fluent[Point])` → *Fluent[Vector[Speed]]*.
Derivative of $\vec{x}(t)$, where $\vec{x}$ is a *Point*-valued *Fluent*, evaluated from the left (see text.)
`VelocAfter(p:Fluent[Point])` → *Fluent[Vector[Speed]]*.
Derivative of $\vec{x}(t)$, where $\vec{x}$ is a *Point*-valued *Fluent*, evaluated from the right.
`DerivOfVeloc(p:Fluent[Vector[Speed]])` → `Fluent[Vector[Acceleration]]`.
First time derivative of $\vec{v}(t)$, where $\vec{v}$ is a velocity, evaluated from the left.
`OPlace(o:Object)` → *Fluent[Point]*.
The fluent tracking the location of *Object* `o` over time.
`QPlace(q:PseudoObject)` → *Fluent[Point]*.
The fluent tracking the location of *PseudoObject* `q` over time.
Predicate symbols:
`Zero(x:α)`—Scalar x has zero value.
`Positive(x:α)`—Scalar x is positive.
`Continuous(p: Fluent[Point]t :Time)`.
Point-valued *Fluent* `p(t)` is a continuous function of time in a neighborhood of time `t`
`TwiceDifferentiable(p:Fluent[Point], t:Time)`.
Point-valued *Fluent* `p(t)` is twice differentiable in a neighborhood of time `t`.

difference is orthogonal to the direction, as in the case when the string is swinging in a circle, that is not considered a yanking.

- If an object is attached to an end of a string undergoing a yanking then it is said to be yanked (axiom S.8)
- A string exerts no force if it is not taut (axiom P.5)

Table 5 Theory of strings

Function symbol:

`Length(s:String)`. Length of *String* `s`.

Predicate Symbols:

`EndOf(q:PseudoObject, s:String)`. *PseudoObject* `q` is an end of *String* `s`.

`Fixed(q:PseudoObject)`. *PseudoObject* `q` is fixed in position.

`Attached(q:PseudoObject, o:Object)`. *Object* `o` is attached to *PseudoObject* `q`.

`Taut(s: String:, t:Time)`. *String* `s` is taut at time `t`.

`Yanking(s:String, t:Time:`. At time `t`, *String* `s` yanks the objects attached to it. See text.

`Yanked(o:Object, t:Time)`. At time `t`, *Object* `o` is yanked by some string it is attached to.

Axioms:

S.1. $\forall_{s:String}\ \exists_{qa,qb}$ `EndOf(qa,s)` $\wedge$ `EndOf(qb,s)` $\wedge$ `qa` $\neq$ `qb` $\wedge$
$[\forall_{qc}$ `EndOf(qc,s)` $\Longrightarrow$ `qc=qa` $\vee$ `qc=qb`.$]$

Every string has exactly two ends.

S.2. $\forall_{t:Time;s:String,qa,qb:PseudoObject}$
`EndOf(qa,s)` $\wedge$ `EndOb(qb,s)` $\wedge$ `qa` $\neq$ `qb` $\Longrightarrow$
`Distance(V(t,QPlace(qa)), V(t,QPlace(qb)))` $\leq$ `Length(s)`.

The distance between the end of a string is at most the length of the string.

S.3. $\forall_{s:String;q:PseudoObject;oa,ob:Object}$ `EndOf(q,s)` $\wedge$ `Attached(oa,q)` $\wedge$ `ob` $\neq$ `oa` $\Longrightarrow$
$\neg$`Attached(ob,q)` $\wedge$ $\neg$`Fixed(q)`.

S.4. $\forall_{o,q}$ `Attached(q,o)` $\Longrightarrow$ $\forall_{t}$ `V(t,OPlace(o)) = V(t,QPlace(q))`.

If *Object* `o` is attached to end `q` of a *String* then `o` and `q` are always in the same place.

S.5. $\forall_{q}$ `Fixed(q)` $\Longrightarrow$ $\forall_{ta,tb:Time}$ `V(ta,QPlace(q)) = V(tb,QPlace(q))`.

A fixed end of a string is always in the same place.

S.6. $\forall_{s:String;t:Time}$ `Taut(s,t)` $\Leftrightarrow$
[[$\forall_{q}$ `EndOf(q,s)` $\Longrightarrow$ `[Fixed(q)` $\vee$ $\exists_{o}$ `Attached(q,o)]]` $\wedge$
`[Distance(V(t,QPlace(qa)), V(t,QPlace(qb))) = Length(s)]]`.

Definition: A string is taut at time `t` if both ends are either fixed or attached to an object and the distance between the ends is equal to its length.

S.7. $\forall_{s:String;t:Time}$ `Yanking(s,t)` $\Leftrightarrow$
`Taut(s,t)` $\wedge$
$\exists_{qa,qb}$ `EndOf(qa,s)` $\wedge$ `EndOf(qb,s)` $\wedge$
`Positive(DotProd(VecMinus(V(t,VelocBefore(Place(qa))),`
`V(t,VelocBefore(Place(qb))))`
`VecFrom(V(t,Place(qa)),V(t,Place(qb)))))`.

Definition: String `s` is yanking at time `t` if it is fully extended at `t`, and if the velocity of the two ends at time `t` are such that it would be overextended if they continued in their motion.

S.8. $\forall_{o:Object;t:Time}$ `Yanked(o,t)` $\Leftrightarrow$
$\exists_{s,q}$ `Attached(o,q)` $\wedge$ `EndOf(q,s)` $\wedge$ `Yanking(s,t)`.

Object `o` is yanked at *Time* `t` if it is attached to some *String* that is yanking.

Table 6 Physics word problem: laws of physics

Constant symbol:
`Kilogram` → *Mass*
Function Symbols:
`MassOf(o:`*Object*`)` → *Mass*. The mass of *Object* `o`.
`GravForceOn(o:`*Object*`)` → *Fluent[Vector[Force]]*. The gravitational force on *Object* `o`.
`ForceOn(oa:`*Object*`, ob:`*Object*`)`. The force executed on `oa` by `ob`.
`TotalForceOn(oa:`*Object*`)` → *Fluent[Vector[Force]]*. The total force executed on `oa`.

Axioms:

P.1. $\forall_{o:Object;t:Time}$ `Continuous(OPlace(o),t)`.
Objects move continuously.

P.2. $\forall_{o:Object;t:Time}$ $\neg$`Yanked(o,t)` $\Longrightarrow$
`TwiceDifferentiable(OPlace(o),t)` $\wedge$
`ScalarTimesVec(Mass(o),V(t,DerivOfVeloc(VelocityBefore(OPlace(o)))))` = `V(t,TotalForceOn(o))`.
Newton's second law, except when there is an impulse from a string.

P.3. $\forall_{o:Object;t:Time}$ `GravForceOn(o,t)` =
`ScalarTimeVec(`-9.8 *`* MassOf(o) * Meter/(Second * Second), Z)`*.
Terrestrial gravitational force.

P.4. $\forall_{o:Object;s:String;qa,qb:PseudoObject;t:Time}$
`Attached(o,qa)` $\wedge$ `EndOf(qa,s)` $\wedge$ `EndOf(qb,s)` $\wedge$ `Fixed(qb)` $\wedge$
`Yanking(s,t)` $\Longrightarrow$ `V(t,VelocAfter(OPlace(o)))` =
`VecMinus(V(t,VelocBefore(OPlace(o))),`
`ScalarTimesVec(DotProd(V(t,VelocBefore(OPlace(o))),`
`Direction(VectorFrom(V(t,QPlace(qb)),`
`V(t,QPlace(qa))))),`
`Direction(VectorFrom(V(t,QPlace(qb)),`
`V(t,QPlace(qa))))))`.
When an object "collides" with the end of a string and the other end is fixed, then the velocity after the collision is the component of the velocity before the collision in the direction tangent to the taut string.

P.5. $\forall_{o:Object;s:String;t:Time}$ $\neg$`Taut(t,s)` $\Longrightarrow$ `Zero(V(t,ForceOn(s,o)))`.
If a string is not taut, it is not exerting any force.

P.6. $\forall_{o:Object;s:String;qa,qb:PseudoObject;t:Time}$
`Taut(t,s)` $\wedge$ `EndOf(qa,s)` $\wedge$ `EndOf(qb,s)` $\wedge$ `qa` $\neq$ `qb` $\wedge$
`Attached(o,qa)` $\Longrightarrow$
`[Zero(Magnitude(V(t,ForceOn(s,o))))` $\vee$
`Direction(V(t,ForceOn(s,o)))` =
`Direction(VectorFrom(V(t,QPlace(qa)),`
`V(t,QPlace(qb))]`.
The force exerted by a taut string on an object attached at one end is parallel to the direction to the other end.

Table 7 Physics word problem: problem formulation

`B`→*Object*—The bob.
`S`→*String*—The string.
`QA`→*PseudoObject*—The end of the string attached to `B`
`QB`→*PseudoObject*—The fixed end of the string.
`T0`→*Time*—The initial time.
`TU`→*Time*—The time when the bob is directly under the attachment point.
Axioms:
F.1. `EndOf(QA,S)` $\wedge$ `EndOf(QB,S)` $\wedge$ `QA` $\neq$ `QB`.
F.2. `Attached(B,QA)`.
F.3. `Fixed(QB)`.
F.4. `Length(S) = Meter`.
F.5. `MassOf(B) = Kilogram`.
F.6. `V(T0,OPlace(B)) = PointPlusVec(V(T0,QPlace(QB)),ScalarTimesVec(0.5*Meter,X))`.
F.7. `Direction(VectorFrom(V(TU,QPlace(QB)),V(TU,OPlace(B)))) = ScalarTimesVec(−1, Z)`.
F.8. $\forall_{t:Time}$ `Direction(VectorFrom(V(t,QPlace(QB)),V(t,OPlace(B)))) = ScalarTimesVec(−1, Z))` $\Longrightarrow$ `t` $\geq$ `TU`. `TU` is the first time when the bob is below the attachment point.
F.9. $\forall_{t:Time}$ `V(t,TotalForceOn(B)) = V(t,ForceOn(S,B)) +V(t,GravForceOn(B))`. Closed world assumption: The only forces on the bob are gravity and the string.
Evaluate: `(TU−T0)`. **Evaluate:** `V(TU, ForceOn(S,B))`.

- A string that is taut and not yanking exerts on an attached object a non-negative force in the direction along the string (axiom P.6)
- If one end of a string yanks on an object, and the other end is fixed, then the velocity of the object changes discontinuously. Specifically, its velocity after the event is equal to the component of its velocity before the event in the direction orthogonal to the direction of the string (axiom P.4—there may well be some more elegant way to axiomatize this.)

Combining these with the statement (axiom P.2) that, when not yanked, the object obeys Newton's second law, these suffice to determine that, after falling vertically

to the length of the string, the bob will swing back and forth on a circle, and the centripetal force that the string exerts on the bob will be exactly what is needed to keep it on that path (the component of gravity in the direction of the string plus the centrifugal force). If the centripetal force were less then that, then the distance between the end of the string would be greater than the length, which is impossible; if it were greater, it would pull the bob within the circle; the string would cease to be taut, and the bob would instantaneously fall back, which is also impossible.

The rule for the changes in velocity if the string is attached to objects at both ends and a yanking event occurs is similar, but more complicated; it is not included here.

The problem formulation requires a closed world assumption (axiom F.9) that the total force on the bob is the sum of gravity and the force from the string. Almost any problem formulation in physical reasoning has to have some kind of closed world assumption, that states that everything that will interfere with the system has been accounted for. In this case, it would be better to have a general rule of physics that the total force on an object is the sum of the forces, and then to have the individual problem statement assert that the only forces on the bob are gravity and the string. However, that would require adding "sets of forces" as a sort and summation over sets as an operation, so we went with this simpler, less general, formulation instead.

In general, there is always a choice to be made about how general to make the formulation of the theory and how much to tailor it to the specifics of the problem at hand. If you have only a single problem in mind, then the decision is essentially stylistic: using a general representation makes the argument that the theory generalizes more plausible, using a more tailored one makes the exposition simpler. The more problems you address, the more is gained by generality, but it remains to some extent a matter of taste. (Tailoring the representation of a general theory to the specifics of one or a few problems violates the "no function in structure" rule of de Kleer and Brown [25]. On the other hand, if one is going to choose among idealizations the one that best fits the problem, as I have argued above, then that principle has been given up in any case.)

On the whole word problems in physics are simpler and more idealized than experimental set ups. A reasonable axiomatization of the Cavendish experiment at a comparable level of detail would probably be two or three times longer.

6 Historical Context and Related Work

There is a long history of work more or less along the lines of PAVEL. That history has three primary threads: in physics, in philosophy, and in AI. The physics and philosophy threads both largely begin with Hilbert; the AI thread is largely separate.

Corry [12] gives a very detailed account of the physics and philosophical work up through the work of Hilbert; I have not found a comprehensive review of the work since Hilbert.

6.1 Before Hilbert

Newton's *Principia* is substantially presented as deductions from axioms, in imitation of Euclid. In modern times Hertz's [43] *Die Prinzipien der Mechanik* was the first attempt to formulate the laws of mechanics in axiomatic form. It was notable for its exclusion of force as a fundamental concept, and using only time, space, and mass.

6.2 Hilbert's Sixth Problem and the Axiomatization of Physics

In Hilbert's famous collection of 23 mathematical problems, proposed at the 1900 International Congress of Mathematicians, number 6 was the axiomatization of physics [12].

> **Mathematical Treatment of the Axioms of Physics.** The investigations on the foundations of geometry suggest the problem: To treat in the same manner, by means of axioms, those physical sciences in which already today mathematics plays an important part; in the first rank are the theory of probabilities and mechanics.

Hilbert further explained:

> As to the axioms of the theory of probabilities, it seems to me desirable that their logical investigation should be accompanied by a rigorous and satisfactory development of the method of mean values in mathematical physics, and in particular in the kinetic theory of gases. ... Boltzmann's work on the principles of mechanics suggests the problem of developing mathematically the limiting processes, there merely indicated, which lead from the atomistic view to the laws of motion of continua.

In general, mathematicians have been unenthusiastic about Hilbert's sixth problems. It is very much an outlier among his 23 problems; whatever it is, it isn't mathematics,[9] and it is not at all clear what would count as a solution. Yandell [92], in his 400-page book on Hilbert's problems, dismissed the sixth problem in a mere four pages.

There seem to be three general projects involved in Hilbert's sixth problem.

First, the axiomatization of probability theory. This was accomplished by Kolmogorov, at least as far as the measure space interpretation goes. As discussed in Sect. 3.2, I am not convinced that the likelihood model, which permits probabilities of individual propositions, is axiomatized to the point that it supports analysis of real-world situations.

Second, the axiomatization of the foundations of physics; these, of course, were radically transformed in the three decades after Hilbert's speech. Hilbert himself

[9] Incidentally, the fact that Hilbert included this problem and spent a great deal of time working on it tells strongly against the common idea that Hilbert was a pure formalist, who viewed the meaning of mathematical symbols as unimportant [12].

devoted substantial research energy to the formulations of quantum theory and general relativity; he and Emmy Noether were in communication with Einstein about general relativity during the years that Einstein was developing the theory.

As best as I can ascertain, the current status is as follows:

- General relativity is completely axiomatized. It would be feasible to formulate the theory as axioms in a proof-verification system and to prove consequences such as the rotation of the perihelion of Mercury, the possibility of black holes, gravitational lenses, gravitational waves, and so on.
- Schrödinger's equation for non-relativistic quantum mechanics is easily axiomatized—it is just a partial differential equation—and its consequences can be proved, up to the limits discussed in Sect. 4.8.1. However, if one adds Born's law, which governs the probabilistic collapse of the wave function following an observation, then the situation becomes much less clear. As far as I can find, most so-called "axiomatizations" of quantum physics that include Born's law (e.g. [11, chap. 3]) are fine as regards the physics, but do not specify what is the probabilistic logic used (if that is necessary) or give a useful characterization of an observation, or state the independence assumptions. It is not clear to me that we are currently in a position to characterize axiomatically experiments whose outcomes depend on Born's law.[10] I do not know how severe a limitation that is; for example, how many, if any, of the explanations of phenomena enumerated in the above quote from Laughlin and Pines would be affected.

  Ludwig's [58, 59] *An Axiomatic Basis for Quantum Mechanics* develops an axiomatic theory, and, further, presents a metatheory of axiomatizations of physical theory. It includes an extensive, though very abstract, discussion of the relation between the theory and its macroscopic manifestations. Unfortunately, I am not at all in a position to evaluate what is the scope of what he accomplished; apparently the discussion is extremely difficult and relentlessly abstract, even for expert readers [87].

  Boender et al. [8] have Coq to verify protocols in quantum communication and quantum cryptography, but this is far from the physics experiments that we are discussing, and though it uses probabilities, it requires only a very limited theory.
- Quantum field theory is in a much less certain state; the axiomatizations that have been proposed, such as the Wightman axioms, have severe limitations. This remains an open problem.

Also, as is well known, finding a satisfactory theory that encompasses both general relativity and quantum theory is unsolved.

[10] It has been suggested to me that it will be easier to find a logical formulation of the "many-worlds" interpretation of quantum mechanics or, alternatively, the theory of quantum decoherence than the Copenhagen interpretation. That may be so; but I can't find that anyone has produced a logical formulation of either of these interpretations either.

Third, the explanation of continuum mechanics in terms of particle mechanics[11]; more generally, the explanation of macroscopic behaviors in terms of foundational theories. This is a more open-ended project, since there are several forms of continuum mechanics, and an open-ended collection of macroscopic behaviors.

One particularly important and difficult problem of this kind has been to complete the derivation of thermodynamics from statistical mechanics begun by Maxwell and Boltzmann. A recent study which develops a substantial formal foundation, is Wallace [88].

In general, it seems to me fair to say that what a physicist usually means by "axiomatization" is quite different from a mathematician means, and still more from what a logician means. When a physicist claims to have "axiomatized" a theory, what he/she generally has done is to have enumerated a set of foundational rules for an abstract theory which, generously supplemented by the physicists' own understanding of the concepts involved and by a variety of facts too obvious to be worth mentioning, will support various kinds of informal arguments. (Ludwig [58, 59] is certainly an exception.)

6.3 Philosophy

There is a long philosophical literature on axiomatizing physics, particularly particle dynamics, either in a strictly logical notation or in some other formalism. Some early work include part VII of Russell's [73] *The Principles of Mathematics*, a precursor to *Principia Mathematica*, entitled "Matter and Motion"; and Hamel [35, 36] *Elementare Mechanik* and *Grundbegriffe der Mechanik*. (Hamel was a student of Hilbert's)

In Vienna in the 1920s, a group of philosophers, mathematicians, and physicists called "The Vienna Circle" [76] embarked on a formidably ambitious project to investigate the foundations of science, called "logical positivism" or "logical empiricism". Following the models of Whitehead and Russell's [90] in *Principia Mathematica*, and of Wittgenstein's [91] *Tractatus*, they attempted to demonstrate that scientific theory could be built up logically from basic observations. They planned to produce a large series of books, the *International Encyclopedia of Unified Science,* which would formalize the foundations of all the sciences—physical, biological, and social. Twenty monographs in the series were published, in two volumes.

The Vienna Circle held regular meetings from 1924 to 1936; at any given time, there were 10–20 people involved. The central figures at the start were the

[11] Slemrod [77] writes, "Historically a canonical interpretation of this '6th problem of Hilbert' has been taken to mean passage from the kinetic Boltzmann equation for a rarefied gas to the continuum Euler equations of compressible gas dynamics as the Knudsen number ϵ approaches zero." I do not know what is the basis for this rather narrow interpretation.

physicist Moritz Schlick, who served as chair, the sociologist Otto Neurath, and the mathematicians Otto Hahn and Philipp Frank. In 1926, the Circle were joined by Rudolf Carnap, who became the leading exponent of logical positivism; his book *The Logical Structure of the World* became a Bible of the movement. (Gödel was also a participant in the meetings of the Circle; however, he does not seem to have ever subscribed to the tenets of logical positivism.)

During the 1950s, there seems to have been an explosion of interest in the subject. Most notably, in 1957, Henkin et al. [42] organized a ten-day international symposium at Berkeley on *The Axiomatic Method: With Special Reference to Geometry and Physics*; Part II consisted of 13 papers on "Foundations of Physics". (Part I had to do with geometry; part III was miscellaneous.) In particular, the papers by Adams on rigid body and particle mechanics, by Noll on continuum mechanics, by Hermes on axiomatizing mechanics, and by Suppes, by Walker, and by Ueno on relativistic kinematics give sets of precise axioms that could easily be formalized in a logical notation, and used in a proof verifier. These all lie within the foundational paradigm; they are concerned with formulating basic axioms, not with drawing connections to experiment or observation, except in a very general sense.

There are a number of striking gaps. Carnap was not involved, despite being a good friend of Tarski's and nearby at UCLA; nor are there any citations to his work or any of the other logical positivist work. Hilbert's sixth problem is never mentioned as a context or motivation. Despite the fact that the organizers were Henkin, Suppes, and Tarski, none of the papers in the physics section use logical notation or refer to the concepts of mathematical logic; of course, it is not an especially congenial notation for physics theories. (Several of the papers in parts I and III do use logical notation and reference mathematical logic.) Feynman's dictum notwithstanding, atoms are never mentioned, as far as I can tell.

In their preface to the proceedings of the symposium, Henkin et al. [42] expressed some reservations about whether the project of axiomatizing physics was a reasonable one:

> Much foundational work in physics is still of the programmatic sort, and it is possible to maintain that the status of axiomatic investigations in physics is not yet past the preliminary stage of philosophical doubt as to its purpose and usefulness.

An even sharper critique arguing for the unsuitability in physical reasoning, not merely of axiomatic logic, but of any kind of rigorous mathematics, was Schwartz [75] "The Pernicious Influence of Mathematics on Science."

A number of important papers along the same lines precede the conference e.g. McKinsey et al. [65]. But after the conference, this line of research seems to have gradually petered out.[12] Montague [68] wrote a paper on deterministic physics,

[12] I am necessarily relying here on the fact that I have failed to find much later work of this flavor, which is obviously an unreliable argument. However, I do have the following concrete evidence. The International Congress of Logic, Methodology, and Philosophy of Science was in some respects the successor to the Symposium on the Axiomatic Method; it has met 15 times since its inception in 1960. Between 1960 and 1999 there was only one paper [66] that presented an

illustrated with an axiomatization of the gravitational theory of a finite collection of particles, written in logical notation. In the last few decades there have been some further sporadic studies of this kind e.g. [74].

In recent years, some philosophers seeking a mathematical framework for science have turned to Bayesianism, discussed earlier in Sect. 4.2.

A fascinating study, not easily characterized in terms of the above categories, is Strevens' [82] *Tychomancy*. Drawing extensively on the cognitive psychology of probabilistic reasoning, Strevens attempts to justify the probabilistic reasoning underlying Maxwell's amazing derivation of the distribution of velocities among particles in a gas; he includes also a discussion of the reasoning involved in Darwinian evolution.

6.3.1 Is PAVEL a Bad Reinvention of Logical Positivism?

In many ways the previous undertaking that most resembles my proposal for PAVEL was logical positivism. Like PAVEL, logical positivism, as applied to physics, attempted to draw a logical line all the way from the theory to the experience of the scientist doing measurements or observations and to characterize the way in which the theory explains the data and the data supports the theory.

That is not the most encouraging of precedents. The general consensus is that logical positivism was thoroughly demolished by Wittgenstein, Popper, Quine, Kuhn, Lakatos, and others, and that it is an entire dead end—a wholly unworkable approach to the analysis of the scientific method. "The fundamental assumptions of the positivist world view . . . lie shattered" [6]. Is PAVEL trying to revive a long-dead horse?

Obviously, I don't think so. I think that there are reasons for optimism.

First, the general consensus may be overstated. A philosophical programme that makes ambitious claims is apt to get strong rejoinders, but demonstrating that it has limitations and flaws does not establish that it has nothing of value to offer. Moreover, part of the disrepute of logical positivism is that it became associated with the psychological theory of behaviorism; but the philosophy of science in no way depends on that. There are some indications that the pendulum in the philosophical world may be swinging back.

Second, one issue that the logical positivists were never able to resolve to their own satisfaction was the nature of the ultimate data. The bedrock data from which theory is built are supposed to be "protocol statements" expressing "direct perception", but that turns out to be a very slippery notion. We are in a better position to deal with that now. Perception is better understood now than in 1930. If we want, we could use computer vision to start with actual sensor input. Whether or not this would have satisfied Carnap or early Wittgenstein as an epistemically

axiomatization of any physical theory (relativistic space-time), though there was a second paper [60] that argued in favor of axiomatizations.

primitive starting point, it is clearly a well-defined and motivated starting point. The analogous question then becomes, at what level do we move from opaque computer vision procedures to representations with semantics, but that is much more of an engineering question.

Finally, PAVEL has the advantage of being AI, not philosophy. It therefore does not have to produce a theory that covers all cases, or to find its way down to the ultimate turtle or to characterize the whole chain of turtles; if it produces a useful partial answer, that is enough to justify the undertaking. We can set the starting point wherever we want, and get a theory that is more or less powerful and rich. For instance, rather than insisting on taking human perception as the grounding point and viewing the validity of experimental measurements as a hypothesis to be tested, we can take the experimental measurements as a given; that will give results that are in some respect more limited but could still be very enlightening. In my discussion of the BACON program, below, I am critical of BACON for using pre-digested data; but there is nothing wrong with taking that as a starting point, as long as one is aware of its limitations. Problem representations like the one in Table 7 are also enormously pre-digested as compared to the actual sensor input, though much richer than the BACON input. The key point is to be aware of the many levels of abstraction that are ultimately involved, and to keep working toward realism.

6.4 Artificial Intelligence

Within AI, there is work of many kinds on physical reasoning [16]; there are AI programs that solve word problems e.g. [32, 50, 51]; that do qualitative reasoning [7]; that design devices e.g. [44], and that design experiments e.g. [52, 67]. Data mining and machine learning are now ubiquitous in scientific research. In this section I will limit the discussion to AI research on developing rich declarative theories of basic physics, and on inferring fundamental theories from data.

6.4.1 Knowledge-Based Physical Reasoning

The AI project closest to PAVEL was the GALILEO project [57]. GALILEO used the Isabelle proof assistant to encode a number of models of physical theories and their experimental consequences, including: Joseph Black's theory of latent heat and heat capacity; the explanations of galactic orbital velocities by positing dark matter and by using Milgrom's proposed modification of Newtonian gravity; Roemer's measurement (in 1676) of the speed of light by delays and advances in the perceived eclipses of Io by Jupiter; the identification of the morning and evening star as the same planet, using observations and Kepler's theory (oddly unhistorical, since the

identification was known to the Babylonians[13]); and Pythagoras' determination that the earth is spherical, based on its shadow on the moon during eclipses.

The primary objective of GALILEO was to characterize how ontologies and theories change as a result of disconfirming evidence, and the examples were used as illustrations of various techniques for changing theories. The details of the representation are therefore only developed necessary to illustrate these meta-level techniques. For instance, in the encoding of the speed-of-light example, the time delay on light coming from Jupiter is taken as a primitive measurement; there is no mention of Io or its revolutions.

A research programme, initiated by Hayes [39, 40] aims toward analyzing physical reasoning, particularly "naive" or "commonsense" physical reasoning, at the knowledge level [69] by formulating theories of physics in a logical form and demonstrating that simple inferences can be justified as inference within the logical theory. This is part of a more general project in AI of using logic to formalize the representation of commonsense knowledge and the process of commonsense reasoning [19, 63, 86]. I myself have continued this direction of research, axiomatizing elementary reasoning about cutting [14], carrying objects in boxes and containers [18, 23], and pouring liquids [17]. The results are axiom sets and problem formulations similar in flavor to Tables 1, 2, 3, 4, 5, 6, and 7. The sample inferences in [23] were automatically verified in SPASS [89], a first-order theorem prover considerably less powerful and expressive than Coq or Isabelle, but easier to use.

Bundy et al. [10] used logic in a quite different way for solving simple physics word problems. Using the "logic programming language" Prolog, they implemented a system that accepted a problem written in English; carried out a "semantic parse" to extract the content of the problem statement; used a rule-based system to find the appropriate equations; and then solved the equations. The program was supplied with schemas for translating categories of problems into equations, for example

```
schema(pullsys
     [Pull,Str,P1,P2], Time
     [ constacc(P1,Time),
       constacc(P2,Time),
       cue stringsys(Str,[Lpart,Rpart]),
       (tension(Lpart,T1,Time)
            <-- coeff(Pull,zero) &
                tension(Rpart,T,Time)   )
     ],
     [ coeff(Pull,zero),
       mass(Pull,zero,Time)
     ]
  )
```

[13]In fact, despite its popularity as an philosophical example since Frege, there is little evidence that anyone who was aware of the existence of the planets has ever thought that Phosphorus and Hesperus were two different planets.

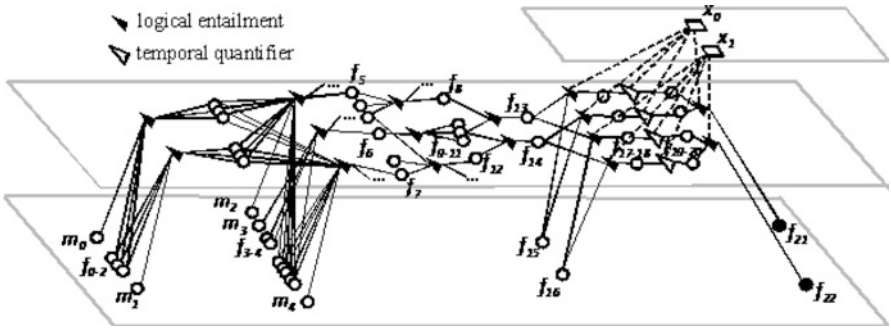


Fig. 3 Network of explanations. From Friedman et al. [29]. This represents the structure of the explanation of the change of temperature over the seasons in terms of the false theory that the earth is closer to the sun in summer

The explanation is thus: "This schema asserts that in a standard pulley problem, the objects undergo constant acceleration, the tension in both parts of the string is equal if there is no friction, and that the friction and mass of the pulley default to zero if not otherwise specified." It is notable here that the general physical law about tension is placed subordinate to the class of pulley problems—that is, at least as done here, it would have to be restated separately in each class of problems where it is used; the general law is placed parallel to the defaults of zero mass and friction on pulleys, which are mostly just conventions about how exercises are written. In a more general knowledge base, it would be better to separate out these levels.

Friedman et al. [29] develop a cognitive model of how student progress from incorrect to correct explanations of physical phenomena. The representation used in that model is a detailed knowledge-based (though not logic-based) structure (Fig. 3) that relates the observation that Chicago is warmer in summer than winter, both to the correct theory of the seasons (the earth's axis is tilted) and to a common misconception (the earth is closer to the sun in summer than in winter).

6.4.2 AI Programs That Induce Scientific Theories

AI programs that have induced broad or fundamental scientific theories from data are few. (There have of course been an enormous number of projects that have used data mining for scientific discovery for very specific projects.)

The largest project of this kind was the BACON project of Langley et al. [54, 55] which modeled the induction of scientific laws from data. BACON, in its various incarnations, took as input data tables of results whose values are either numerical or uninterpreted symbolic values. It had heuristics for formulating numerical laws which can depend on inferred intrinsic properties. For instance, if resistors A, B, and

Table 8 Chemical data input to BACON (from Langley et al. [55])

Element	Compound	w_E	w_C	v_E	v_C	w_E/w_C	w_E/v_E	w_E/v_C
Hydrogen	Water	10.0	90.0	112.08	112.08	0.1111	0.0892	0.0892
Hydrogen	Water	20.0	180.0	224.16	224.16	0.1111	0.0892	0.0892
Hydrogen	Water	30.0	270.0	336.25	336.25	0.1111	0.0892	0.0892
Hydrogen	Ammonia	10.0	56.79	112.08	74.72	0.1761	0.1338	0.1338
Hydrogen	Ammonia	20.0	113.58	224.16	149.44	0.1761	0.1338	0.1338
Hydrogen	Ammonia	30.0	170.37	336.25	224.16	0.1761	0.1338	0.1338
Hydrogen	Ethylene	10.0	140.10	112.08	112.08	0.0714	0.0892	0.0892
Hydrogen	Ethylene	20.0	280.21	224.16	224.16	0.0714	0.0892	0.0892
Hydrogen	Ethylene	30.0	420.31	336.25	336.25	0.0714	0.0892	0.0892

C each give rise to a linear relation between voltage and current, then BACON can formulate the rule $V = IR$, conjecturing that each of the resistors has a different value for R.

BACON's tabulated clean data is, of course, extremely remote from the realities of experiment interpretation that scientists had to deal with. For instance, Table 8 shows the input from which BACON inferred Prout's law of definite proportion in chemical composition. This contrasts starkly with the actual situation of eighteenth and nineteenth century chemists (Fig. 4), who had to identify chemicals and elements and to distinguish them from mixtures using the techniques and methods available in the labs of the time. Langley, Bradshaw, and Simon do point out that Bacon had the advantage of using clean data, while the data available to the historical scientist used included both noise and significant errors; and that Bacon was presented with only the relevant variables, while a large part of the task facing the scientists was figuring out which variables were critical. But, despite a long historical discussion, they don't address the enormous epistemic gap between a table of numbers and a laboratory set up.

I argued above that in examining the relation of theory to data, it was reasonable to take the grounding data at any level of abstraction. So there is nothing inherently invalid with BACON having taken the data in Table 8 as the starting point for theory construction; only, it is important to realize how much that leaves out, as a model of science.

More recently, Bridewell and Langley [9] has been working on inducing process models characterized by differential equations from traces of parameters over time, across a wide range of domains, including aquatic eco-systems, biochemical kinetics, and molecular biology.

6.4.3 Bayesian Inference of Structure

Kemp and Tenenbaum [49] implemented a program that quite directly follows the Bayesian program described in Sect. 4.2 to infer theories from data. Their space of

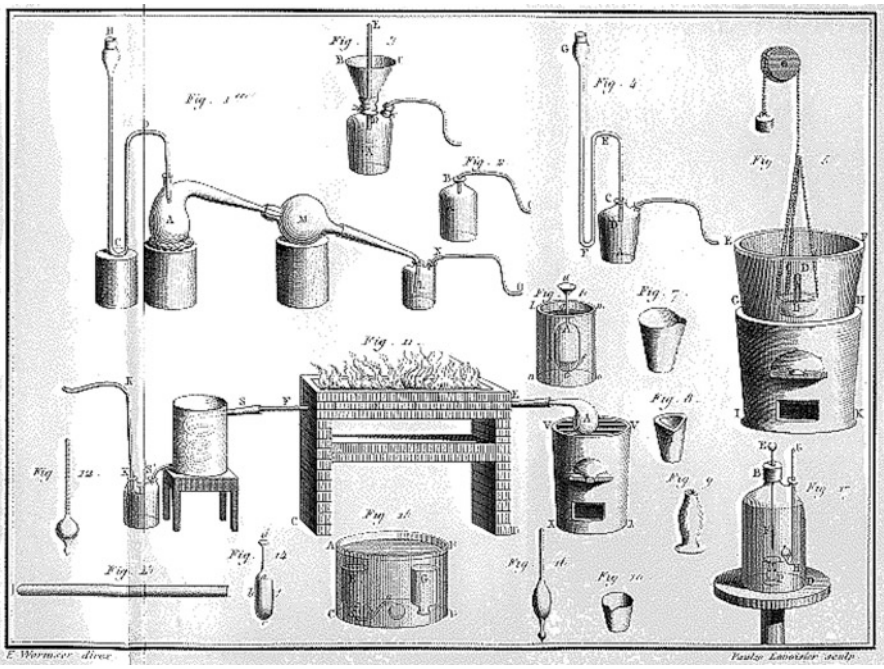


Fig. 4 Lavoisier's equipment. From Lavoisier *Ouevres*, Paris, 1862

theories Φ is the space of graph structures. The prior $P(H)$ for $h \in \Phi$ is given by a generative process that generates graph structures with various kinds of regularities. The likelihood function $P(d|h)$ is a measure of how well the data fits the structure. The program uses heuristic search to approximately find the most probably structure given the data.

The program was applied to a variety of induction problems. As Fig. 5 illustrates, it inferred from a table of animal features that animal species conform to a tree structure; it inferred from a table of features of Supreme Court opinions that Supreme Court justices conform to a linear structure (conservative to liberal); it inferred from a table of similarity judgments over colors that that colors follow a ring structure; it inferred that a collection of images of faces varying along masculinity and race conforms to a two-dimensional grid; and it inferred from a table of distances between world cities that the position of cities corresponds to a graph that is the cross-product of a ring structure for latitude with a linear structure for longitude.

The last of these, however, inadvertently points out the dangers of using an inappropriate space of models in this kind of study. They write as follows:

> We applied the model to a dataset of distances between 35 world cities. Our model chooses a cylinder where the chain component corresponds approximately to latitude and the ring component corresponds approximately to longitude.

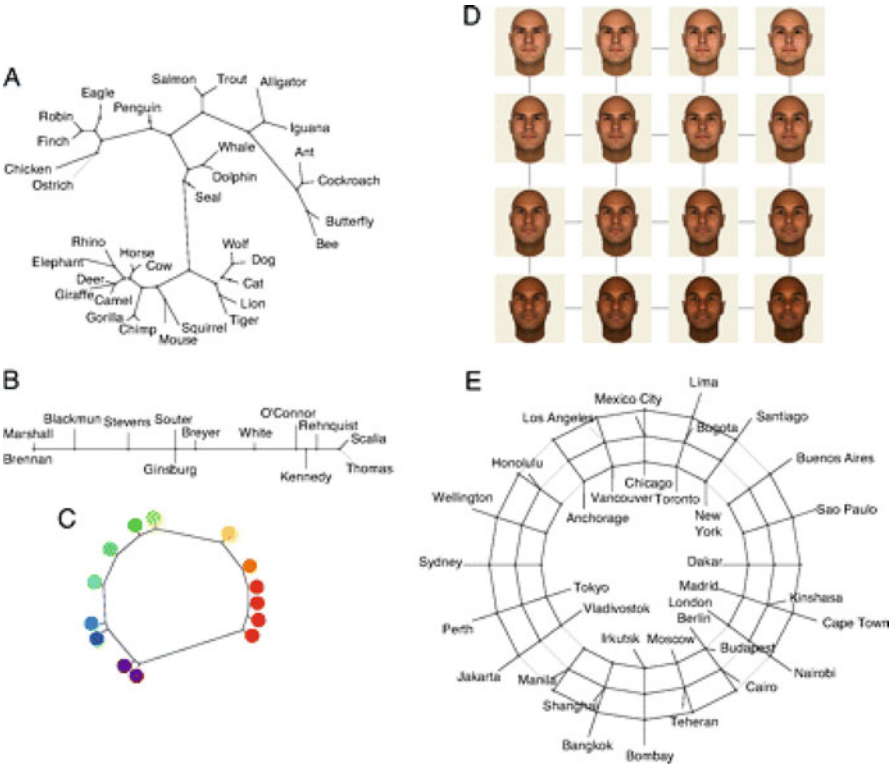


Fig. 5 Results of structure induction. From Kemp and Tenenbaum [49]

This outcome is so far from reality that one wonders why they would think it supports their theory. The correct model for the geodesic distances between cities on the globe, accurate to within the precision of measurement, is that they are points or small regions on the surface of a sphere; this model, however, is not even in the space of discrete models that they are searching over. Optimizing a model of the distance between cities is not, historically, how the shape of the earth was induced, or could have been induced. In general it is mathematically impossible to induce the concepts of latitude and longitude from city distances, because the choice of the particular grid for latitude and longitude has essentially no connection to the position of cities, except insofar as there are no cities close to the poles, and that some major coastlines lie roughly north-south. There is no particular reason that a graph of cities should give one a cylindrical structure, rather than any other planar graph, since any planar graph can be embedded in the sphere. In fact, you will only get a cylindrical graph structure corresponding to latitude and longitude if you pick the cities rather carefully with that outcome in mind. If you actually look for structure in the distances between cities in the world, what will be most conspicuous is their tendency to cluster; cities are dense in some areas and very

sparse in others—completely absent in the oceans that make up 7/10 of the earth's surface. The area in the South Pacific where there are no large cities is considerably larger than the areas around the North or South Pole.

In short, what Kemp and Tenenbaum did in this example is that they cherry-picked data to induce a structure that sound impressive but is actually meaningless in terms of the semantics of the data, using an inductive bias that bears no relation to the semantics of the data, searching through a space of models that does not contain models of the correct type.

6.4.4 Domingos and the Master Algorithm

The techniques of corpus-based machine learning that have recently been particularly successful, such as deep learning, are mostly highly specific in their focus and do not attempt to induce symbolic theories. Thus they are not directly relevant to PAVEL. However, Domingos [26], in his book *The Master Algorithm*, a survey of machine learning techniques, speculates as follows:

> The Master Algorithm is the germ of every theory: all we need to add to it to obtain theory x is the minimum amount of data required to induce it. (In the case of physics, that would be the results of perhaps a few hundred key experiments).

Domingos' "Master Algorithm" is a universal machine learning algorithm, which can optimally induce theories from data. He takes this as the Holy Grail of machine learning, and considers that it may well be found in the not very distant future. So his claim is that, in principle, one could choose a few hundred experiments that, given as input to the Master Algorithm, would enable the algorithm to induce all of physics.

I presume that Domingos is thinking here of something akin to the formulation in BACON; the input is a digested table of numbers, the target output is the foundational theories. Even so, "a few hundred" seems to me a huge underestimate. If the intended input is something close to a realistic description of the experiment, then the estimate of the number of experiments is surely off by at least a couple of orders of magnitude. (Not that it is always easy to individuate or count number of experiments; how are astronomical observations counted, for example?) Finally, cherry-picking only the evidence supporting the eventual theories is an unrealistic and ecologically invalid undertaking; a true logical reconstruction of science would have to take into account all the evidence that doesn't fit well, or is irrelevant. Still, in general what Domingos is suggesting here is somewhat comparable to PAVEL.

7 Potential Philosophical Impact

It seems to me that implementing some part of PAVEL might well yield insights that would be of interest to philosophers of science, on issues such as the nature of informal argumentation in physics, the sufficiency of physics as an explanation, the

nature of the reduction of the other sciences to physics, and the universalizing claims made by physics. Even if PAVEL takes an approach to issues that the philosophers found unacceptable from a philosophical standpoint, still the existence of one clear-cut approach to these issues would be valuable, if only as a point of comparison.

Another point where the construction of PAVEL might shed light is on the nature of the prior expectations. There seems to have been an enormous pressure over the centuries of the development of physics to find theories that are governed by a small, simple dynamic theory, at an almost arbitrary cost in the complexity of the boundary conditions; that are local in time and space; that are universal; that obey various kinds of symmetry; that conform to mathematically elegant equations[14]; and that are mechanistic. It would seem, moreover, that the preferences for these kinds of theories are stronger than be accounted for in terms of minimum description length or other such general principles. One evidence for this is that, historically, scientists were eager to claim the universality of physics long before, in retrospect, it would seem that the state of the data or theory came close to justifying it. For example, in 1814, Laplace contemplated

> An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies in the universe and those of the tiniest atom; for such an intellect, nothing would be uncertain and the future just like the past would be present before its eyes.

Laplace had every justification to say this of the solar system, having worked it out himself. But what evidence did he have that this applied to all the other motions in the universe, considering what a small fraction of motions the science of his time could actually explain or predict?

Assuming that this is right, are these preferences necessary, as prior preferences? Are they in any well-defined sense rational? Perhaps they are merely expressions of the existing power structure, in a Foucaultian sense.

At this point, I have to confess, I find myself seduced by the siren song of Bayesianism. It would be so wonderful to be able to assign a numerical confidence to the theory of gravity, or to Schrödinger's equation, or to the universalizing claims discussed in Sect. 4.7! Or to determine to what extent any particular experimental finding should increase or decrease our confidence in any particular theory. It seems like it should be so close, comparatively speaking! The equation is sitting there, in Sect. 4.2; all we have to do is to find well-founded values for the numbers.

[14] Hossenfelder [45] argues that the fetishizing of mathematical elegance is responsible for the stagnation of fundamental physics over the last few decades.

8 Conclusions: Whither PAVEL?

There is no lack of things to do: there are easy things to do in the short term and harder things to do in the long term. The most important directions, it seems to me, would be:

- To increase the collection of physical theories we have in forms that can be used in a theorem prover.
- To develop techniques for choosing suitable idealizations, approximations, and abstractions for a given situation.
- To analyze the nature of the informal argumentations used in physics.
- To validate the approach by showing how word problems and experiments can be verified in these theories.
- To further validate the representation of word problems by developing natural language system that can translate verbal statements into formal representations.

Acknowledgements Thanks for useful information and helpful feedback to Scott Aaronson, Alan Bundy, Ken Forbus, Tom LaGatta, Michael Strevens, David Tena Cucala, Peter Winkler, and the anonymous reviewer.

References

1. Appel K, Haken W (1977) The solution of the four-color-map problem. Sci Am 237(4):108–121
2. Avigad J, Donnelly K, Gray D, Raff P (2007) A formally verified proof of the prime number theorem. ACM Trans Comput Log 9(1):1–23
3. Bailey DH, Borwein J (2015) Experimental computation as an ontological game changer: the impact of modern mathematical computation tools on the ontology of mathematics. In: Davis E, Davis P (eds) Mathematics, substance and surmise: views on the meaning and ontology of mathematics. Springer, Cham
4. Beech M (2014) the pendulum paradigm: Variations on a theme and the measure of heaven and earth. Brown Walker Press, Boca Raton
5. Berger M (2010) Geometry revealed: a Jacob's ladder to modern higher geometry. Springer, Berlin
6. Bhaskar R (1979) Realism in the natural sciences. In: Proceedings of the sixth international congress of logic, methodology and philosophy of science
7. Bobrow D (ed) (1985) Qualitative reasoning about physical systems. MIT Press, Cambridge
8. Boender J, Kammüller F, Nagarajan R (2015) Formalization of quantum protocols using Coq. In: Huenen C, Selinger P, Vicary J (eds) 12th international workshop on quantum physics and logic, pp 71–83.
9. Bridewell W, Langley P (2010) Two kinds of knowledge in scientific discovery. Top Cogn Sci 2:36–52
10. Bundy A, Byrd L, Luger G, Mellish C, Palmer M (1979) Solving mechanics problems using meta-level inference. In: IJCAI-79, pp 1017–1027
11. Cappellaro P (2012) Quantum theory of radiation interaction. MIT OpenCourseware. https://ocw.mit.edu/courses/nuclear-engineering/22-51-quantum-theory-of-radiation-interactions-fall-2012/lecture-notes/

12. Corry L (2004) David Hilbert and the axiomatization of physics (1898–1918): from Grundlagen der Geometrie to Grundlagen der Physik. Kluwer, Dordrecht
13. Davis E (1988) A logical framework for commonsense predictions of solid object behavior. AI Eng 3(3):125–140
14. Davis E (1993) The kinematics of cutting solid objects. Ann Math Artif Intell 9(3–4):253–305
15. Davis PJ (1993) Visual theorems. Educ Stud Math 24(4):333–344
16. Davis E (2008) Physical reasoning. In: van Harmelen F, Lifschitz V, Porter B (eds) The handbook of knowledge engineering. Elsevier, Amsterdam, pp 597–620
17. Davis E (2008) Pouring liquids: a study in commonsense physical reasoning. Artif Intell 172:1540–1578
18. Davis E (2011) How does a box work? a study in the qualitative dynamics of solid objects. Artif Intell 175:299–345
19. Davis E (2017) Logical formalizations of commonsense reasoning: a survey. J Artif Intell Res 59:651–723
20. Davis E (in preparation) The logic of coal, iron, air, and water: representing common sense and elementary science
21. Davis E, Marcus G (2014) The scope and limits of simulation in cognition. https://arxiv.org/abs/1506.04956
22. Davis E, Marcus G (2016) The scope and limits of simulation in automated reasoning. Artif Intell 233:60–72
23. Davis E, Marcus G, Frazier-Logue N (2017) Commonsense reasoning about containers using radically incomplete information. Artif Intell 248:46–84
24. Dehaene S (1997) The number sense: how the mind creates mathematics. Oxford University Press, Oxford
25. de Kleer J, Brown JS (1985) A qualitative physics based on confluences. In: Bobrow D (ed) Qualitative reasoning about physical systems. MIT Press, Cambridge
26. Domingos P (2015) The master algorithm: how the quest for the ultimate learning machine will remake our world. Basic Books, New York
27. Dwyer J, Uman MA (2013) The physics of lightning. Phys Rep 534:147–241. https://doi.org/10.1016/j.physrep.2013.09.004
28. Feynman R, Leighton RB, Sands M (1964) The Feynman lectures on physics. Addison-Wesley, Boston
29. Friedman S, Forbus K, Sherin B (2017) Representing, running, and revising mental models: a computational model. Cogn Sci 42(2):1–36
30. Gonthier G et al (2013) A machine-checked proof of the odd order theorem. In: International conference on interactive theorem proving (Lecture notes in computer science), vol 7998. Springer, Berlin, pp 163–179.
31. Gopnik A (2012) Scientific thinking in young children: theoretical advances, empirical research, and policy implications. Science 337(6102):1623–1627
32. Gunning D et al (2010) Project Halo update—progress toward Digital Aristotle. AI Mag 31(3):33–58
33. Gutierrez TD (1999) Standard model lagrangian. http://nuclear.ucdavis.edu/~tgutierr/files/stmL1.html
34. Hales T et al (2015) A formal proof of the Kepler conjecture. http://arxiv.org/abs/1501.02155
35. Hamel GKW (1912) Elementare Mechanik. Teubner, Leipzig
36. Hamel GKW (1921) Grundbegriffe der Mechanik. Teubner, Leipzig
37. Harrison J (2006) Formal verification of floating point trigonometric functions. In: International conference on formal methods in computer-aided design, pp 254–270
38. Harrison J (2009) Formalizing an analytic proof of the prime number theorem. J Autom Reason 43(3):243–261
39. Hayes P (1979) The naïve physics manifesto. In: Michie D (ed) Expert systems in the microelectronic age. Edinburgh University Press, Edinburgh
40. Hayes P (1985) Ontology for liquids. In: Hobbs J, Moore R (eds) Formal theories of the commonsense world. Ablex Publishing, New York

41. Hendry RF (1999) Chemistry and the completeness of physics. In: Symons J, van Dalen D, Davidson D (eds) Philosophical dimensions of logic and science: selected contributed papers from the 11th international congress of logic, methodology, and philosophy of science, Kraków, pp 165–178
42. Henkin L, Suppes P, Tarski A (1958) The axiomatic method with special reference to geometry and physics. North-Holland, Amsterdam
43. Hertz H (1894) Die Prinzipien der Mechanik. Johann Ambrosius Barth, Leipzig
44. Hornby GS, Globus A, Linden DS, Lohn JD (2006) Automated antenna design with evolutionary algorithms. In: AIAA space, pp 19–21
45. Hossenfelder S (2018) Lost in math: how beauty leads physics astray. Basic Books, New York
46. Howson C, Urbach P (2006) Scientific reasoning: the Bayesian approach. Open Court Publishing, Chicago
47. Jaynes ET (2003) Probability theory: the logic of science. Cambridge University Press, Cambridge
48. Jeannin JB et al (2015) A formally verified hybrid system for safe advisories in the next-generation airborne collision avoidance system. Int J Softw Tools Technol Transfer 19(6):717–741
49. Kemp C, Tenenbaum J (2009) The discovery of structural form. Proc Natl Acad Sci 105(31):10687–10692
50. Khashabi D, Khot T, Sabharwal A, Roth D (2018) Question answering as global reasoning over semantic abstractions. In: AAAI-18
51. Khot T, Sabharwal A, Clark P (2018) SciTail: a textual entailment dataset from science question answering. In: AAAI-18
52. Krenn M et al (2016) Automated search for new quantum experiments. Phys Rev Lett 116(9):090405
53. Kushman N, Artzi Y, Zettlemoyer L, Barzilay R (2014) Learning to automatically solve algebra word problems. In: ACL-2014
54. Langley P, Bradshaw GL, Simon HA (1981) BACON.5: The discovery of conservation laws. IJCAI-81, pp 121–126
55. Langley P, Bradshaw GL, Simon HA (1983) Rediscovering chemistry with the BACON system. In: Michalski RA et al (eds) Machine learning. Springer, Berlin
56. Laughlin R, Pines D (2000) The theory of everything. Proc Natl Acad Sci 97(1):28–31
57. Lehmann J, Chan M, Bundy A (2013) A higher-order approach to ontology evolution in physics. J Data Semant 2:163–187
58. Ludwig G (1985) An axiomatic basis for quantum mechanics: volume 1, derivation of Hilbert space structure. Springer, Berlin
59. Ludwig G (1987) An axiomatic basis for quantum mechanics: volume 2, quantum mechanics and Macrosystems. Springer, Berlin
60. Ludwig G (1989) An axiomatic basis as a desired form of a physical theory. In: Proceedings of the 1987 international congress for logic, methodology, and the philosophy of science. North Holland, Amsterdam
61. Martin U, Pease A (2015) Hardy, Littlewood, and polymath. In: Davis E, Davis P (eds) Mathematics, substance and surmise: views on the meaning and ontology of mathematics. Springer, Berlin
62. MathOverflow (2016) Is it possible to have a research career while checking the proof of every theorem you cite? https://mathoverflow.net/questions/237987/is-it-possible-to-have-a-research-career-while-checking-the-proof-of-every-theor
63. McCarthy J (1968) Programs with common sense. In: Minsky M (ed) Semantic information processing. MIT Press, Cambridge
64. McCune W (1997) Solution of the Robbins problem. J Autom Reason 193, 264–276
65. McKinsey JCC, Sugar AC, Suppes P (1953) Axiomatic foundations of classical particle mechanics. J Ration Mech Anal 2:253–272
66. Mehlberg H (1965) Space, time, relativity. In: Proceedings of the 1964 international congress for logic, methodology, and the philosophy of science. North Holland, Amsterdam

67. Melnikov A et al (2018) Active learning machine learns to create new quantum experiments. PNAS 115(6):1221–1226
68. Montague R (1974) Deterministic theories. In: Thomason R (ed) Formal philosophy: selected papers of Richard Montague. Yale University Press, New Haven
69. Newell A (1982) The knowledge level. Artif Intell 18(1):87–127
70. Nipkow T, Paulson LC, Wenzel M (2002) Isabelle/HOL: a proof assistant for higher-order logic. Lecture notes in computer science, vol 2283. Springer, Berlin
71. Paleo BW (2012) Physics and proof theory. Appl Math Comput 219:45–53
72. Rosenkrantz R (1977) Inference, method, and decision: towards a bayesian philosophy of science. Springer, Berlin
73. Russell B (1903) The principles of mathematics. Cambridge University Press, Cambridge
74. Sant'Anna A (1999) An axiomatic framework for classical particle mechanics without space-time. Philos Nat 36:307–319
75. Schwartz J (1960) The pernicious influence of mathematics on science. In: Proceedings of the 1960 international congress for logic, methodology, and the philosophy of science. North Holland, Amsterdam
76. Sigmund K (2017) Exact thinking in demented times: the Vienna circle and the epic quest for the foundations of science. Basic Books, New York
77. Slemrod M (2013) From Boltzmann to Euler: Hilbert's 6th problem revisited. Comput Math Appl 65(10):1497–1501
78. Smith KA, Battaglia P, Vul E (2013) Consistent physics underlying ballistic motion prediction. In: Cognitive science
79. Sober E (2002): Bayesianism—its scope and limits. Proc British Acad 113:21–38
80. Souyris J, Wiels V, Delmas D, Delseny H (2009) Formal verification of avionics software products. In: International symposium on formal methods, pp 532–546
81. Strevens M (2005) The Bayesian approach in the philosophy of science. In: Borchet DM (ed) Encyclopedia of philosophy, 2nd ed. Macmillan, Detroit
82. Strevens M (2013) Tychomancy: inferring probability from causal structure. Harvard University Press, Cambridge
83. Suppes P, Luce RD, Krantz D, Tversky A (1974) Foundations of measurement. Dover Publications, Mineola
84. Tegmark M (2009) The multiverse hierarchy. ArXiv preprint arXiv:0905.1283
85. Tenenbaum JB, Kemp C, Griffiths TL, Goodman ND (2011). How to grow a mind: Statistics, structure, and abstraction. Science 331(6022):1279–1285
86. van Harmelen F, Lifschitz V, Porter B (eds) (2008) Handbook of knowledge representation. Elsevier, Amsterdam
87. Vogt A (1997) Review of *an axiomatic basis for quantum mechanics: vol 1, derivation of hilbert space structure* by Günther Ludwig, SIAM Rev 29(3):499–501
88. Wallace D (to appear) The logic of the past hypothesis. In: Loewer B, Weslake B, Winsberg E (eds) Time's arrows and the probability structure of the world Harvard University Press, Cambridge
89. Weidenbach C, Dimova D, Fietzke A, Kumar R, Suda M, Wischnewski C (2009) Spass Version 3.5. In: International conference on automated deduction (CADE) (Lecture Notes in Computer Science), vol 5563. pp 140–145
90. Whitehead AN, Russell B (1910) Principia Mathematica. Cambridge University Press, Cambridge
91. Wittgenstein L (1922) Tractatus Logico-Philosophicus. Harcourt, Brace, and Company, San Diego
92. Yandell BH (2002) The honors class: Hilbert's problems and their solvers. A.K. Peters. Taylor & Francis, Milton Park

An Applied/Computational Mathematician's View of Uncertainty Quantification for Complex Systems

Max Gunzburger

Abstract Uncertainty quantification (UQ) is defined differently by different disciplines. Here, we first review an applied and computational mathematician's definition of UQ for complex systems, especially in the context of partial differential equations (PDEs) with random inputs. We then discuss the types of stochastic noises that are used as inputs to the PDEs and, for the case of infinite stochastic processes, how those inputs are approximated so that they are amenable to computations. We then review methods that are used to obtain approximations of solutions of PDEs with random inputs, with special emphases given to stochastic Galerkin and stochastic sampling methods, including sparse-grid methods in the latter case. We close with a brief foray into where UQ in the PDE setting is going moving forwards.

1 Introduction

We begin with some general comments that serve to introduce, set up, and focus the material presented in subsequent sections. We do not provide copious citations of the literature; instead, we list [1–5] a few general references in that provide additional details and more comprehensive mathematical and algorithmic expositions of what is written about in this paper.

Uncertainty Is Everywhere Physical, biological, social, economic, financial, etc. systems always involve uncertainties. Certainly, mathematical models of these systems should account for uncertainty. Uncertainties are often classified into two classes. *Epistemic* uncertainty is due to *incomplete knowledge* of the system so that, at least in principle, uncertainty can be reduced by additional measurements, improvements in measuring devices, etc. Although such uncertainties are, again in principle, predictable, it may be too difficult, perhaps impossible, or too costly

M. Gunzburger (✉)
Department of Scientific Computing, Florida State University, Tallahassee, FL, USA
e-mail: mgunzburger@fsu.edu

© Springer Science+Business Media, LLC, part of Springer Nature 2019

N. Fillion et al. (eds.), *Algorithms and Complexity in Mathematics, Epistemology, and Science*, Fields Institute Communications 82,
https://doi.org/10.1007/978-1-4939-9051-1_5

to obtain additional measurements. Subsurface media properties in oil reservoirs or aquifers are an example of this type of uncertainty. *Aleatoric* uncertainty is *intrinsic* in the system so that uncertainty cannot be reduced through additional measurements, improvements in measuring devices, etc. Running an experiment twice with exactly the same settings still results in different outcomes. The distinction between the two is certainly fuzzy; one person's aleatoric uncertainty may be another's epistemic uncertainty.

If computations are involved in uncertainty quantification (UQ), there is a third type of uncertainty which we refer to as *computational epistemic* uncertainty. In this case, not everything in the data and/or solutions can or should be resolved because it is too difficult, perhaps impossible, or too costly to do so in a computational simulation; turbulent flows are an example of this situation. Alternately, some scales may not be of interest, e.g., surface roughness, hourly stock prices; in such cases, uncertainty, e.g. randomness, is sometimes *artificially introduced* into the system to model the effects that behaviors at the unresolved scales have on that can be resolved.

Everyone realizes that laboratory experiments are not precisely repeatable which they often (and always should) be reported with error bars. But, are computer experiments repeatable? Running the same code with exactly the same inputs and on exactly the same computer should result in the same outputs. This statement can remain true even if there is randomness in the inputs because on a computer, one uses quasi-random number generators which one can reproduce from one computational run to the next. However, running the same code with exactly the same inputs on different computers or using different software can result in different outputs, not just because the two computers may use different pseudo-random number generators, but also for other hardware and software differences. With regards to the believability of experimentally determined data vs. data determined computationally, one should recall the quote[1]

> Experimental results are believed by everyone,
> except by the person who ran the experiment.
> Computational results are believed by no one,
> except by the person who wrote the code.

Who Does Uncertainty Quantification and Why Does Everyone Now Want to Do It? Certainly statisticians do; some statisticians even aver that statistics = uncertainty quantification. However, below, the way we define UQ is perhaps not the most common way UQ is thought of by statisticians. Perhaps certain philosophers and probabilists feel the same was as do statisticians. Scientists and engineers of all types do UQ as do some mathematicians, especially those involved in algorithmic and mathematical model development and analysis. So, indeed, everyone does UQ, but not all do it well or honestly.

[1]Some attribute this quote as a slightly modification of a quote attributed to Albert Einstein: A theory is something nobody believes, except the person who made it. An experiment is something everybody believes, except the person who made it.

These days "everyone" wants to do UQ. For many it is because there is money in it but, being less mercenary, some may genuinely believe it is interesting and important.

What Is a Complex System? *Complex system* is a terminology introduced to make it look like one is solving a difficult problem. Thus, it is very useful to write, when you are preparing a proposal for funding, that your work deals with a complex system; that will certainly do you more good than if you write that you work on simple systems. Here is a working definition of what constitutes a complex system.

Let us denote the system input by $\boldsymbol{y}$ and the system output by u. If we change the system input we change the system output so that we can think of the output u as being a function of the system input $\boldsymbol{y}$. *A complex system is one for which determining the dependence of the output on the input requires copious computational resources.* For example, the evaluation of a known function, i.e., $u = f(\boldsymbol{y})$ with $f(\boldsymbol{y})$ a known function, or the integral over a domain D of a known function $f(\boldsymbol{y})$, i.e., $u = \int_D f(\boldsymbol{y})\, d\boldsymbol{y}$ with $f(\boldsymbol{y})$ a known function, are examples of what is not a complex system.

Differential equations, especially partial differential equations, and especially nonlinear partial differential equations provide many examples of what qualifies as a complex system according to our definition, For example, the Navier-Stokes equations for fluid flow

$$\begin{cases} \rho\Big(\dfrac{\partial \mathbf{u}}{\partial t} + \mathbf{u}\cdot\nabla\mathbf{u}\Big) - \nu\Delta\mathbf{u} + \nabla p = \mathbf{0} & \text{in a domain } D \\ \nabla\cdot\mathbf{u} = 0 & \text{in a domain } D \\ \mathbf{u} = \mathbf{f}(\boldsymbol{y}) & \text{on the boundary of } D \end{cases}$$

is a complex system for which the outputs (the velocity $\mathbf{u}$ and the pressure p) must be solved for.

Thus, throughout this paper, it is assumed that *function evaluations*, i.e., determining outputs from the inputs, is an expensive proposition.

A Word About Simulations *Simulation*, a word we have already used without giving its definition, is a noun derived from the verb simulate. In turn, the Oxford English Dictionary definitions of simulate include:

1. pretend to be, have, or feel
2. imitate or counterfeit.

In other words, to simulate is to cheat. However, as a fourth definition, that dictionary has

4. produce a computer model of (a process)

Thus, for us, a simulation is an approximation of the solution, i.e., the output, of a mathematical model that is obtained using computers, especially in situations in which the mathematical model itself is not exactly solvable.

How Do Uncertainties Enter into the Mathematical Descriptions of Systems? Before answering this question, it is useful to keep in mind what the great statistician George Box said about models:

> All models are wrong but some models are useful.

Less well known but equally powerful is another quote of George Box:

> Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. On the contrary, following William of Occam he should seek an economical description of natural phenomena. Just as the ability to devise simple but evocative models is the signature of the great scientist, so overelaboration and overparameterization is often the mark of mediocrity.

Now, back to the question of how uncertainties enter into mathematical models.

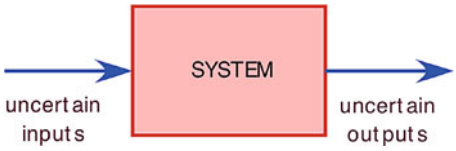
First, *the structure of model itself may not be known precisely.* A trivial example would be if all we knew about a function is that it is continuous and odd, so we could model it as $u = \sin y$, or $u = \tan y$, or $u = y^{2n+1}$ for some positive integer n, or as any of the other countless possibilities. Clearly, we need more information about the function to narrow down the choices. In reality, things are more subtle than that, especially in the context of phenomenological models, i.e., models that cannot be derived with mathematical precision from more basic or more generally accepted models. We also do not often know if simplified models that are cheaper to compute with provide good enough answers.

Even if the model form is agreed upon, it may contain *parameters whose values are not precisely known*, e.g., we know we have a function of the form $u = x^{\alpha}$ but α is a number whose value is not precisely known or we know our model form is $-\alpha \frac{d^2u}{dx^2} = x^2$ but again α is not precisely known.

Beyond input parameters, the model may contain *inputs functions that are uncertain* at every point in their domains, e.g., the coefficient and right-hand side of the differential equation $-\frac{d}{dx}\left(\alpha(x)\frac{du}{dx}\right) = f(x)$.

2 Uncertainty Quantification (UQ)

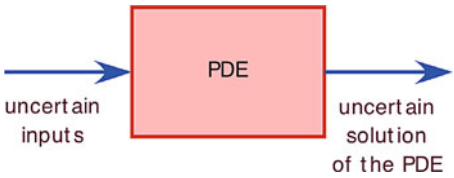
In this section, we provide definitions and discussions about UQ in its many guises as viewed by many applied and computational mathematicians. For that community, a general definition that is illustrated in the figure below is that



UQ is the task of determining information about the uncertainty in the outputs of a system, given information about the uncertainty in its inputs

Of course, a system may have additional inputs that are known with certainty. Let's narrow down this definition a little. We consider systems governed by *partial differential equations* (PDEs) so that now

UQ is the task of determining information about the uncertainty in the solution of a PDE, given information about the uncertainty in its inputs



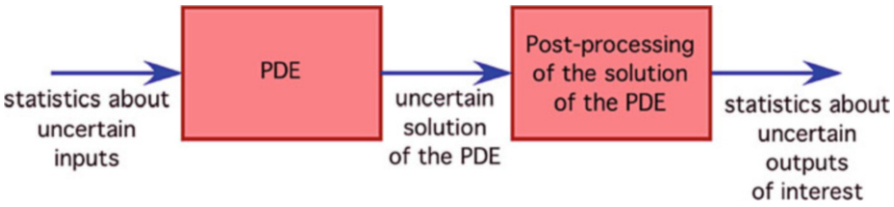
The solution of the PDE defines the mapping from the input variables to the output variables; as already mentioned, very often, PDEs do indeed model complex systems. Determining approximate solutions of a PDE usually requires costly computations. In fact, solving PDES was the reason modern computers were invented in the 1940s and even today, remain a major driving force behind the development new supercomputers.

Often, the solution of the PDE is not the primary output of interest. Of more interest are quantities obtained by post-processing solutions of the PDE to determine outputs of interest. Of course, one still has to obtain a solution of the PDE to determine the output of interest. Thus, now

UQ is the task of determining information about the uncertainty in an output of interest that depends on the solution of a PDE, given information about the uncertainty in its inputs

The desired information about the output is referred to as a *quantity of interest* (QoI).

There are several approaches towards UQ, including but not limited to, fuzzy sets and possibility theory, interval arithmetic, probabilistic approaches, evidence theory (e.g., Dempster-Shafer theory), etc. We consider *probabilistic approaches*, i.e., the uncertainty in the inputs of the PDE are described in terms of statistical quantities, i.e., probability density functions (PDFs), expected values, variances, covariance functions, higher moments, etc. Thus, now



UQ is the task of determining statistical information about the uncertainty in an output of interest that depends on the solution of a PDE, given statistical information about the uncertainty in its inputs

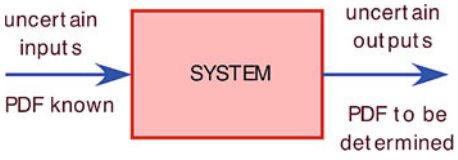
For example, an input parameter α could take the form of a given value plus noise, e.g. $\alpha = \alpha_0 + \eta$, where α_0 is a deterministic number and η is a random number whose value is selected by sampling a given PDF.

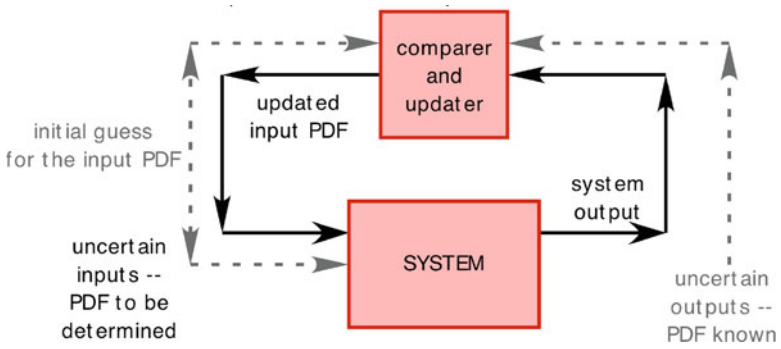
It seems we are in business: someone from on high gives us a mathematical model and statistical information about the inputs to the model. Then, to obtain statistical information about an output of interest depending on the solution of the model, all we have to do is devise a means to (perhaps approximately) solve the model equations. There is small problem however: often that someone on high disappoints us, i.e., often we do now know the needed statistical information about the inputs. What can one do? The most common approach is to make an educated (but sometimes an out-of-the-blue) guess as to what is that information. One can try to do something better such as use field or laboratory observations to make a more informed guess, but one should keep in mind that such observations also come with uncertainty (a troublesome fact that is often ignored),

Model Calibration/Parameter Identification Our discussion so far has been about the forward (or direct) problem of determining information about model outputs, given information about model inputs. Model calibration is about the reverse path, i.e., it is the task of determining statistical information about the inputs of a system, given statistical information about the outputs. On could, e.g., use experimental or field observations to determine the statistical information about the outputs. In particular, one would like to identify the PDF of the input variables. In case the inputs take the form of parameters appearing in the model that needs calibration, model calibration is often referred to as *parameter identification.*

Of course, the system still maps the inputs to the outputs so that determining the input PDF is an *inverse problem* whose solution usually requires multiple forward simulations of the system equations.

UQ: the direct (or forward) problem





Model calibration/parameter identification—inverse problem

Model calibration problems are a particular case of more general stochastic inverse, stochastic control, stochastic optimization, or stochastic design problems.

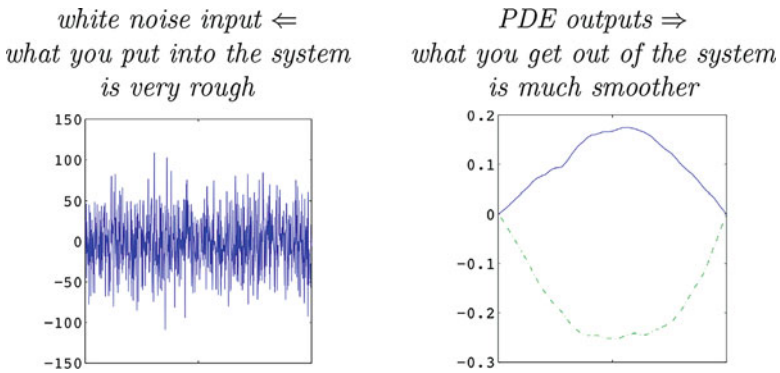
3 Types of Input Noises

We differentiate between the three types of noises that can be used as uncertain inputs that appear in mathematical models such as PDEs.

Random Parameters The input data could depend on a *finite number of random parameters*. Examples are flow rates in an HVAC system, voltages in an electric circuit, load on a beam, and the like. One can think of the parameters as begin "knobs" in an experiment which one has to set before running the experiment, but which in practice cannot be set at the exact value one wants. Each parameter may vary independently according to its own given PDF. Alternately, the parameters may vary according to a given joint PDF or through conditional probabilities.

White Noise Random Fields The value of the input data varies randomly and *independently* from one point of the physical domain to another and/or from one time instant to another. Thus, a white noise random field can be viewed as a function $\eta(\mathbf{x}, t)$ whose value at a point $\mathbf{x}$ and/or at a time t is sampled independently according to a single given PDF from any other point and any other time. Because the values at different times and at different times are independent of each other and are determined by drawing from the same PDF, such fields are referred as being independently and identically distributed, or i.i.d., for short.

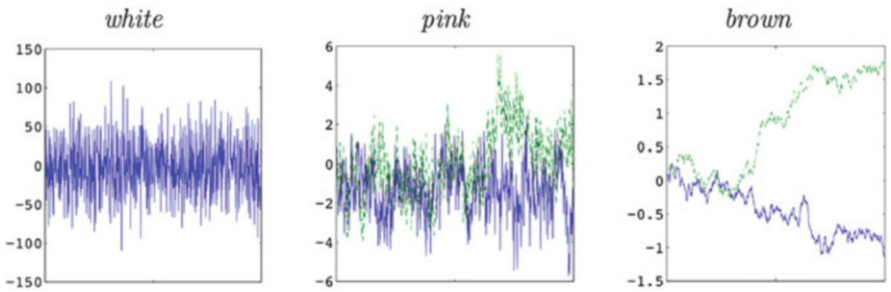
In practice, white noise is by far the most used means for introducing randomness into a system. However, white noise has infinite energy so that it cannot naturally exist. It continuous to be in ubiquitous use because discretizations and truncations of white noise have finite energy and are very easy to implement on a computer. Furthermore, in many settings, the solution of the system equations driven by white noise has finite energy and is much smoother than the input white noise, as is illustrated in the figure below, so that the potential problems with white noise inputs are glossed over.



Colored or Correlated Random Fields The value of the input data varies randomly from one point of the physical domain to another and/or from one time instant to another and is identically distributed but is *not independent* from the values at other points and other time instants. Instead, the values obey a given (spatial/temporal) *correlation structure.*

Colored noises are ubiquitously present. Three well-known colored noises are Brownian or brown noise that are continuous random processes and are related to diffusion; Lèvy noises that are jump processes and are related to anomalous diffusion; and Ornstein-Uhlenbeck or mean-reverting processes.

The figure below contains plots two realizations from each of three one-dimensional random fields, one is white and the other two are colored,[2] with increasing correlation going from left to right. We see that increasing correlation results in increasing smoothness of the field.



Approximate realizations of one-dimensional random fields

[2]Pink noise is "half-way" between white and brown noise; it is referred to as pink because in some circles brown noise is referred to as red noise.

4 Discretization of Stochastic Processes

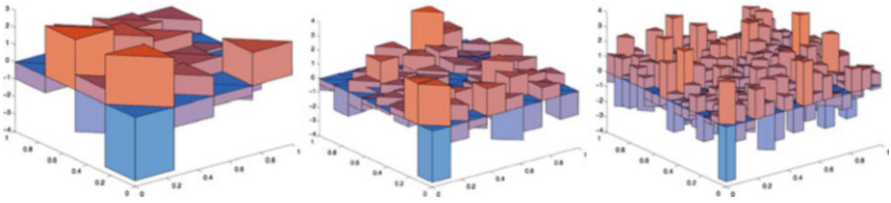
We started with the random parameter case in which the problem inputs involve a *finite* number of random numbers. If we choose values for these parameters, e.g., by sampling them according to their PDFs, we can then solve for the corresponding solution of the system equations, i.e., the PDE.

White noise fields are defined by a given PDF (usually a Gaussian PDF) and are easy to evaluate at a given point and/or at a given instant in time because the values may be chosen independently from the values previously sampled at other points and time instants; one merely samples from a given PDF.

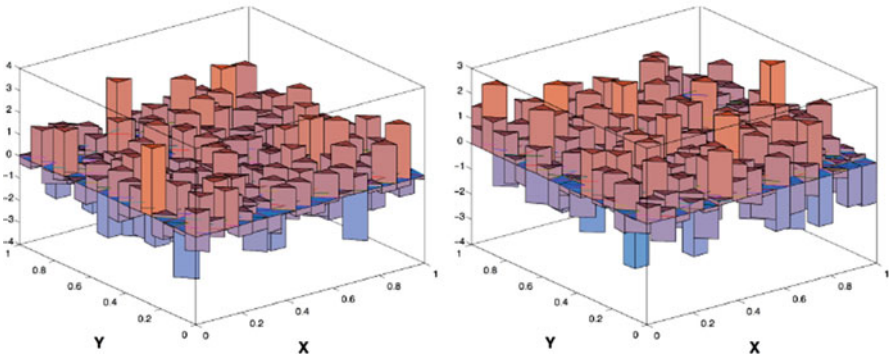
Colored noise (correlated random fields) are usually defined in terms of a PDF and correlation function. One is usually not given a "formula" that allows one to evaluate the random field at a given point and/or at a given instant in time. However, spatially and/or time-dependent random fields, whether the are correlated or white, can also be described in terms of parameters. But, because they are *infinite* stochastic processes, it requires an infinite number of random parameters to describe them.

Of course, on a computer one can only solve problems involving a finite number of random parameters. In the white and colored noise cases, one discretizes the noise so that the noise is approximated in terms of a finite number of parameters.

Discretizing White Noise The most common means for discretizing white noise is to draw independent samples from its PDF and use those samples to assign values of the field at each grid point (or in each grid cell) and/or each time interval used in discretizing the PDE. In this case, if J denotes the number of grid points (or grid cells) and K denotes the number of time intervals used to solve the PDEs, then the number of independent samples drawn is J for steady-state problems and JK for time-dependent problems. Thus, if one refines the spatial or temporal grids, the number of random samples increases. For example, in three dimensions, it one halves the spatial grid size, one would increase the number of parameters by a factor of 8. The following figures illustrate realizations of discretized white noise. Note that discretized white noise is a piecewise constant function in space and time; a piecewise constant function is much smoother than the white noise random field it approximates.



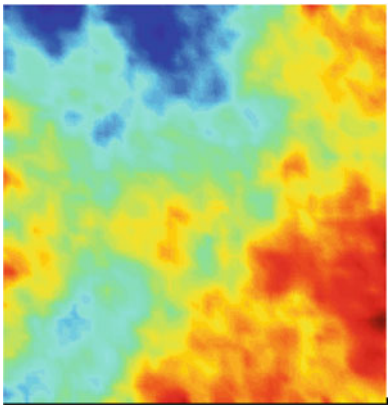
Realizations of discretized white noise at a same time interval in a square subdivided into 32, 128, and 512 triangles



Realizations of discretized white noise at two different time intervals in a square subdivided into the same number of triangles

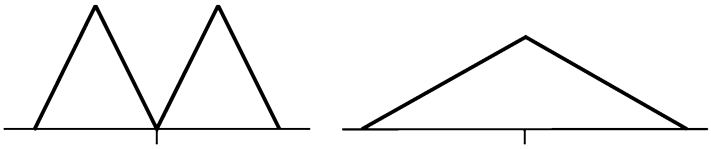
Discretizing Colored Noise In the colored noise case, one can use grid-based discretizations as well, but more often one takes advantage of the fact that one can express the random field in terms of infinite expansions in terms of orthogonal functions. There is more one way to do this, the most commonly used being *Karhunen-Loève expansions* that use the eigenvalues and eigenfunctions of the correlation function. As a result, the random fields are approximated by truncating the infinite expansions so that the approximate field then involves a finite number of parameters. In the case of colored noise, the number of parameters has a weaker dependence, compared to the white noise case, on the spatial/temporal grid sizes. The figure below provides an illustration of a realization of an approximated colored random field.

Realization of an approximated colored random field



A Little Sweeping-Under-the-Rug Part 1 We have tacitly assumed, as is often done, that we know the PDFs or other statistical information about the input

parameters. Actually, in practice, one usually does not know much about the statistics of the input variables. One is lucky if one knows a range of values, e.g., maximum and minimum values, for an input parameter in which case one often assumes that the parameter is uniformly distributed over that range. If one is luckier, one knows the mean and variance for the input parameter in which case one often assumes that the parameter is normally distributed. Of course, one may be completely wrong in assuming such simple probability distributions for a parameter as the figure below illustrates. This, as we have already noted, leads to the need to solve stochastic model calibration problems.



Two PDFs with the same mean and variance

A Little Sweeping-Under-the-Rug Part 2 Input variables could be distributed independently or jointly and could be correlated or uncorrelated. Without proper justification and sometimes incorrectly, it is often assumed that the parameters are independent. Based on empirical evidence, sometimes this is a justifiable assumption in the parameters-are-"knobs" case. But often, independence is a simplifying assumption that is invoked for the sake of convenience, e.g., because of a lack of knowledge.

Let us consider the case of correlated random fields. The Karhunen-Loève expansion does two wonderful things. First, it gives us a *formula*, albeit one with an infinite number of terms, that enables us to evaluate the random field at any point. Second, it expresses a correlated random field in terms of *uncorrelated parameters.* Unfortunately, what KL does not necessarily do is give us a formula involving *independent* parameters because although independence implies uncorrelated, uncorrelated does not necessarily imply independence as the following well-known example shows.

Let y_1 be uniformly distributed on $[-1, 1]$ so that the expected values $\mathbb{E}(y_1) = \int_{-1}^{1} y_1 \, dy_1 = 0$ and $\mathbb{E}(y_1^3) = \int_{-1}^{1} y_1^3 \, dy_1 = 0$. Now, let $y_2 = \frac{3}{2} y_1^2$. We have that the correlation $\mathbb{C}_{12} = \mathbb{E}(\boldsymbol{y}_1 \boldsymbol{y}_2) - \mathbb{E}(\boldsymbol{y}_1)\mathbb{E}(\boldsymbol{y}_2) = \frac{3}{2}\mathbb{E}(\boldsymbol{y}_1^3) = 0$ so that $\{\boldsymbol{y}_1, \boldsymbol{y}_2\}$ are uncorrelated. However, clearly $\{\boldsymbol{y}_1, \boldsymbol{y}_2\}$ are not independent. In fact, uncorrelated guarantees independence if and only if the variables follow a multivariate Gaussian distribution.

Revisiting Quantities of Interest We now can assume that we are given a random input parameterized in terms of a finite number of random parameters $\{y_1, y_2, \ldots, y_N\}$; we use the abbreviation $\boldsymbol{y} = \{y_1, y_2, \ldots, y_N\}$. These parameters

can be actual parameters appearing in the specification of the PDE or could arise from approximations of random field inputs.

The solution of the PDE is not only a function of the spatial variable $\mathbf{x}$ and perhaps also time t, it also a function of the random parameters $\boldsymbol{y}$. A *realization* is a solution $u(\mathbf{x}, t; \boldsymbol{y})$ of a PDE for a specific choice $\boldsymbol{y} = \{y_n\}_{n=1}^N$ of the random parameters.

There almost never is any interest in individual realizations. Instead, one is interested in *statistical information*, e.g., expected values, variances, standard deviations, higher moments, PDFs, CDFs. Most often one is not interested in such statistical information about the solution of the PDE itself. Instead, one is interested in statistical information about outputs depending on the solution of the PDE. The outputs of interest are often functionals of the solution of the PDE. For example, one is not interested in the velocity of a flow around a wing at points in the flow field but rather one is interested in things like the drag and lift on the wing. Again, one is almost never interested in individual realizations of such outputs but are interested in statistical information about such outputs which are referred to as *quantities of interest*.

The Curse of Dimensionality Applied and computational mathematicians live in one or two or three dimensions, with very rare excursions to higher dimensions. They think three dimensions is hard enough; even using today's supercomputers, some three dimensional problems cannot be addressed, e.g., the direct numerical simulations of practical turbulent flows. But, they are not well prepared to deal with the shock of having to find approximations of solutions $u(\mathbf{x}, t, \boldsymbol{y})$ of parameterized PDEs. Discretization has to be effected with respect to all three arguments so that one has to discretize in parameter space as well as in physical space and time and, moreover, the dimension of the parameter space, i.e., the number of parameters, may be large and certainly often much larger than three.

Statisticians are familiar with high-dimensional problems but they are not so used to dealing with problems for which realizations are very costly, as is the case when a realization involves the solution of a discretized PDE.

Let us leave PDEs alone for a minute and consider interpolation in N dimensions using polynomials of total degree at most p, e.g., for $N = 2$ and $p = 1$ we have the linear polynomial $a + by_1 + cy_2$ and also consider interpolation in N dimensions using tensor product polynomials of degree at most p in each direction, e.g., for $N = 2$, $p = 1$ we have the bilinear polynomial $a + by_1 + cy_2 + dy_1y_2$. For the same p, both types have same the approximation properties, i.e., for interpolating sufficiently smooth functions, the rate of convergence is the same.

What about the complexity? In the next table, one sees the *explosive* growth in the number degrees of freedom one needs for both types of approximations as the number N of parameters and the degree p of the polynomials increase. Note that, for the same rate of convergence, total degree interpolation is relatively much more economical compared to total tensor product interpolation but that even total degree interpolation suffers the explosive growth in the number of degrees of freedom, i.e., suffers from the curse of dimensionality.

Returning to the PDE setting, M in the table is the number of PDE solves one needs to determine the polynomial approximation. *The curse of dimensionality* refers to the explosive growth in the number of parameter degrees of freedom and therefore in the number of (costly) PDE solves needed for a certain accuracy as the number of parameters N and the degree p of the polynomials increases.

5 Approximation of Solutions of PDEs with Random Inputs

Stochastic finite element methods refer to the use of finite element methods (FEMs) to *spatially* discretize a PDE with random inputs. Of course, spatial discretization may be effected by any of a number of other methods, e.g., finite difference, finite volume, spectral, radial basis function, etc., so that, e.g., when using the first of these, one can also use the terminology "stochastic finite difference methods."

With respect to discretizing the parameter dependence of the problem, we discuss two approaches.

Global polynomial approximation in parameter space

$N =$ number of variables	$p =$ maximal degree of polynomials	$M =$ number of degrees of freedom	
		Using total degree polynomial basis	Using tensor product basis
3	3	20	64
	5	56	216
5	3	56	1024
	5	252	7776
10	3	286	1,048,576
	5	3003	60,046,176
20	3	1771	$>1\times10^{12}$
	5	53,130	$>3\times10^{15}$
100	3	176,851	$>1\times10^{60}$
	5	96,560,646	$>6\times10^{77}$
		$\Uparrow$ $\frac{(N+p)!}{N!p!}$	$\Uparrow$ $(p+1)^N$

Stochastic Galerkin Methods (SGMs) refer to effecting the discretization with respect to the random parameters using a Galerkin method. Galerkin methods are variational or projection methods which are commonly used for FEM and spectral method spatial discretization of the PDE.

The good news about SGMs is that only a single discrete system has to be solved to determine both the spatial/temporal and parameter dependencies of the

approximate solution. In particular, no parameter sampling is needed. As a result, one obtains an approximation that can be evaluated at any point in the parameter domain, e.g., to determine a quantity of interest.

The bad news is that the physical and parameter degrees of freedom are coupled. Thus if J denotes the number of finite element degrees of freedom and if M denotes the number of degrees of freedom used for parameter approximation, then the size of the discrete system that has to be solved is $JM \times JM$. Given than in a practical calculations J could be in the millions and we already saw that M could be gazillions, one can quickly get to systems of formidable size when using SGMs.

Another disadvantage of SGMs are that their implementation requires extensive recoding of a deterministic PDE solver because the discretizations of the spatial and parameter dependences are tightly coupled. For this reason, such methods are referred to as being *intrusive.*

The usual choice for effecting parameter approximations are orthogonal polynomials; in such cases, in the mathematical UQ community, the method unfortunately[3] goes by the name polynomial chaos. To take advantage of the high-accuracy of orthogonal polynomial approximations, whatever is approximated has be smooth with respect to the parameters.

Stochastic Sampling Methods (SSMs) refer to methods in which the PDE is solved at each member of a set of sample points in the parameter domain. Here, any spatial discretization of the PDE can be used because that discretization is uncoupled from how the parameter dependence is treated. For this reason, a deterministic PDE solver can be used as a black box for determining approximate solution at the selected parameter points. For this reason, SSMs are referred as being *non-intrusive.*

Stochastic sampling methods proceed as follows:

- sample M points $\{\boldsymbol{y}_m\}_{m=1}^M$ in the parameter domain Γ;
- for $m = 1, \ldots, M$, solve the spatial/temporal discretized PDE for each of the sample points;
- then use the M solutions so obtained to build ensemble averages or other statistical information about the approximate output of interest that depends on the approximate solution of the PDE.

Thus, to determine statistical information in the sampling setting one solves M discrete systems, each of size $J \times J$, compared to the stochastic Galerkin case for which one solves a single discrete system of size $JM \times JM$.

[3] *Polynomial chaos* was a term coined by Norbert Weiner when he studied PDEs driven by white noise and whose solution displayed chaotic behaviors. He expressed white noise random fields by truncated expansions in terms of orthogonal polynomials. In the mathematical and engineering UQ communities, polynomial chaos is used as a substitute name for orthogonal polynomial approximation, even though very few of the problems addressed by those communities have solutions that display any chaotic tendencies. We ourselves eschew the use of "polynomial chaos" and instead call it by what it actually is, namely, orthogonal polynomial approximation.

The list of possible sampling schemes is, of course, very, very, very long.[4] Obfuscating the situation is that what makes a set of points in parameter space "good or bad" depends on whether the points are used for quadrature, interpolation, regression, or some other purpose.

Sampling + Simple Averaging Methods for Quadrature By far, the most used sampling method for parameter domain quadrature is the *Monte Carlo* (MC) method for which one randomly chooses the set of sample points and simply averages the values of the integrand over those points. There is a lot of good news about MC. Perhaps uniquely among sampling strategies (and other approaches to UQ), the convergence rate of MC is independent of the number N of parameters so in this sense it does not suffer from the curse of dimensionality. Also, MC does not care about the smoothness of the integrand in the sense that the convergence rate is unaffected by smoothness. In addition, MC does not care much about the shape of the domain of integration. However, there is also bad news about MC. Convergence (which is in expectation) is very slow, with a rate $1/\sqrt{M}$, where M denotes the number of sample points. Furthermore, that remains the convergence rate regardless of how smooth is the integrand, i.e., MC cannot take advantage of any smoothness the integrand possesses.

The slow convergence of MC has spawned a huge industry aimed at devising alternative quadrature schemes that are "better" than MC, i.e., that converge faster, but which still are simple sampling and averaging schemes. Both deterministic and probabilistic, some sequential and some not, alternative sampling strategies have been invented. A non-exhaustive list includes variance reduction techniques, quasi-Monte Carlo sequences (e.g., Halton, Sobol, Faure, ... ad infinitum), Hammersley, Latin hypercube, importance sampling, stratified sampling, lattice sampling, orthogonal arrays, multilevel Monte Carlo, etc. For several of these methods, the error is proportional to $\frac{(\ln M)^{N+s}}{M}$ for some $s > 0$. For a "small" number of parameters N, one can ignore the logarithmic term and obtain close to linear convergence, i.e., $1/M$, which is a decided improvement over the $1/\sqrt{M}$ convergence rate of MC. However, for "large" N, the logarithmic term dominates so that the curse of dimensionality bites us again.

Interpolation and "better" Quadrature Rules It is well known that in one dimension and for smooth integrands one can do "better" compared to simple averaging rules $\frac{1}{M}\sum_{m=1}^{M} f(\boldsymbol{y}_m)$ by using weighted quadrature rules $\sum_{m=1}^{M} w_m f(\boldsymbol{y}_m)$, where the quadrature points $\{\boldsymbol{y}_m\}_{m=1}^{M}$ and weights $\{w_m\}_{m=1}^{M}$ are judiciously chosen. In one dimension, Gauss rules are beautiful examples of how to define "better" rules. This is why for a (very!) small number N of parameters, tensor products of Gauss rules have proven to be very useful. But, as we have seen, tensor products should be avoided like the plague, even for moderate N.

[4] There is even a non-intrusive version of SGMs, but that version is better viewed as a sampling method.

We consider *interpolatory quadrature rules* which are constructed by first constructing an interpolation method and the approximating an integral of a function by the integral of the interpolant of the function. Thus, the discussion here applies to both interpolation and to quadrature. In particular, we focus on polynomial interpolation and the quadrature rules that they engender.

Interpolants are built by requiring that they match the value of a function at sample points that, in this context, are referred to as interpolation points. When interpolants are used to define quadrature rules, the interpolation points become the quadrature points. Thus, we first must define "good" interpolants. Without any knowledge about the function being interpolated other than its smoothness, total degree interpolation requires the least number of interpolation points (i.e., in our context, the least number of PDE solves) to achieve the best rate of convergence possible by polynomial interpolation. Unfortunately, "good" interpolation points for total degree interpolation in as simple domain as a hypercube are not known, even in three dimensions, and there is some controversy about them even in two dimensions.[5] Fortunately, the along came Sergey Smolyak.

Stochastic Collocation or Sparse Grid Methods Smolyak (or sparse)[6] grids[7] are a judiciously chosen subset of tensor product grids. For the same precision, i.e., for integrating the same polynomial space exactly, sparse-grid quadrature rules require more points that do interpolatory quadrature rules based on total degree interpolation but require substantially fewer points than does tensor product interpolation or quadrature; see the table below. Note that as N and/or p increase, the gap between the number of total degree and sparse grid points grows quickly, but still at a much slower pace compared to the gap between sparse grids and tensor product grids. Note also that because total degree quadrature rules suffer from the curse of dimensionality, so do sparse grids rules.

[5]It is well known that in one dimension, evenly spaced points are "bad" interpolation points for general smooth function, bad because the interpolation error can get can get worse as the degree of the polynomial increase and the point spacing decreases. On the other hand, the unevenly spaced Chebyshev points are known to be ideal for the interpolation of smooth function in one dimension.

[6]In truth, any sampling method can be referred to as being a stochastic collocation methods because "stochastic" simply refers to the fact that we are dealing with random variables within some domain in parameter space and "collocation" simply means that we are evaluating the function, in our case the solution of the PDE, at points in the domain, ergo, we are sampling the solution at points in the parameter domain. However, stochastic collocation methods is now thought of as referring to a class of methods for which deterministic sampling is done on a structured set of points that are much fewer, e.g., much sparser, than, e.g., a tensor product of points, ergo, the synonymous moniker "sparse grids."

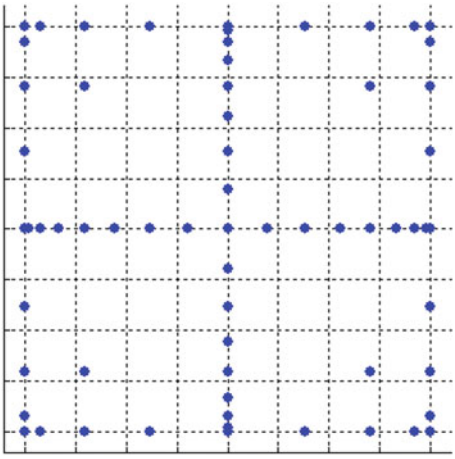
[7]The quadrature rules that use Smolyak of sparse grids as the quadrature points are not always interpolatory quadrature rules, but they are all built using combinations such rules.

For three types of grids in N-dimensional hypercubes, the degrees of freedom for quadrature rules having the same convergence behavior as total degree quadrature using polynomials of degree at most p

	Total degree		Sparse grid		Tensor product
Degrees of freedom	$\frac{(N+p)!}{N!p!}$	$<$	$O\left(p(\ln p)^{N-1}\right)$	$\ll$	$(p+1)^N$

The figure below is an illustration of a sparses grid. Note the big holes in the grid, i.e., the large areas containing no points. If the function being interpolated or integrated is very smooth, the big holes do not matter. For moderate parameter dimension and polynomial degree, sparse grids beat Monte Carlo, quasi-Monte Carlo, etc. but, of course, the curse of dimensionality rather quickly kicks in so that for quadrature,[8] sparse grids start losing to MC sampling because, as we have already mentions, MC does not suffer from the curse.

A 65 point sparse grid



The need for smoothness in sparse grid quadrature and the lack of such need for Monte Carlo quadrature is illustrated in the table below.

Comparison of sparse grid and Monte Carlo approximations of the integral of a discontinuous function

M	SG estimate	SG error	MC estimate	MC error
1	4.000	1.167	0.00000	5.16771
13	64.000	58.832	0.00000	5.16771
85	−42.667	47.834	3.01176	2.15595
389	−118.519	123.686	4.77121	0.39650
1457	148.250	143.082	5.15216	0.01555
Exact	5.16771	–	5.16771	–

[8] MC and QMC points are useless for interpolation.

6 Quo Vadis Uncertainty Quantification?

It is clear the what we have written above does not completely exorcize the curse of dimensionality, although in some situations, e.g., for smooth dependences on the random parameters, progress has been made in reducing the cost of UQ in the PDE setting, at least for a moderate number of parameters. However, there are efforts out there devoted to make further progress towards defeating the curse. In addition, there are other important aspects of UQ that we have not touched upon. We close by giving some very brief comments about some of these topics, with the comments ordered in decreasing length but not necessarily in decreasing importance.

Informed Sampling In the algorithms discussed so far, the sample points can be pre-selected to yield useful, i.e., efficient and accurate, approximations of function belonging to certain classes of functions such as functions with a certain number of continuous derivatives. However, they do not take into account any features of the specific function, e.g., an approximation of the solution of the PDE, that one wishes to integrate or approximate.

If one has some information about that function, as one often does, one can take advantage of that information to lessen the cost of integration or interpolation. For example, if one knows that some parameters are more influential than others, one can do anisotropic sampling so that there is a lower density of points in directions in parameter space that correspond to less important parameters compared to that for more important parameters.

Even before that, there are techniques available, e.g., screening methods, sensitivity analyses, etc., that can be used to determine the relative importance of the parameters.

Adaptive point selection is another promising approach. Here, one sequentially samples the point, with the position of a new point in parameter space selected so that it minimizes some function that can be computed from the current set of points and which in some (approximate) way represents the error that the use of the augmented set of points would induce.

Surrogates Our focus has been on becoming more efficient in approximating in parameter space. However, the large cost of approximate PDE solves, e.g., function evaluations in our setting, also contributes to the curse of dimensionality. The idea here is to use a relatively *few* approximate solutions of the PDE to construct a cheap-to-evaluate surrogate that can be used, instead of additional PDE solves, to provide the *huge* number of PDE solutions needed to do UQ. Actually, it is even more efficient to directly build surrogates for output of interest. Many types of surrogates approaches have been studied, including interpolation, least-squares approximation (regression), greedy algorithms, reduced-basis methods, proper orthogonal decomposition, etc.

Risk Assessment As has already been mentioned more than once, with the possible exception of simple sampling and averaging methods, what we have talked about, including orthogonal polynomial and sparse grid collocation methods, requires

smooth dependences on the parameter. As a result, these approaches have limited usefulness for quantifying risk assessment because that often involves integration of discontinuous functions. This includes, e.g., most approaches for determining probabilities of failure. In the UQ for PDE setting, little has been successfully done to move away from sampling and simple averaging approaches.

Compressed Sensing Compressed sensing is one of several approaches that try to determine the least number of terms that are needed in a polynomial approximation to achieve a certain accuracy. Of course, this has to be done without first having to compute all the terms in the polynomial and then throwing out those that are insignificant. Instead, a priori estimates for the coefficients are used to determine which terms to not include. The idea behind approaches such as compressed sensing is to reduce the number of terms below that needed for total degree interpolation without sacrificing accuracy.

Inverse Problems We have already mentioned stochastic optimization, identification, calibration, and control problems, all of which are obviously of great interest. Often these problems suffer from an even worse curse because of the possible need to compute the quantities of interest several times during a control or optimization process. Still, inverse problems constrained by PDEs with random inputs is a quickly growing industry. Baysean approaches are quite seductive and enjoy widespread popularity, but direct optimization approaches are also used.

Rare Events Quantifying rare events is another task for which not much progress has been made in the UQ for PDEs setting, a setting in which function evaluations are very expensive. The techniques being used are mostly standard, well-known ones in the statistics community.

Acknowledgements The author "Max Gunzburger" was supported in part by the US Department of Energy Office of Advanced Scientific Computing Research and the US Air Force Office of Scientific Research.

References

1. Ghanem R, Higdon D, Owhadi H (eds) (2017) Handbook of uncertainty quantification. Springer, Berlin
2. Gunzburger M, Webster C, Zhang G (2014) Stochastic finite element methods for partial differential equations with random input data. Acta Numer 23:521–650
3. Smith R (2014) Uncertainty quantification: theory, implementation, and applications. SIAM, Philadelphia
4. Soize C, (2017) Uncertainty quantification: an accelerated course with advanced applications in computational engineering. Springer, Berlin
5. Sullivan R (2015) Introduction to uncertainty quantification. Springer, Berlin

Dynamical Symmetries and Model Validation

Benjamin C. Jantzen

Abstract I introduce a new method for validating models—including stochastic models—that gets at the reliability of a model's predictions under intervention or manipulation of its inputs and not merely at its predictive reliability under passive observation. The method is derived from philosophical work on natural kinds, and turns on comparing the *dynamical symmetries* of a model with those of its target, where dynamical symmetries are interventions on model variables that commute with time evolution. I demonstrate that this method succeeds in testing aspects of model validity for which few other tools exist.

1 Introduction

Scientists of all stripes are in the business of building models as tools for predicting, controlling, and explaining phenomena. For each of these purposes, it is generally not sufficient that a model merely "save the phenomena." That is, it's not enough that a model successfully summarize the data already in hand. Rather, the model builder wants some sort of assurance that the model accurately represents the world, at least with respect to those features pertinent to her epistemic goals. The most common approaches to establishing such a warrant of reliability focus on comparing features of the solutions or predictions of a model and states (or time-series of states) of the world. For example, one would typically validate a regression model that predicts lifetime earnings on the basis of socioeconomic factors like education by comparing the model's predictions against a collection of fresh data not used in its construction. The more such predictions match, the more confident we are in the model, at least as a tool for prediction.

While the bulk of statistical tools are designed with such comparisons in mind, there is often value in comparing relations amongst accessible states (or relations

B. C. Jantzen (✉)
Department of Philosophy, Virginia Tech, Blacksburg, VA, USA
e-mail: bjantzen@vt.edu

© Springer Science+Business Media, LLC, part of Springer Nature 2019

N. Fillion et al. (eds.), *Algorithms and Complexity in Mathematics, Epistemology, and Science*, Fields Institute Communications 82,
https://doi.org/10.1007/978-1-4939-9051-1_6

amongst time-series of states) rather than states themselves. That is, it can be helpful to ask whether the *change* in a model's prediction given a *change* in input or initial conditions matches the change in the target system given a corresponding change in it's initial conditions or external inputs. This sort of relation is exactly what one needs to know if a model is to be used for control. That is, if one wants a model to reliably reflect causal relations amongst variables—as opposed to offering purely correlative predictions—it is essential to verify that its gets these relations right.

In this paper, I introduce a new approach to validating dynamical models—including stochastic models—using 'dynamical symmetries'. This method is focused not on static features of states or time-series, but rather on relations amongst such things under changes in input or initial conditions. This is a tool for checking the causal information explicitly or implicitly contained in a model, and is therefore useful for validating models for control as well as for prediction. My narrow aim is to argue that this method is, in many circumstances, an unusually powerful tool for model validation that gets at causal structure in a way most statistical methods do not. More broadly, I want to suggest that the success of this method is evidence of the practical, methodological relevance of philosophical work on natural kinds. Consequently, the technical results presented here amount to a sort of advertisement for the mutual benefits philosophers, applied mathematicians, and data analysts can offer one another.

To meet these aims, the rest of the essay is laid out as follows. In Sect. 2, I clarify the problem of model validation, and present a concise but somewhat more detailed overview of standard methods that focus on comparing static properties of model solutions or predictions with single measurements of the target system. This is then contrasted with what I call "structural approaches" that consider relations amongst model predictions. I summarize a variety of methods that are plausibly viewed as structural. In Sects. 3 and 4, I introduce the theory of dynamical symmetries, and present previously published methods for comparing them for different systems given empirical data. In Sect. 5, I outline the way in which comparison of dynamical symmetries can be used as a powerful tool for model validation, and illustrate the method in a variety of contexts with concrete examples. Finally, I conclude with a discussion of the scope and limitations of this new method.

2 The Problem of Validation

2.1 Verification and Validation

Any model—if it is to be useful for predicting or controlling its target system—needs both verification and validation. Verification is the process of assessing whether a given model possesses the intended properties. That is, does the actual instrument or mechanism for generating predictions instantiate that which was intended; do the outputs instantiate the intended mapping from inputs? Verification

is not a significant concern in the case of analytic models, since there is generally little doubt that a set of equations is in fact the set intended. It becomes critical, however, when numerical approximations are used in extracting predictions or solutions from the equations, and even more pressing in the case of complex computational models. It is not at all obvious that a program correctly implements the numerical integration of a set of equations, or represents the intended set of functional relationships between variables. It is even less clear whether a multi-physics or agent-based model captures the intended set of approximations of law-like interactions amongst constituents. Verification poses a fascinating collection of epistemic problems, and there exist large literatures on model verification in engineering, software development, and mathematics.[1] However, for the purposes of this essay, I'll set aside the problem of verification, and assume that models are correctly implemented.

Validation concerns the accuracy of the model in representing the intended aspects of the target. There are three principal senses in which a model can be accurate. First, it can more or less successfully describe the target system. That is, it can reproduce the known data with varying degrees of fidelity. Second, it can more or less accurately predict passively observed features of the target system at later times, or for different boundary conditions or inputs. And finally, it can more or less accurately predict the behavior of the target system under interventions or alterations of the environment, boundary conditions, or input. The first two tend to be the focus of many standard statistical methods. The novel method described below concerns the third.

There are, of course, many kinds of models with distinct epistemic aims, and the suitability of a particular method of validation will depend on the kind of model under consideration. The method below is appropriate for models in which each variable varies continuously, possibly as a function of other variables in the system. Of particular interest are dynamical systems, i.e., systems that change through time. The bulk of the examples below will deal with systems for which the values of variables change continuously over time.

A brief word on terminology is also important. The term "model" tends to be used in subtly different ways in different disciplines—ecologists tend to use the term differently than machine learning practitioners, and the same group tends to use the term differently depending on context. For clarity, I will use the term as follows: a *model* is the specification of a class of mappings from input to output.[2] The mappings may be via an explicit function or set of equations (as in differential

[1]For an influential engineering perspective, see [1]. For a recent and comprehensive overview of both software and systems modeling aspects from the National Research Council, see [5]. For a pithy and very current overview of verification in the world of software design, see [26]. Finally, for an accessible and illuminating discussion of the state of the art from the perspective of applied mathematics, see [7].

[2]Note that this terminology is at odds with machine learning, where each specific set of parameter values constitutes a model. What I'm calling a model is, in the context of machine learning, or statistical learning theory a space of hypotheses or class of models.

models of thermodynamic phenomena) or via simulations of varying complexity. To produce an output, a model requires two things: (1) a set of parameter values, and (2) a set of inputs. Parameters are understood to represent features of the target system that may vary from system to system but do not vary for a given system. The intrinsic growth rate of a population or the Young's modulus of a given material are examples of parameters. Sometimes these can be measured independently, but often have to be estimated from data about other properties of the system that depend upon these parameters. Inputs are a set of initial or boundary conditions that can differ across time or contexts for a given system. The temperature of a reaction vessel or the current population size are typical inputs for chemical engineering or ecological models, respectively. I will refer to a particular output for a given input and choice of parameter values as a *solution* of a model.

2.2 Static Fit Approaches to Validation

There are a wide variety of methods for model validation that appeal to a single output of the model and one or more datasets. If the model is designed to make point estimates—e.g., of a particular property of the target system such as the ionization energy of a molecule, the vibration frequency of a nano-beam, or the biodiversity of a region—then a wealth of standard statistical hypothesis testing methods are available. In the classical mode, these involve choosing a test statistic (e.g., χ^2), computing the distribution this statistic should have under the hypothesis that the model is correct, and then deciding whether the value of the test statistic for a set of validation data is sufficiently unlikely to reject the adequacy of the model. The procedure and interpretation of results are a bit different from a Bayesian point of view, but the emphasis remains on comparing a single model output with a validation data set. See [15] for a recent overview of both approaches to validating models in the case of point estimates. What's relevant to the discussion here is that no consideration is given to the answers the model *would* give under different inputs. That is, the model typically has parameter values and boundary conditions estimated from one data set, and a single point estimate is tested with a second dataset from the same target system in the same configuration. There is no attention paid to how the model's estimate varies with variations in properties or initial states of the target system.

Of course, point estimates are just one special class of model output. Many models explicitly represent one or more functional relations amongst variables. The simplest such model is a regression curve (a functional form fit to a single dataset), but one could include simulations and agent-based models in this category. For models like these, it is typical to construct the model using a single data set describing the target system under one set of initial conditions. For example, one could measure the growth of bacteria on a petri dish over time, and then use that data to assign values to parameters in a model that proposes an exponential functional relationship between time and population. Once the parameters of the model have been set, the predicted curves relating variables of the model are compared, either to

the original dataset or to a validation dataset taken under identical conditions (often, one simply splits the original dataset into training and testing pieces).

There are a variety of approaches to comparing the curve predicted by a model with data from a target system. Commonly, simple measures of agreement, such as the Sum of Squared Errors (SSE) or the coefficient of determination (R^2) are deployed to assess how well the model captures the variation in the data. For example, Fujikawa et al. [9] use the Mean Squared Error (MSE)—given by SSE/n where n is the number of samples—as a measure of goodness of fit for their growth model.

By themselves, these measures of fit only get at how well the model describes the target system (in one particular context). This is the first, descriptive sense of validity I mentioned above. To get a sense for how well the model is likely to generalize (how reliable it will be for prediction) we need other tools. Analysis of residuals is one such tool. More specifically, the distribution of the errors (the differences between values predicted by a model and the actual data) can tell one a lot about whether there is systematic error in the model of the sort that would impugn its ability to make accurate predictions outside the original data. Methods of residual analysis include hypothesis tests for bias (the errors all tend to be in one direction), skew (there is a trend in the errors, even though there may be no bias), and curvature (there may be no bias or skew, but the envelope of the errors exhibits curvature around the correct values) (see [19, ch. 16] for a concise overview).

When we turn our attention to stochastic models, validation gets more complicated, at least insofar as we continue to directly compare particular model outputs with acquired data. This class of model has received less attention with respect to methods of validation, and the literature on modeling across disciplines harbors a consensus that it's difficult, particularly when data is limited. This difficulty seems to have stymied the emergence of anything that would be considered a standard method. As McCarthy and Broome [17, p. 600] put it, "there are no established methods for validating stochastic population models, but useful methods are required."

Some useful methods include generalizations of those described above for the deterministic case. In the most straightforward approach, one would compare predicted distributions of the dependent variables (e.g., population size) for each value of the independent variables (e.g., time) with the observed data. But this requires many replicates of the target system so that such distributions can be estimated. That sort of data is usually not forthcoming. Typically, all one has for comparison is one or a few series of measurements, with at best a handful of measurements for each value of the independent variable. But a variety of approaches have been proposed to overcome this difficulty. For example, Sokal and Rohlf [21] propose a method of 'standard deviates' for assessing whether the stochastic variation predicted by a model coincides with a dataset. This is, in fact, the method used by McCarthy and Broome [17] for a model of population viability (i.e., of the risk that a population will go extinct). The salient point is that this method and others like it still focus on features of a single output of a model.

This is not quite the case for cross-validation, a powerful tool for validating both deterministic and stochastic models. This technique iterates partitioning of the data into training and testing portions in order to estimate the error of a model on unobserved data.[3] Roughly, cross-validation provides an estimate of the generalization error of a model[4] by assessing how well a model, after being fit to a sample of data from a system, will do in predicting unseen data. Cross-validation thus does not focus on a single solution of a model, but rather the reliability of the model (and the method for setting its parameter values). Nonetheless, it is indifferent to the way in which model solutions relate to one another. Similarly, the method of "active nonlinear tests" [18] amounts to probing the space of parameter values and inputs to assess the robustness of features of a model's output to variations in parameter values, and the model's stability and plausibility for inputs not observed. Here again, there is no attention paid to the details of how solutions relate to one another, only how robust a given solution is to variations of model features.

Why is this problematic? If the aim is prediction, it's not a problem at all. These are all effective approaches to predictive validation. But if one wants to be confident that a model which fits a given data set will get its predictions right when someone intervenes and changes the boundary conditions or inputs, more is needed.

2.3 Structural Approaches to Validation

Some models do aspire to capture more about a target system than is necessary to predict its behavior under passive observation. In particular, some models are intended to capture something about the structure of a target system, and the extent to which they do so has been called the "structural validity" of a model (see, e.g., Zeigler et al. [28, ch. 2]).[5] In the dynamical systems, engineering, and operations research literatures, the notion of structural validity seems to have a rather narrow and stringent sense. A model is only structurally valid if the structure of the model is isomorphic to that of the target system. As Zeigler et al. [28, p. 31] put it, saying that a model is structurally valid "…means that the model not only is capable of replicating the data observed from the system, but also mimics in step-by-step, component-by-component fashion the way in which the system does its transitions."

[3]Most textbooks on machine learning include descriptions of cross-validation. An especially lucid presentation can be found in Flach [8, ch. 12].

[4]The estimate of the generalization error of a model is biased for cross-validation, but in the direction of over-estimating the error (see [11, ch. 7.10]).

[5]Attention to structural validation is curiously discipline dependent. Concepts (such as those pertaining to testing "white-box" models in systems engineering) seem to have relatively little penetration in other fields such as ecology. This is probably partly due to the quantity and precision of data available in these different fields. Structural tests tend to be data-hungry or to require manipulations of the target system that are not available to, e.g., field ecologists.

This sort of validity can be assessed in a variety of ways.[6] What Barlas [3] calls "structure-oriented behavior tests" include a variety of comparisons of those qualitative features of a model thought to be tied to its structure with those of the target system. For example, one can assess how well a model captures temporal patterns in the target system such as the period, phase, and amplitude of oscillatory behavior, or the presence of trends [2]. A failure of the model to generate periodic behavior of approximately the right frequency, for instance, might suggest that a feedback in the structure of the model is incorrect. Another sort of structure-oriented behavior test involves assigning extreme values to model inputs or parameters and comparing the resulting output to the behavior of the target system under correspondingly extreme conditions.[7] Note that to conduct this sort of test with respect to parameter values, the parameters must be meaningful (and both measurable and manipulable) outside of the model.

In "direct tests", one attempts to establish the accuracy of structural components of the model (e.g., the existence and values of certain parameters) by directly testing hypotheses about these components against experiments on the target system, or even established knowledge in the relevant field. For example, one might attempt to empirically ascertain whether the form of the equations in an equation-based model match the functional form of the relations among variables in the target system [3].

Whether by direct or indirect approaches, structural validation in the narrow sense is often the wrong epistemic goal. Narrow-sense structural validity is frequently more than one needs to meet the epistemic aims of modelers. That is, there is a broader sense of structural validity that gets at what a model needs in order to accurately characterize the behavior of system under intervention or manipulation, and nothing more. In this broader sense, a model is structurally valid if it correctly characterizes the *change* in a behavior of a system under changes in inputs or boundary conditions. Whether the model does so in the same way the target system does is irrelevant.

Presumably, if a model is structurally valid in the narrow sense, then it is structurally valid in this broader sense as well. But the broader sense is easier to satisfy in that it doesn't matter how a model captures this information, only that it does. Consequently, tests that reject this sort of validity rule out a bigger class of potential models in one go. And yet, so far as I can tell, it is largely neglected in the model validation literature. Of course, the entire field of causal discovery is concerned with methods for *building* models that capture structure in something like this broad sense (see, e.g., [22]). But those models tend not to capture the sort of fine-grained temporal detail that engineers or dynamical systems folks are interested in. Nor do methods of causal discovery directly help us to validate existing models that use, e.g., differential equations or complex agent-based computations. When I say that scant attention is paid to broad-sense structural validity, I mean there

[6] See [3] for a widely-cited review.

[7] Balci [1] calls this "stress testing."

are few if any tools in the modeling literature for validating models of arbitrary structure—especially dynamical models—with respect to counterfactual behavior. No one looks at which *changes* in model behavior follow from changing conditions or inputs, and whether this pattern of change (reflective of causal structure) matches the world. This is, however, exactly what a comparison of dynamical symmetries can do for us.

3 Dynamical Symmetries

3.1 Theory

As I indicated above, the new approach to validation described here is focused on the structure of a model or, more specifically, on the relations among solutions of a model that are implied by its structure. The important set of relations are what I previously dubbed *dynamical symmetries* [13]. Qualitatively, a dynamical symmetry is an intervention on one or more variables in a system that commutes with the incrementation of another variable in the system. More precisely, I define a dynamical symmetry as follows [14]:

Definition 1 (Dynamical Symmetry) Let V be a set of variables and Ω be the space of states that can be jointly realized by the variables in V. Let $\sigma : \Omega \rightarrow \Omega$ be an intervention[8] on the variables in $Int \subset V$. The transformation σ is a dynamical symmetry with respect to some index variable $X \in V - Int$ if and only if σ has the following property: for all values x_i and x_f of X and for all initial states $\omega_i \in \Omega$, the final state of the system $\tilde{\omega_f} \in \Omega$ is the same whether σ is applied when $X = x_i$ and then an intervention $\Lambda_{x_i,x_f} : \Omega \rightarrow \Omega$ on X makes it such that $X = x_f$, or the intervention on X is applied first, changing its value from x_i to x_f, and then σ is applied. This property is represented by the following commutation diagram:

$$\begin{array}{ccc} \omega_i & \xrightarrow{\sigma} & \tilde{\omega}_i \\ {\scriptstyle \Lambda_{x_i,x_f}}\downarrow & & \downarrow{\scriptstyle \Lambda_{x_i,x_f}} \\ \omega_f & \xrightarrow{\sigma} & \tilde{\omega}_f \end{array} \tag{1}$$

[8] As indicated in [13], I am using the term "intervention" in its technical sense as it appears in the literature on causation. In this context, "…an intervention on X (with respect to Y) is a causal process that directly changes the value of X in such a way that, if a change in the value of Y should occur, it will occur only through the change in the value of X and not in some other way"[27].

For example, suppose we have a pressure tank full of fluid and attached to a pump that can increase or decrease the pressure in the tank. Inside the fluid-filled pressure tank, there is a vertical rail on which is mounted a pressure gauge. Initially, this gauge is at the top of the tank where the pressure is P. If we use h to represent the depth of the gauge relative to the top of the tank and p to indicate the pressure read by the gauge, then at the outset, $h = 0$ and $p = P$. Now consider two different sequences of interventions on this system. In the first, we leave the gauge where it is, and then turn on the pump until the pressure at the gauge is $P + c$. Then we lower the gauge until it is a distance h_f below the top of the tank. At that point, it reads a pressure of $P + c + \rho g h_f$, where ρ is the density of the fluid in the tank and g is the gravitational constant ($9.81\,\mathrm{ms}^{-2}$). This sequence of manipulations and results is summarized in Table 1.

Now suppose that we start over with our tank in the same initial state, and reverse the order in which we manipulate the pump and the gauge. That is, suppose we first lower the gauge so that its depth relative to the top of the tank goes from 0 to h_f and then turn on the pump to increase the pressure at the gauge by an amount c. As Table 2 indicates, we end up in exactly the same final state after performing these actions. Thus, increasing the pressure at the gauge by an additive constant is a dynamical symmetry with respect to the index variable h. Note, however, that scaling pressure by a multiplicative constant (i.e., an intervention of the functional form $\sigma(P) = kP$) is *not* a dynamical symmetry. The result of applying transformations of this sort in either order with respect to moving the gauge is shown in Tables 3 and 4. Unlike in the additive case, the bottom rows of these two tables are not the same.

Table 1 Sequence of states when pressure is adjusted by an additive constant first and then the gauge is lowered a vertical distance h_f

p	h
P	0
$P + c$	0
$P + c + \rho g h_f$	h_f

Table 2 Sequence of states when the gauge is first lowered a vertical distance h_f and then the pressure is adjusted by an additive constant

p	h
P	0
$P + \rho g h_f$	h_f
$P + c + \rho g h_f$	h_f

Table 3 Sequence of states when pressure is adjusted by a multiplicative constant first and then the gauge is lowered a vertical distance h_f

p	h
P	0
kP	0
$kP + \rho g h_f$	h_f

Table 4 Sequence of states when the gauge is first lowered a vertical distance h_f and then the pressure is adjusted by a multiplicative constant

p	h
P	0
$P + \rho g h_f$	h_f
$k(P + \rho g h_f)$	h_f

Since many models of interest are models of dynamical systems in the more restrictive sense of variables that evolve through time under a fixed law, I offer the following definition of a special dynamical symmetry [14]:

Definition 2 (Dynamical Symmetry with Respect to Time) Let t be the variable representing time, and let V be a set of additional dynamical variables such that $t \notin V$ and Ω is the space of states that can be jointly realized by the variables in V. Let $\sigma : \Omega \to \Omega$ be an intervention on the variables in $Int \subseteq V$, and Λ_{t_0,t_1} the time-evolution operator that advances the state of the system from t_0 to t_1. The transformation σ is a dynamical symmetry with respect to time if and only if for all intervals Δt and initial states $\omega_i \in \Omega$, the final state of the system $\tilde{\omega_f} \in \Omega$ is the same whether σ is applied at some time t_0 and the system evolved until $t_0 + \Delta t$, or the system first allowed to evolve from t_0 to $t_0 + \Delta t$ and then σ is applied. This property is represented by the following commutation diagram:

$$\begin{array}{ccc} \omega_i & \xrightarrow{\sigma} & \tilde{\omega}_i \\ {\scriptstyle \Lambda_{t_0,t_0+\Delta}}\downarrow & & \downarrow{\scriptstyle \Lambda_{t_0,t_0+\Delta}} \\ \omega_f & \xrightarrow{\sigma} & \tilde{\omega}_f \end{array} \tag{2}$$

For example, consider a microbial population whose growth is governed by:

$$\frac{dx}{dt} = rx\left(1 - (x/k)^2\right). \tag{3}$$

Such a population exhibits a whole family of dynamical symmetries with respect to time. Specifically, if we take an initial population of x_0 and add or subtract enough microbial stock to raise the population to $\tilde{x}(x_0)$, where

$$\tilde{x}(x) = \frac{ke^{pk^2}x}{\sqrt{k^2 - x^2 + e^{2pk^2}x^2}}, \tag{4}$$

for any real value of p, and then allow the colony to grow for an hour, we would end up with the same final population size as if we allowed the population to grow for an hour starting from x_0 and then added (or subtracted) enough to scale the result according to Eq. (4) (with the same value of p).

3.2 Motivation and Generalization

The dynamical symmetries of a system depend upon and thus reflect its detailed causal structure. But dynamical symmetries are just one sort of feature of the causal structure of a model, and there are indefinitely many other features of causal structure that one could deploy for structural validation. So why focus on this one? There are at least three reasons to do so. The first is theoretical relevance. The notion of a dynamical symmetry is central to a general theory of *projectible kinds* [13]. Projectible kinds are categories or ways of binning portions of the world that are narrow enough that the members of a category share sufficient features in common to support generalizations of the sort we tend to call laws of nature, but broad enough to encompass sufficient variety in the world to make the law useful. In [13], I propose that we use symmetry structures—collections of dynamical symmetries along with an algebra describing how these dynamical symmetries interact under composition—to pick out projectible kinds. Two systems belong to the same projectible kind (what I call a *dynamical kind*) just if they exhibit all of the same dynamical symmetries, and these dynamical symmetries compose with one another in the same way. The categories picked out by dynamical kinds align well with those carved out informally by scientific practice. For example, the categories corresponding to the order of a chemical reaction are also dynamical kinds. So all reacting systems that obey a first-order reaction rate law belong to the same dynamical kind. I argue in [13] that recognizing dynamical kinds as the sort of projectible kinds scientists are after offers a variety of advantages for automated scientific discovery. In particular, systems can be sorted into kinds without first learning detailed models of their dynamics. Thus, one can learn how to delineate a new scientific domain pre-theoretically. The details are beyond the scope of our present concerns, but the point is that a focus on dynamical symmetries in model validation is not arbitrary. Rather, it is motivated by a broader program in the logic of scientific discovery.

The second reason is the specific nature of the relation of dynamical symmetries to causal structure. In addition to the bare causal skeleton of which variable is a cause of which, dynamical symmetries are sensitive to the functional form of relations amongst variables. This makes them a discriminating tool for comparing models with target systems in a manner relevant to fine-grained prediction *and* control.

The final reason for emphasizing dynamical symmetries, and perhaps the most practically salient, is the extensibility of the concept. I'll focus on one particularly important extension of the basic notion of a dynamical symmetry: stochastic systems. As indicated above, stochastic models of (presumably stochastic) target systems are difficult to validate with respect to their predictive reliability. This is because there are more dimensions to a model's output—where before we had point values or trajectories of point values over time, now we have distributions characterized by indefinitely many non-vanishing moments (e.g., mean, variance, skew, etc.). Validating such models with respect to structure is even harder. But

dynamical symmetries can be generalized to the stochastic case in a way that makes their application to validation straightforward.

So how do we extend the notion of dynamical symmetry beyond the deterministic case? In [14], I provide one proposal. Specifically, Definition 5 of that paper shifts the focus from values of variables to distributions over variables. However, in hindsight it's clear that Definition 5 is ambiguous in important respects. I thus offer the following refinement:

Definition 3 (Dynamical Symmetry) Let V be a set of random variables, Ω the set of states that can be jointly realized by the variables in V, and Γ the space of probability distributions over Ω. Let $\sigma : \Gamma \rightarrow \Gamma$ be an intervention on the variables in $Int \subset V$. The transformation σ is a dynamical symmetry with respect to some index variable $X \in V - Int$ if and only if σ has the following property: for all initial joint distributions $\gamma_i \in \Gamma$ and marginal probability distributions f and g, the final joint probability distribution over V, $\tilde{\gamma_f} \in \Gamma$, is the same whether σ is applied when the marginal distribution over X is given by $p_x(x) = f(x)$ and then an intervention $\Lambda_{f(x),g(x)} : \Gamma \rightarrow \Gamma$ on X makes it such that $p_x(x) = g(x)$, or the intervention on X is applied first, changing its marginal distribution from $f(x)$ to $g(x)$, and then σ is applied. This property is represented in the following commutation diagram:

$$\begin{array}{ccc} \gamma_i & \xrightarrow{\sigma} & \tilde{\gamma}_i \\ {\scriptstyle \Lambda_{f(x),g(x)}}\downarrow & & \downarrow{\scriptstyle \Lambda_{f(x),g(x)}} \\ \gamma_f & \xrightarrow{\sigma} & \tilde{\gamma}_f \end{array} \tag{5}$$

Note that this definition captures the deterministic dynamical symmetries as a special case (at least insofar as one is willing to entertain degenerate probability distributions). As we'll see below, this more general notion of dynamical symmetry is useful because it allows us to check the causal structure of a model against that of a target system, even when the underlying dynamics is fundamentally stochastic.

4 Comparing Dynamical Symmetries

The dynamical symmetries of two systems can be directly compared without first learning a detailed model of how the variables of either system interact. The first published algorithm to implement such a test appears in [14]. To use the algorithm one must, of course, first obtain data about the dynamical symmetries to be compared. The most direct way to do so is to acquire two time series for System A (and two more for System B) starting at two different initial values. The initial values, let's call them x_0 and $\tilde{x}_0$, must be the same for A and B, though of

course the rest of the time series may differ between them.[9] It is a consequence of the definition of a dynamical symmetry that, for systems that are deterministic, the function which maps the points of one time series to the points of the other time series corresponding to the same time is a dynamical symmetry. Furthermore, any two symmetry functions of a given system that agree on the initial values (any symmetry functions that map x_0 to $\tilde{x}_0$) must agree for the rest of the time-series.

The algorithm I reported in [14] compares the dynamical symmetries exhibited by System A and System B using such pairs of time series. In broad strokes, the algorithm involves nested cross-validations. Cross-validation in general involves dividing the available data into training and testing portions. In tenfold cross-validation, one partitions the data into ten segments, nine of which are used for training and one of which is set aside for testing. With the training data, a particular solution of the model is fit. Then the fit model is used to predict the testing data, and the squared errors of these predictions are saved. Then the process is repeated using a different element of the partition as the testing data and the remaining nine elements for training. After each of the ten data segments has been used once as the testing data, the mean of the accumulated squared errors (MSE) is used as an estimate of the error of the model (or really, of the model plus the method used for fitting a solution).

In my algorithm, the outer cross-validation loop estimates the errors for two different models trained on data reflecting the dynamical symmetries exhibited by two systems of interest (call them A and B). The first model—called *sep* for "separate"—assumes that the data represent two different symmetries. That is, *sep* fits the data from A and B with two different and independent sets of parameters. The other model—called *joint*—assumes that the data from systems A and B derive from the very same dynamical symmetry, and fits a solution involving only a single set of parameters. The inner cross-validation loop is used for fitting polynomial models to the training data. Specifically, cross-validation is used to choose the order of the polynomial that should be fit to the data. Higher orders can fit a training set better but generalize poorly (in statistical parlance, they 'overfit' the data), and lower orders ignore salient variations (they are overly 'biased'). When the outer cross-validation is complete, the algorithm declares the symmetries to be different just if the MSE of the *joint* model is significantly larger than the *sep* model. That is, the dynamical symmetries are judged to be different if cross-validation estimates a higher error when the data are treated as coming from a single function than when they are treated as separate.

This algorithm was originally developed to compare symmetries of two physical systems. In the next section, I demonstrate how it can also be used to structurally validate a model by comparing the symmetries of the model with those of the target system.

[9] In principle, one could take a single long time series for each system and cut it in half to obtain two such curves, but for ease of exposition, I assume the time series are obtained separately.

5 Dynamical Kinds and Model Validation

5.1 Growth Models

To provide a concrete sense for how dynamical symmetries can contribute to model validation, I present three case studies in this section. In each case, the target system involves biological growth of a single species. More specifically, the models I'll consider are aimed at predicting population size—of mammals or microbes—as a function of time for a given environment. Perhaps the most influential model of this sort was published in the early nineteenth century by Verhulst [24].[10] In Verhulst's "logistic model", the instantaneous rate of population growth is proportional to a quadratic function of the current population:

$$\frac{dx}{dt} = rx\left(1 - \frac{x}{K}\right). \tag{6}$$

The parameter r is generally interpreted as representing fecundity (or intrinsic growth rate) and K is viewed as the carrying capacity (the maximum sustainable population). The solutions of this equation are curves with a familiar sigmoid shape—they begin with a nearly exponential phase, pass through an inflection point, and level off in an asymptote to the carrying capacity. It is important to note that, although r and K can be given a biological interpretation, they are in general not directly measurable, and must be estimated by fitting one of these sigmoidal curves to the data.

Verhulst's original model has spawned a menagerie of generalized, extended, or otherwise modified logistic models. The bulk of these can be gathered under a single class of models that [23] call "generalized logistic" functions.[11] These have the form,

$$\frac{dx}{dt} = rx^{\alpha}\left(1 - \left(\frac{x}{K}\right)^{\beta}\right)^{\gamma}, \tag{7}$$

where the additional parameters α, β, and γ have no obvious biological interpretation. For parameter values not too far from 1 (e.g., $\alpha = \beta = 1$; $\gamma = 2$), the solutions of generalized logistic models are only subtly different in shape from the original Verhulst model, at least when one is free to choose values of r and K (see [23] for a thorough review). This fact—coupled with the fact that r and K cannot be independently measured or estimated—leads to a profound underdetermination and a persistent problem for model validation. Which model is the right model of population growth for a given species in a given context? Lest the reader get

[10] For an English translation of the French, see [25].

[11] Another equally old and venerable model is that of Gompertz [10]. This model also continues to be deployed for growth modeling.

the impression that this question is merely academic and this example merely a "toy", note that papers continue to be published in biological and industrial process journals addressing this question [4, 9, 29]. Researchers actually want to know the answer so that they can not only predict but control and optimize the growth of, e.g., microbial stock species or virulent microbial contaminants. In the case studies that follow, I demonstrate how methods of assessing the sameness of dynamical symmetries can aid in model selection in the context of bacterial growth.

5.2 *Example: Deterministic Generalized Logistic Models*

In the first case, consider a simulated population whose actual growth is deterministic and dictated by a generalized logistic equation with $\alpha = 1$, $\beta = 3/2$, $\gamma = 2$. We can use this simulated population to generate data for which we know the ground truth. In Fig. 1a, you can see two samples from this system—for two different starting populations—where Gaussian noise of standard deviation 0.3 has been added in order to accurately reflect the noise inherent in measurement.

Now imagine yourself as a researcher interested in learning the "right" model of population growth. For one reason or another, you've decided to consider two models: the Verhulst logistic equation (Eq. (6)) and a generalized logistic (Eq. (7))

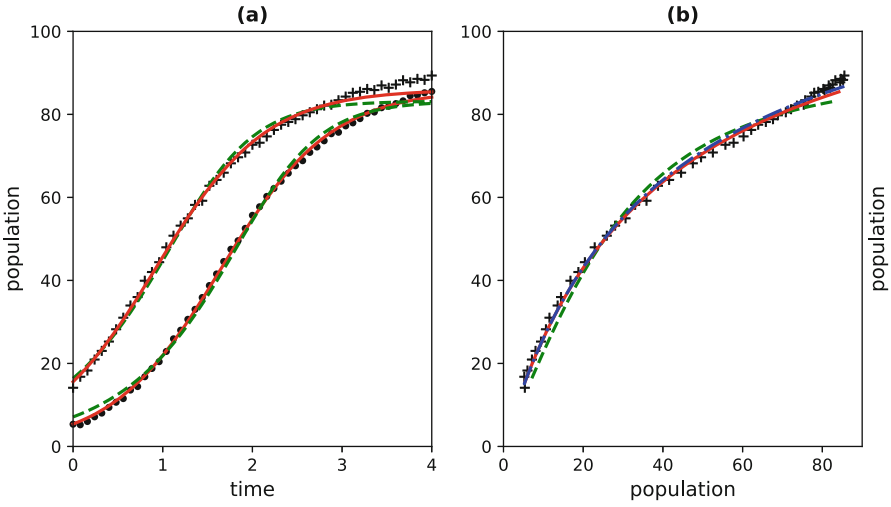


Fig. 1 (**a**) Noisily sampled measurements for a simulated system governed by a generalized logistic equation starting from $x_0 = 5$ (black dots) and $x_0 = 15$ (crosses). The best fit Verhulst model is depicted with a solid red line and the best fit $\beta = 2$ model with a dashed green line. (**b**) The empirical symmetry function computed from the two trajectories in (**a**) is shown with black crosses. The theoretical dynamical symmetries implied by the Verhulst and $\beta = 2$ models are shown with solid red and dashed green lines, respectively

for which $\alpha = 1$, $\beta = 2$, $\gamma = 1$. I'll call the latter the $\beta = 2$ model. Of course, neither of those reflects the true dynamics, but the scientist never gets to know this a priori (that would make inductive inference rather trivial). The point here is to examine what can be learned by different inferential methods in a realistic, relatively simple case where we happen to know the ground truth and can thus assess the performance of each method.

While there are myriad ways to fit and validate models of either sort, we'll follow a particularly simple procedure that exhibits the core features of most common statistical methods. In particular, we'll work with parameterized analytic solutions to the above differential equations. Specifically, Eq. (6) has solutions of the form,

$$x(t) = \frac{K}{1 + \left(\frac{K}{x_0} - 1\right) e^{-rt}}, \tag{8}$$

while the $\beta = 2$ model has solutions of the form,

$$x(t) = \frac{K}{\left(1 + \left((\frac{K}{x_0})^2 - 1\right) e^{-2rt}\right)^{1/2}}. \tag{9}$$

We'll use one sample from our target system to fit parameters for each model. That is, we'll use one set of measurements to determine r and K using nonlinear least squares regression. We'll then use those fit parameter values to try and predict the data in the second set of measurements.[12] The sum of the squared errors (SSE) for the predictions made by each model can be used as a simple measure of goodness fit.

The results of carrying out this procedure are shown graphically in Fig. 1a. The best fit of the Verhulst model (fit to the data of the lower curve) is depicted with a solid red line, and the best fit of the $\beta = 2$ model is shown with a dashed green line. Visually, it's clear that both models can be used to fit the initial curve very well. The SSE for the fit Verhulst logistic is 22.1, and 128 for the alternative model. The Verhulst has an advantage, but both do a decent job of at least summarizing the data. However, when we use the parameters from the first fit to predict the second data set, the $\beta = 2$ model clearly falls apart. The sum of squared errors are 95.9 and 310 for the Verhulst and $\beta = 2$ models, respectively. Note that the logistic is not merely better than the $\beta = 2$ model, but it does a compelling job of predicting the data. On the basis of this information—exactly the sort of information standard methods provide the scientist trying to infer a model of growth—you might reasonably be

[12] Note that the initial value of the population, x_0 is fit independently in each case. That's because, while the other parameters are presumed to be intrinsic features of the growing population, the initial population size is variable and assumed to have different (unknown) values in each case.

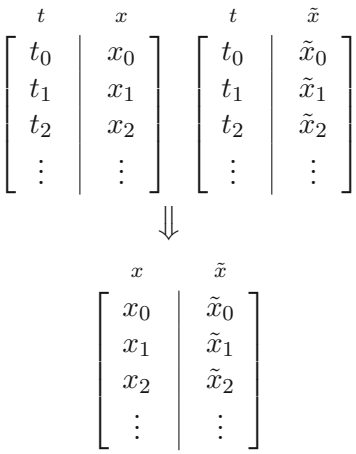


Fig. 2 Schematic showing how samples from two time-series are restructured to obtain an implicit model of the dynamical symmetry that maps one trajectory into the other (adapted from Figure 1 in [14])

inclined to conclude the Verhulst model is not just the best of the available options, but also a fairly reliable representation of the structure of the growth dynamics.[13]

But this would be a mistake. We know that the Verhulst model is wrong in this case, and that it will systematically lead us astray for growing populations not yet observed. Here is where attending to dynamical symmetries can help. To extract information about one of the dynamical symmetries (with respect to time) of a system from two trajectories of that system, one can simply build a new curve by matching each value of the variable of interest (x, or population size, in this case) in one trajectory with its contemporaneous value in the other. This operation is shown schematically in Fig. 2. The resulting empirical curve ($\tilde{x} = \sigma(x)$) is shown by the black crosses in Fig. 1b.

To use this information about the dynamical symmetries of our unknown growth system, we need to compare the symmetry function predicted by each of the models we have already fit to the data. These predicted symmetries, computed numerically for the models in precisely the same way as for the experimental data, are shown as solid red and dashed green lines in Fig. 1b. The comparison algorithm discussed above in Sect. 4 judges both theoretical symmetries to be significantly different from the empirical symmetry, and thus rejects the hypothesis that either of them accurately describes the structure of the target system. In other words, neither the Verhulst model (using the best-fit parameter values), nor the $\beta = 2$ model (again, using the best-fit parameters) accurately represent the target system. We have learned that they are both wrong.

But an even stronger result can be established. In general, one can consider the entire space of dynamical symmetries implied by a model, and ask whether there exist *any* parameter values that could account for the observed symmetry function,

[13] This is the line of reasoning presented in [29], where the Gompertz model is favored.

regardless of how well the associated solutions describe individual trajectories. In this case, it is possible to solve analytically for the set of all dynamical symmetries for each of the two classes.[14] For the Verhulst logistic model, the dynamical symmetries are given by

$$\sigma_p(x) = Kx/\left((1-e^{-p})x + e^{-p}K\right), \tag{10}$$

where each real value of p corresponds to a distinct symmetry transformation. Using this analytic form, it's possible to search for a set of parameter values (including p) that best fit the empirical symmetry directly. The optimal fit can then be compared, via the comparison algorithm described above, with the empirical symmetry. Doing so in this case leads to a *rejection*. In other words, we can with confidence reject the claim that *any* parameterization of the Verhulst logistic model accurately represents the dynamics of the target system.

5.3 Example: Real Populations

The procedure for checking the symmetries of a theoretical model against the empirical, measured symmetries of a dynamical system was demonstrated in the previous section for artificial data—data for which we know the ground truth about the governing dynamics. I crafted the artificial data to be as faithful to the messiness of real-world data as possible, but there is always a concern that a method will break down when confronted with real data. So let's take a look at an example of actual biological (or microbiological) growth.[15] Figure 3 shows two segments of data from a growth experiment. The experiment was designed to answer a question about the fitness of bacterial strains in a variety of environments. As such, it involved many populations of three bacterial strains, each tested in three distinct environments. But for our purposes, I have selected time-series measurements indicating the size of just one of these populations of bacteria growing on a microtiter plate.

In the interests of full disclosure, this particular population was not selected at random from the available datasets. Rather, I focused on this particular population because it was the one with a growth curve most plausibly described by one of the models considered above. In other words, it was chosen to maximize the difficulty of rejecting a logistic or $\beta = 2$ model. Other curves were clearly poorly fit by such models, and one would not have been inclined to try. It's also important to note that though two curves are shown in Fig. 3a, there was really only a single measured

[14] It's generally possible to determine and fit symmetries numerically, without an analytic, closed form solution. But since one is available in this case, I use it to simplify the analysis.

[15] This data was obtained from Connelly [6] and is used here with permission (and gratitude). The dataset can be found at https://zenodo.org/record/1171129. I am specifically considering the sixteenth row of the table.

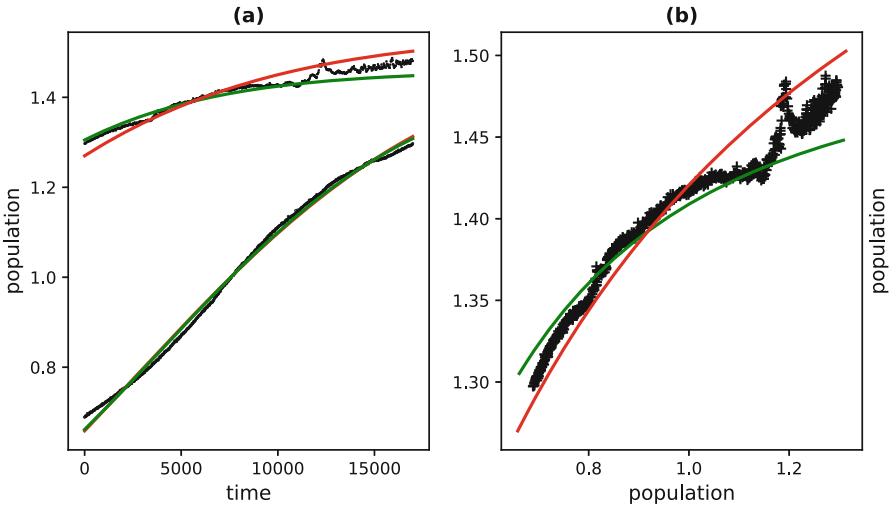


Fig. 3 (**a**) Measurements of population size for a real bacterial colony growing on a microtiter plate. The growth trajectory was divided in two to indicate how the population changes starting from two different initial conditions. The best fits Verhulst model is depicted with a solid red line and the best fit $\beta = 2$ model with a dashed green line for each measured curve. (**b**) The empirical symmetry function computed from the two trajectories in (**a**) is shown with black crosses. The theoretical dynamical symmetries implied by the Verhulst and $\beta = 2$ models are shown with solid red and dashed green lines, respectively

time series. What I've done is to split the time series in half, and translate the time-values of the second half so that it begins at $t = 0$. The validity of such a procedure rests on the assumption that the dynamics is autonomous. When such an assumption is warranted, it means that a dynamical symmetry can be directly estimated from purely observational data, without any interventions.

With the pair of sampled time-series curves, we can proceed as before and use them to estimate a symmetry of the growing population. Figure 3a shows the best-fit models as solid red and green lines for the Verhulst and $\beta = 2$ models, respectively. The fit models are nearly indistinguishable in terms of their SSE values, and it's clear from visual inspection that neither provides an exact fit. In fact, given the obvious curvature in the residuals, both models would likely be rejected by methods that focus on single trajectory analysis. Nonetheless, the $\beta = 2$ model provides a better relative fit, and might seem a reasonable approximation to the data. However, the failure to represent the target system is quite pronounced when we examine how well the symmetries of the theoretical models fit the symmetry estimated from the data. The latter is shown in Fig. 3b, along with the theoretical symmetries implied by the best-fit logistic and $\beta = 2$ models. The decision procedure we've been considering strongly rejects the hypothesis that either theoretical symmetry is equivalent to that in the data. In other words, neither model accurately represents the causal structure of this growing population.

5.4 Example: Stochastic Logistic Models

As discussed above, stochastic models present special problems for validation. To demonstrate the efficacy of the symmetry comparison approach, I simulated a stochastic version of a generalized logistic equation. Specifically, I built a (simulated) target system governed by the following stochastic differential equation (SDE),

$$\frac{dx}{dt} = rx\left(1 - \left(\frac{x}{K}\right)^2\right) + sxdW_t, \tag{11}$$

where W_t is a one-dimensional Wiener process, and s is a constant determining the amount of multiplicative noise. Solutions to this equation were generated numerically using the discrete equations derived in [16] (based on the Milstein method mentioned in [12]).

Figure 4a shows data from times series measured for multiple replicates of the target system. There are ten replicates of the system for initial condition $x_0 = 5$, and ten replicates for the initial condition $\tilde{x}_0 = 15$. That is, the target system was

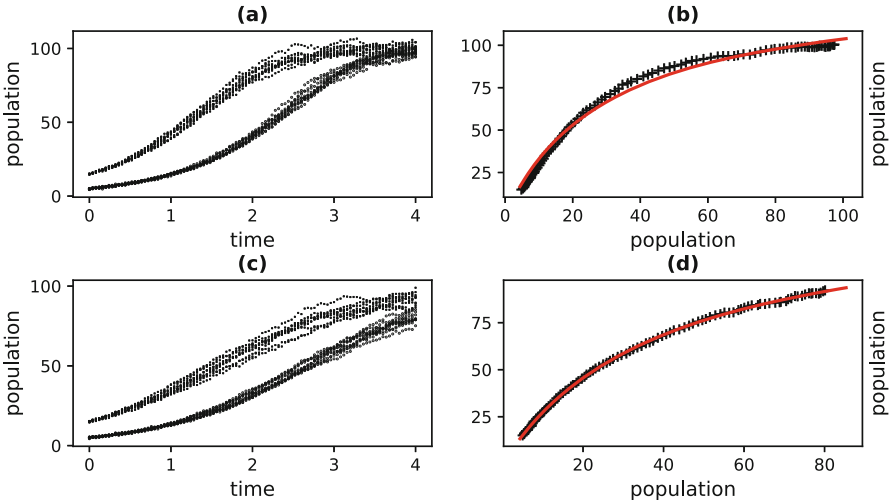


Fig. 4 (**a**) Measurements for a simulated system governed by a stochastic generalized logistic equation for which $\alpha = 1$, $\beta = 2$, and $\gamma = 1$ (black crosses). The lower set of curves starts from $x_0 = 5$, and the upper from $x_0 = 15$. (**b**) The empirical symmetry function computed from the means of each of the two trajectories in (**a**) is shown with black crosses. The theoretical dynamical symmetry implied by the Verhulst model is shown with a solid red line. (**c**) Measurements for a simulated system governed by a stochastic Verhulst logistic equation (black crosses). (**d**) The empirical symmetry function computed from the means of each of the two trajectories in (**c**) is shown with black crosses. The theoretical dynamical symmetry implied by the Verhulst model is shown with a solid red line

evolved through time for ten iterations starting from each of two initial population sizes. Data from all iterations and initial conditions are plotted together.

There are a variety of ways in which to fit a stochastic model to such data (assuming we don't know the ground truth of Eq. (11)). One might build a numerical simulation and then attempt to optimize parameters with respect to one or another of the measures of fit like those discussed above in Sect. 2.2. The generalized definition of dynamical symmetry (see Definition 3 above), however, suggests that we focus on the expected value of the population at a given time (for a given initial condition). Consider the two sets of replicates corresponding to the two distinct initial conditions. If one averages the values measured at a given time for each set, one obtains two curves indicating expected population as a function of time. The function mapping one of these curves into the other is entailed by (though not a full specification of) a dynamical symmetry. The result of this procedure is shown by the black crosses in Fig. 4b. For whatever our model is, we can similarly compute a connection between expected value curves (as implied by a dynamical symmetry) and then compare the two as before. A significant difference would allow us to reject the structural validity of our model. In this case, I have chosen to try to model the system with a stochastic Verhulst equation:

$$\frac{dx}{dt} = rx\left(1 - \left(\frac{x}{K}\right)\right) + sxdW_t, \tag{12}$$

For such a model, it is possible to find an analytic expression for the expected value of x as a function of time. This is given by:

$$E[x(t)] = \frac{k}{1 + ((k/x_0) - 1)e^{(-rt)}}, \tag{13}$$

where x_0 is the value of x at $t = 0$ [20, Sec 4.4]. This, as it happens, is exactly the form of the solutions of the deterministic Verhulst equation. Functions connecting expected value curves of the stochastic model are thus identical to the dynamical symmetries of the deterministic model (see Eq. (10)). The red solid lines in Fig. 4b show the result of performing a least squares fit for symmetries of the Verhulst equation on the empirical data obtained from the target system. When the comparison algorithm is applied to this pair of curves—the one estimated for the target system and the best fit solution for our model—it rejects the model. In other words, even in the stochastic case, comparison of dynamical symmetries can lead to definitive rejection of a model.

Of course, this would be useless if the method also rejects the true model. In Fig. 4c, d, results are shown for the same procedure carried out when the underlying system really is governed by a stochastic Verhulst equation. In this case, the comparison algorithm tentatively declares that the model and the target system share the same symmetry (and thus the model may be structurally valid).

6 Conclusion

I have argued above for the need to recognize a sort of structural validity for models that is broader and more forgiving than exact structural isomorphism of model and target (whatever that might mean), but that is nonetheless sufficient for establishing the reliability of a model's predictions regarding outcomes under a manipulation of inputs or boundary conditions. In other words, the sort of model reliability that is needed for the confident prediction and control of a target system is looser than the strict notion of structural validity that can be found in much of the scientific and engineering literature.

Furthermore, I've shown by way of a series of concrete examples how dynamical symmetries can be used to test for this broader sense of structural validity. Roughly, one compares the theoretical dynamical symmetries entailed by the model with estimates of real dynamical symmetries exhibited by the target system. This method, with a suitably generalized definition of dynamical symmetry, even applies to the case in which both the model and target system are stochastic. The method has important limitations. For one, it assumes that there are no latent variables driving the dynamics. Nonetheless, it represents a rigorous new tool in a field that is increasingly in need of new tools as computational models grow ever more complex.

While these results are, I think, important in their own right, I wish to draw out some implications of the mode of origination and practical success of these methods. In a sense, the development of this method of model validation is an exercise in applied philosophy. The notion of a dynamical symmetry derives from philosophical work on natural kinds (i.e., projectible kinds) [13]. A specific tool for model validation was derived from a very general answer to an epistemic puzzle fundamental to scientific inquiry: how do we recognize clusterings of systems or phenomena that are good candidates for instantiating scientific laws? This is obviously not the first philosophical project to contribute to scientific practice. But it is a reminder that philosophers can be helpful partners in developing well-motivated tools for empirical inquiry, not just sideline commentators.

But the converse is also true; the development of the philosophical idea into an operative tool has provided a variety of lessons that philosophers should heed. For example, the focus on approaches to narrow-sense structural validity to the exclusion of methods that would more effectively satisfy the aims of a modeler is largely a product of philosophical myopia, not a quirk of those working with models. Prominent authors in the literature on model validation frequently and explicitly take their cue from the philosophers. For instance, in speaking of methods for establishing structural validity, [3, p. 186] frames the project of structural validation this way:

> Validation of a system dynamics model is much more complicated than that of a black-box model, because judging the validity of the internal structure of a model is very problematic, both philosophically and technically. It is philosophically difficult, because, as we shall briefly review in the next section, the problem is directly related to the unresolved philosophical issue of verifying the truth of a (scientific) statement.

Thus, it is a philosophical lesson of the applied work presented here that there exists an important feature of models (and scientific theories) that sits between mere predictive success and perfect representational fidelity. Specifically, models can more or less reliably make judgments about what would be the case under intervention or manipulation without using a mechanism exactly isomorphic to whatever drives the real-world target system. This is an aspect of modeling that philosophers often ignore but shouldn't given its demonstrable utility in making sure models do what we need them to do. In other words, philosophers interested in foundational epistemic problems would do well to listen closely to their colleagues in applied math, data analysis, and statistics.

Acknowledgements I am grateful to the participants in the 2015 Algorithms and Complexity in Mathematics, Epistemology and Science (ACMES) conference for insightful discussion of an early algorithm for discovering dynamical kinds, to Cosmo Grant for pointing out a physical inconsistency in the first version of one of my examples, and to Nicolas Fillion for helpful comments on a previous draft of this paper. The work presented here was supported by the National Science Foundation under Grant No. 1454190.

References

1. Balci O (1994) Validation, verification, and testing techniques throughout the life cycle of a simulation study. Ann Oper Res 53(1):121–173
2. Barlas Y (1989) Multiple tests for validation of system dynamics type of simulation models. Eur J Oper Res 42(1):59–87
3. Barlas Y (1996) Formal aspects of model validity and validation in system dynamics. Syst Dyn Rev 12(3):183–210
4. Buchanan RL, Whiting RC, Damert WC (1997) When is simple good enough: a comparison of the Gompertz, Baranyi, and three-phase linear models for fitting bacterial growth curves. Food Microbiol 14(4):313–326.
5. Committee on Mathematical Foundations of Verification, Validation, and Uncertainty Quantification (2012) Assessing the reliability of complex models: mathematical and statistical foundations of verification, validation, and uncertainty quantification. National Academy Press, Washington
6. Connelly B (2014) Data set for 'analyzing microbial growth with R'. https://doi.org/10.5281/zenodo.1171129
7. Fillion N (2017) The vindication of computer simulations. In: Lenhard J, Carrier M (eds) Mathematics as a tool: tracing new roles of mathematics in the sciences. Springer, Cham, pp 137–155
8. Flach P (2012) Machine learning: the art and science of algorithms that make sense of data. Cambridge University Press, Cambridge
9. Fujikawa H, Kai A, Morozumi S (2004) A new logistic model for Escherichia coli growth at constant and dynamic temperatures. Food Microbiol 21(5):501–509
10. Gompertz B (1825) XXIV. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. In a letter to Francis Baily, Esq. F. R. S. &c. Philos Trans R Soc Lond 115:513–583
11. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer series in statistics, 2nd edn. Springer, New York

12. Higham DJ (2001) An algorithmic introduction to numerical simulation of stochastic differential equations. SIAM Rev 43(3):525–546
13. Jantzen BC (2014) Projection, symmetry, and natural kinds. Synthese 192(11):3617–3646
14. Jantzen BC (2017) Dynamical kinds and their discovery. In: Proceedings of the UAI 2016 workshop on causation: foundation to application. ArXiv: 1612.04933
15. Ling Y, Mahadevan S (2013) Quantitative model validation techniques: new insights. Reliab Eng Syst Saf 111:217–231
16. Liu M, Fan M (2017) Permanence of stochastic Lotka–Volterra systems. J Nonlinear Sci 27(2):425–452
17. McCarthy MA, Broome LS (2000) A method for validating stochastic models of population viability: a case study of the mountain pygmy-possum (Burramys parvus). J Anim Ecol 69(4):599–607
18. Miller JH (1998) Active nonlinear tests (ANTs) of complex simulation models. Manag Sci 44(6):820–830
19. Rhinehart RR (2016) Nonlinear regression modeling for engineering applications: modeling, model validation, and enabling design of experiments. Wiley, Hoboken
20. Skiadas CH (2010) Exact solutions of stochastic differential equations: Gompertz, generalized logistic and revised exponential. Methodol Comput Appl Probab 12(2):261–270
21. Sokal RR, Rohlf FJ (1994) Biometry: the principles and practices of statistics in biological research, 3rd edn. W. H. Freeman, New York
22. Spirtes P, Glymour CN, Scheines R (2000) Causation, prediction, and search. Adaptive computation and machine learning, 2nd edn. MIT Press, Cambridge
23. Tsoularis A, Wallace J (2002) Analysis of logistic growth models. Math Biosci 179(1):21–55
24. Verhulst PF (1838) Notice sur la loi que la populations suit dans son accroissement. Correspondence Mathématique et Physique. X:113–121
25. Vogels M et al (1975) P. F. Verhulst's 'Notice sur la loi que la populations suit dans son accroissement' from correspondence mathematique et physique. Ghent, vol. X, 1838. J Biol Phys 3(4):183–192
26. Wilcox JR (2018) Research for practice: highlights in systems verification. Commun ACM 61(2):48–49
27. Woodward J (2001) Law and explanation in biology: invariance is the kind of stability that matters. Philos Sci 68(1):1–20
28. Zeigler BP, Praehofer H, Kim TG (2000) Theory of modeling and simulation, 2nd edn. Academic, San Diego
29. Zwietering MH et al (1990) Modeling of the bacterial growth curve. Appl Environ Microbiol 56(6):1875–1881

Modeling the Biases in Last Digit Distributions of Consecutive Primes

Daniel Lichtblau

Abstract Recent work by Lemke Oliver and Soundararajan, as well as earlier results by Ko, have brought to light an unexpected asymmetry in the distribution of last digits of consecutive primes. For example, in the first 10^8 pairs of consecutive primes, around 4.6 million end with {1,1} respectively, whereas more than 7.4 million end with {1,3}. This disparity is not explained by the fact that opportunities for the next prime come sooner for n+2 than for n+1. This leaves open the question: what accounts for this sizable bias? We provide justification based on a mix of elementary theory and computation. The model we develop moreover accurately predicts crossovers in relative frequencies of certain pairs-of-pairs.

1 Introduction

It was observed in [1, 4] that last digit pairs of consecutive primes do not appear with the same (approximate) frequencies, and, in particular, equal last digits seem to arise far less frequently than other possibilities. This also holds for last digit pair frequencies in bases other than 10. This behavior is in contrast to the (perhaps naive) assumption that such last digit pairs would be, to reasonable approximation, independently uniformly distributed. Indeed, when a detailed study of this finding was more recently announced in a preprint version of [5], it caught the attention of the scientific popular press, with write-ups in several less technical journals: notable amongst these is [3].

The analysis in [5] is based on the Hardy-Littlewood prime k tuples conjecture [2]. A partial analysis was also presented in [1], using the Prime Number Theorem (PNT), elementary number theory and various heuristics by Polya and the authors to approximate these consecutive prime last digit pair frequencies modulo a given base. The present work will examine a few specific cases using computational methods.

D. Lichtblau (✉)
Wolfram Research, Champaign, IL, USA
e-mail: danl@wolfram.com

© Springer Science+Business Media, LLC, part of Springer Nature 2019

N. Fillion et al. (eds.), *Algorithms and Complexity in Mathematics, Epistemology, and Science*, Fields Institute Communications 82,
https://doi.org/10.1007/978-1-4939-9051-1_7

The approach shares some features of [1] in that the basic tools also include the PNT and elementary number theory. Different heuristics are brought into play, primarily based on straightforward sieving. We also set up a recurrence to better approximate the behavior for some particular bases. We solve the recurrences and show the very close agreement with computed results.

I thank Robert Lemke Oliver for email discussion of some aspects of his joint work on this problem. This volume is dedicated to the memory of ACMES 2016 plenary speakers Jon Borwein and Ann Johnson. In addition to the many prior accomplishments in their lives and work, they helped to make this conference the excellent event that it was. I am grateful to the editors for providing the opportunity to submit work herein. I thank the two anonymous reviewers for their very detailed remarks, which were helpful for streamlining and in other respects improving the exposition.

We first check last digits of consecutive primes in base 3. Ignoring the first few, the last digits of pairs of consecutive primes are in the set (1,1), (1,2), (2,1), (2,2). We begin with a simple observation. We take the second ten million consecutive prime last digit (base 3) pairs. There are 2,222,836 such pairs equal to $(1, 1)$, and 2,223,517 equal to $(2, 2)$. In contrast, there are 2,776,823 equal to $(1, 2)$ and also 2,776,823 equal to $(2, 1)$. So the pairs with equal last digits have very nearly the same frequency, and the counts for pairs with unequal last digits are actually identical. The ratio of the two counts is around 0.80, which is quite far from unity (and surprisingly so, if one expected these four classes to be approximately equally distributed).

For base 4 the last digit pairs are in the set (1,1), (1,3), (3,1), (3,3). Using the same ten million consecutive pairs as above, the count for each of the four possibilities is 2,242,490, 2,757,418, 2,757,418, 2242673 respectively.

In base 10 the last digits can be 1, 3, 7, or 9, giving 16 possible pairs. The set of frequencies for the second ten million consecutive prime pairs is shown in Table 1.

For shorthand we will use the probability symbol $\mathcal{P}$ to model expected frequencies of particular increments of p_1 being prime, since these frequencies behave in a way that is similar to actual probability frequencies. Correspondingly we (mis)use the term "probability" to mean the count of events in a given class divided by total number of possibilities (one might wish to think of this as "heuristic probabilities" or perhaps "probability surrogates"; it is a model intended to capture actual behavior to reasonable approximation). We denote by $\#(a, b, m, v, w)$ the number of consecutive pairs between the vth and wth primes with the first equal

Table 1 Base 10 pair frequencies

Last digit by (row, col)	1	3	7	9
1	456,358	747,558	757,467	538,552
3	598,659	437,164	707,808	756,429
7	638,122	677,528	436,339	747,968
9	806,795	637,811	598,343	457,098

to a and the second equal to b modulo m (that is to say, we consider all pairs of consecutive primes between prime(v) and prime(w)). Again, one might think of this count, divided by the number of consecutive pairs, as a surrogate for a probability distribution. A symmetry may be observed in the examples above: $\#(a, b, m, v, w)$ is very close to $\#(m - b, m - a, m, v, w)$ (in tabular form as above this is seen as a symmetry across the antidiagonal). This is discussed as proposition 4.2 in [1] (where a slightly different notation is used). The model developed in later sections will also have this symmetry.

The next few sections will work with specific bases, also specializing to the case where the first prime in each pair is equal to 1 modulo the base under consideration. In the next section we show a simple numeric last digits frequency approximation for base 3 that turns out to be fairly accurate. More importantly we develop some of the frequency estimation machinery to use in the following sections, where we develop more powerful analytic estimates using recurrence relations. The main power and novelty of this approach is that it provides closed forms for asymptotic pair frequency estimates. As an added bonus, some perhaps surprising behavior appears in the predicted asymptotics, including a case where crossovers in frequencies are such that the self-avoidance effect is not immediately apparent. That these are validated by actual frequency computations is further evidence of the usefulness of these analytic forms. A short section explains the symmetry across the antidiagonal. All computations are performed in the Wolfram Language running in version 11 of Mathematica [6]. Representative code, specifically for bases 3 and 5, is provided in an appendix.

2 Base 3

We have consecutive prime pairs (p_1, p_2) with the further stipulation that $p_1$is equal to 1 modulo 3 (henceforth written $p_1 \equiv_3 1$). Since these are odd primes we also obviously have $p_1 \equiv_2 1$. Clearly $p_1 + 2$ and $p_1 + 8$ are composite since they are divisible by 3. We now consider the residue classes for p_1 modulo 5. These are just 1, 2, 3, or 4, and from the PNT it can be shown that these four cases are to good approximation uniformly distributed (we will refer to this as Approximation 1, or A1 for short). We consider each in turn, in order to see when these increments of p_1 are forced to be divisible by 5 (and thus composite). The cases break down as follows.

1. If $p_1 \equiv_5 1$ then $p_1 + 4$ is divisible by 5, hence composite.
2. If $p_1 \equiv_5 2$ then $p_1 + 8$ is divisible by 5, hence composite (it is also divisible by 3, but the divisibility by 5 will play an important role in subsequent analysis).
3. If $p_1 \equiv_5 3$ then $p_1 + 2$ is divisible by 5, hence composite (it is also divisible by 3).
4. If $p_1 \equiv_5 4$ then $p_1 + 6$ is divisible by 5, hence composite.

From A1 we infer a distribution where $p_1 + 4$ and $p_1 + 6$ are forced to be composite in (roughly) $\frac{1}{4}$ of all cases, and both p_1+2 and p_1+8 are composite in all cases. In each class, comprising $\frac{1}{4}$ of all cases, we have that some p_1+k is divisible by 5, and we can add multiples of 10 and still have divisibility by 5. For example, each of $p_1 + \{12, 14, 16, 18\}$ are a priori composite in $\frac{1}{4}$ of all cases. Notable for its absence in this analysis is $p_1 + 10$. This is an important detail and we will return to it presently.

We know $p_2 \neq p_1 + 2$. We now wish to approximate $\mathcal{P}(p_2 = p_1 + 4)$. Since one in four cases force $p_1 + 4$ to be composite, it should be $\frac{3}{4}$ times the conditional probability that no prime larger than 5 divides it. By "conditional probability" we again have in mind a number-of-occurrences-over-total-count interpretation, where we also take into account the condition that it is not divisible by 2 or 3 and moreover is in the three-in-four cases where it is not divisible by 5. Since $\frac{4}{5}$ of the numbers in the range under consideration are not divisible by 2, 3, or 5, the frequency that p_1+4 is prime is $\frac{15}{4}$ times larger than that of a "random" integer in the size range of p_1. In our example we consider values around $n = \pi\left(10^7\right)$ and the PNT allows us to approximate this as $\frac{1}{\log n}$ (we show this factor for notational convenience below but use the more accurate averaged log integral for actual computations). Putting all this together gives (1).

$$\mathcal{P}(p_2 = p_1 + 4) = \frac{3}{4}\frac{15}{4}\frac{1}{\log n} \tag{1}$$

Obviously p_2 cannot be $p_1 + 6$ if it is $p_1 + 4$. So we must account for this factor in deriving $\mathcal{P}(p_2 = p_1 + 6)$. Again we have a three-in-four situation because $\frac{1}{4}$ of the mod 5 residue classes force it to be composite (2).

$$\mathcal{P}(p_2 = p_1 + 6) = (1 - \mathcal{P}(p_2 = p_1 + 4))\,\frac{3}{4}\frac{15}{4}\frac{1}{\log n} \tag{2}$$

A priori $\mathcal{P}(p_2 = p_1 + 8) = 0$. We now consider $\mathcal{P}(p_2 = p_1 + 10)$. As noted above, this case is different: we lose the factor of $\frac{3}{4}$ because there is no residue class modulo 5 forcing this to be composite in $\frac{1}{4}$ of the cases (3).

$$\mathcal{P}(p_2 = p_1 + 10) = (1 - \mathcal{P}(p_2 = p_1 + 4) - \mathcal{P}(p_2 = p_1 + 6))\,\frac{15}{4}\frac{1}{\log n} \tag{3}$$

The form for general increment $2k$ is obtained by using a factor to insure that a prior increment was not the next consecutive prime (so we subtract the sum of those prior probabilities from 1), as well as a factor $m(k)$ which we now define. It is based on equivalence classes modulo 3 and 5. If $k \equiv_3 1$ then $p_1 + 2k$ is of necessity composite, so $m(k) = 0$. If $k \equiv_5 0$ and $k \not\equiv_3 2$ then $m(k) = 1$. In all other cases we define $m(k) = \frac{3}{4}$. Really m should also be regarded as a function of the base, residue class of p_1, and moduli under consideration, that is, we have shown what

might be denoted $m_{(3,1,\{5\})}(k)$. We will usually omit the subscript since its values will be clear from context. With this notation, the general term is (4).

$$\mathcal{P}\left(p_2 = p_1 + 2k\right) = \left(1 - \sum_{j=1}^{k-1} \mathcal{P}\left(p_2 = p_1 + 2j\right)\right) m(k) \frac{15}{4} \frac{1}{\log n} \tag{4}$$

For computational purposes we want to consider sufficiently many increments to account for nearly all possible values for p_2. If we allow for $p_2 = p_1 + \{2, 4, \ldots, 90\}$ it is not hard to show that we hit close to 99.4% of the possible next primes when p_1 ranges between the ten and twenty millionth primes.

The next step is to compare frequencies of $p_2 \equiv_3 1$ to those of $p_2 \equiv_3 2$. The first case aggregates estimated frequencies for $p_2 = p_1 + 6, p_1 + 12, \ldots$ while the second aggregates those for $p_2 = p_1 + 4, p_1 + 10, \ldots$. A straightforward computation shows that the ratio of these is 0.802, so we have come close to the observed frequency ratio. This is with a simple model wherein we only consider residue classes modulo 5. We will next extend the model to use more prime residue classes.

We begin by recalling some sieving frequency formulas. The frequency of numbers that are composite, with smallest prime factor being 2, is obviously $\frac{1}{2}$ (here we use "frequency" in the standard limiting sense). Similarly the frequency of composites with smallest factor 3 is $\frac{1}{3}$ of those not sieved out by 2, or $\frac{1}{3}\frac{1}{2} = \frac{1}{6}$. Those sieved by 5 comprise $\frac{1}{5}$ of the $\frac{1}{2}\frac{2}{3}$ remaining composites. For the k^{th} prime p_k, denoting by $s(k)$ the frequency of composites with smallest prime p_k, it is $\frac{1}{p_k}$ times the relative number not already removed, which gives (5) below.

$$s(k) = \prod_{j=1}^{k-2} \left(p_j - 1\right) / \prod_{j=1}^{k-1} p_j \tag{5}$$

The frequency of composites that remain after sieving out the first k primes is thus $1 - \sum_{j=1}^{k-1} s(j)$. For $k = 6$ this is $\frac{192}{1001}$ or approximately 0.192. So one correction factor we use in this case is the reciprocal of this value. We remark that this approximation is only reasonable when the product of primes thus utilized is at least modestly small compared to the range under consideration. As the range for our main example is around 10^8, we restrict to the first six primes to be well within the desired inequality.

In order to work with the modulo p equivalence classes for $p \in \{5, 7, 11, 13\}$, we will create a factor $m_{(3,1,\{5,7,11,13\})}(k)$. For a given such prime p, there are $p - 1$ nonzero equivalence classes. We already showed in detail how to handle $p = 5$. Recall that increments that are even multiples of 5, such as $p_1 + 10$, are a priori neither even nor divisible by 5. We will call an increment "special mod 5" if there is no residue class for p_1 forcing that increment to be either even or divisible by 5. It is straightforward to see these increments are simply the even multiples of 5. Since we will work with more primes, we will need to employ correction factors for non-special increments of each such prime.

For $p = 7$ we only show two cases, as the basic idea should be clear by now. (1) If $p_1 \equiv_7 1$ then $p_1 + 6$ is divisible by 7. (2) If $p_1 \equiv_7 2$ then $p_1 + 12$ is divisible by 7.

The important points are as follows. (1) There are 6 nonzero equivalence classes modulo 7. (2) If p_1 is in a given class, that forces certain subsequent values to be composite. (3) With frequency $\frac{5}{6}$, p_1 is not in a given such class. For an increment k that is non-special for 7, the frequency of $p_1 + k$ being prime thus gets adjusted by a factor of $\frac{5}{6}$.

For an arbitrary prime p this adjustment factor is of course $\frac{p-2}{p-1}$. That is to say, given an increment k that is non-special for p, the frequency of $p_1 + k$ being prime must be adjusted by this factor $\frac{p-2}{p-1}$ since with frequency $\frac{1}{p-1}$ it is forced to be composite. With these considerations we see that $m(k)$ is zero for $k \in \{1, 4, 7, \ldots\}$ (since $p_1 + 2k$ is composite for these), and it is $\prod_{p \nmid k} \frac{p-2}{p-1}$ otherwise, adjusting for those primes for which k is not special. This approximation uses an extension of A1 which we call Approximation 2 (A2): The frequencies of primes in a given set of equivalence classes modulo a given set of primes is approximately the product of the frequencies of it being in each separate equivalence class.

A similar use of these factors is shown in [1] and, as noted therein, goes back to Hardy and Littlewood [2]. We use these adjustment factors to again approximate the relative frequencies of consecutive prime pair last digits mod 3, still under the assumption that the first is equal to 1 mod 3 and the range of consideration is the second ten million primes. The ratio (see appendix code) is 0.789, which is slightly worse than the (more naive) first estimate, but still quite close to the mark.

3 Base 4

Simple computations of the sort done for last digit pairs mod 3 give estimates that appear to be fairly close to actual frequencies. For base 4 we develop a more complete model, deriving a closed form to approximate the expected frequencies.

The key parts are as follows. We assume the range of interest is the interval $(n, 2n)$ for some n. There is thus the usual factor of $\frac{1}{\log n}$. For a given first prime in a consecutive pair, p_1, we consider the interval from there to $p_1 + k \log n$ for modest size k, in order to have a reasonable probability of reaching p_2 (in actual computations we can sum to infinity; we use the fact that "most" cases have successors in such a range, and so the factor of $\frac{1}{\log n}$ is justified). As before, we consider the effect of different prime equivalence classes of p_1, excluding 2 (since it plays no role for this case) and restricting to primes no larger than $\log n$ in order for the sieve probability estimate to be valid. Thus we will have a factor of $\prod_{2<\text{prime}(j)<\log n} \frac{\text{prime}(j)-2}{\text{prime}(j)-1}$. Again we take into account the fraction of values not sieved by these first primes, which we denoted $1 - \sum_{j=1}^{\pi(\log n)} s(j)$, with $s(j)$ as defined in (5). For frequency estimates the fraction of interest, $F(k)$, is the reciprocal of this value:

$$F(k) = \left(1 - \sum_{j=1}^{\pi(\log n)} s(j)\right)^{-1}$$

With this notation, a recurrence to approximate $\mathcal{P}$ ($p_2 = p_1 + 2k$) can be written as (6).

$$\mathcal{P}(k+1) = \left(1 - \sum_{j=1}^{k} \mathcal{P}(j)\right) \frac{F(\log n)}{\log n} \prod_{2<\text{prime}(j)<\log n} \frac{\text{prime}(j) - 2}{\text{prime}(j) - 1} \prod_{p|k+1,\, p \text{ prime}, p<\log n} \frac{p-1}{p-2} \tag{6}$$

This is not quite tractable for the purpose of finding a closed form. The primary simplification will be to restrict the last factor to primes 3 and 5. This brings about a related simplification in that we now use $F(3)$ (although, were it needed, $F(\log n)$ could be estimated). We will separately handle the eight explicit equivalence classes where p_1 is nonzero both modulo 3 and modulo 5, and then average the resulting frequency estimates. This also means we drop the next to last product, since it is used only when lumping together equivalence classes and we are now treating them separately.

As written, each term depends on all prior terms. Since they all appear in a way that is unweighted (that is, coefficients are equal), we can conveniently remove most of the dependency. We simply define a new function $g(k)$ in (7) as the sum of the $\mathcal{P}(j \leq k)$.

$$g(k) = \sum_{j=1}^{k} \mathcal{P}(k) \tag{7}$$

We show the resulting approximation to the recurrence in some detail for the case where $p_1 \equiv_3 1$ and $p_1 \equiv_5 1$. First note that certain increments of p_1 are forced to have a frequency of zero, e.g. $p_1 + \{2, 4, 8\}$ (the first and third are of necessity divisible by 3, while the middle one is divisible by 5). Were we to ignore such slots, our recurrence would be as in (8).

$$g(k+1) - g(k) = (1 - g(k))\frac{F(3)}{\log n} \tag{8}$$

Since $F(3)$ is $\frac{15}{4}$, we have a factor $\frac{15}{4 \log n}$ that is independent of k. For notational convenience we will replace it with a new constant λ. The solution to the simplified recurrence in (9) is $1 - (1 - \lambda)^k$. To use this for the actual recurrence of interest, we adjust so that increments with frequency zero amount to positions of repeated values in the recurrence solution (since $\mathcal{P}(k) = g(k) - g(k-1)$). To construct

a general solution to the actual recurrence, one uses the solution to (9), placing repeated values where required, and incrementing only in slots corresponding to nonzero frequencies.

Once increments corresponding to frequencies forced to be zero due to divisibility by 3 meet those forced by divisibility by 5, the frequencies hit a pattern of length 15 with zeros in specific slots, having the form (*,*,0,*,0,0,*,*,0,0,*,0,*,*,0).

A consequence of the eventual repeating (which of course applies to all eight cases) is that we can benefit from a split into two steps. The first handles the "initial" segments, and the second deals with the repeated runs thereafter. Also we want both parts to end with a position corresponding to an increment of frequency zero, in order that taking neighboring differences to recover $\mathcal{P}(k)$ for even parity k does not involve crossing boundaries of segments. Thus if the initial part were to have odd length we just add the next 15 positions. Similarly we aggregate the repeating pattern part in chunks of 30.

The initial sequence for the case $p_1 \equiv_3 1$ and $p_1 \equiv_5 1$ has $g(k)$ as in (9).

$$\begin{aligned} &0, 0, \lambda, \lambda, 1-(1-\lambda)^2, 1-(1-\lambda)^3, 1-(1-\lambda)^3, 1-(1-\lambda)^4,\\ &1-(1-\lambda)^5, 1-(1-\lambda)^5, 1-(1-\lambda)^6, 1-(1-\lambda)^6, 1-(1-\lambda)^6,\\ &1-(1-\lambda)^7, 1-(1-\lambda)^8, 1-(1-\lambda)^8, 1-(1-\lambda)^8, 1-(1-\lambda)^9,\\ &1-(1-\lambda)^9, 1-(1-\lambda)^{10}, 1-(1-\lambda)^{11}, 1-(1-\lambda)^{11} \end{aligned} \tag{9}$$

The first repeating chunk is as below. Subsequent ones look the same except the exponents increase by 16; this is because each repeating segment corresponds to 30 frequencies, of which 16 are nonzero (this follows from the fact that we get nonzero values in positions that correspond to the 16 odd values relatively prime to 60).

$$\begin{aligned} &1-(1-\lambda)^{12}, 1-(1-\lambda)^{13}, 1-(1-\lambda)^{13}, 1-(1-\lambda)^{14}, 1-(1-\lambda)^{14},\\ &1-(1-\lambda)^{14}, 1-(1-\lambda)^{15}, 1-(1-\lambda)^{16}, 1-(1-\lambda)^{16}, 1-(1-\lambda)^{16},\\ &1-(1-\lambda)^{17}, 1-(1-\lambda)^{17}, 1-(1-\lambda)^{18}, 1-(1-\lambda)^{19}, 1-(1-\lambda)^{19},\\ &1-(1-\lambda)^{20}, 1-(1-\lambda)^{21}, 1-(1-\lambda)^{21}, 1-(1-\lambda)^{22}, 1-(1-\lambda)^{22},\\ &1-(1-\lambda)^{22}, 1-(1-\lambda)^{23}, 1-(1-\lambda)^{24}, 1-(1-\lambda)^{24}, 1-(1-\lambda)^{24},\\ &1-(1-\lambda)^{25}, 1-(1-\lambda)^{25}, 1-(1-\lambda)^{26}, 1-(1-\lambda)^{27}, 1-(1-\lambda)^{27} \end{aligned}$$

The actual usage for these will be to separate frequencies for even increments (corresponding to consecutive prime pairs with the same last digit mod 4) from frequencies for odd increments (for pairs with opposite last digits mod 4). Since $\mathcal{P}(k) = g(k) - g(k-1)$ we sum the even parity frequencies as $\sum_k (-1)^k g(k)$. We do this for initial segments and successive chunks of length 30. That latter has the form below, where $n = 11 + 16m$ for the mth chunk (the 11 accounts for the terms from the initial segment).

$$-(1-\lambda)^n(-2+\lambda)(-1+\lambda)\lambda\left(2-2\lambda+\lambda^2\right)\left(2-4\lambda+6\lambda^2-4\lambda^3+\lambda^4\right)$$
$$\left(1-3\lambda+11\lambda^2-17\lambda^3+14\lambda^4-6\lambda^5+\lambda^6\right)$$

This is a tractable formula for the purpose at hand. We now add the initial contribution to the sum over m repeating chunks (we abbreviate for conciseness), obtaining (10).

$$\begin{aligned}(1-\lambda)^2-(1-\lambda)^4+(1-\lambda)^6-(1-\lambda)^7+(1-\lambda)^8-(1-\lambda)^{10}+\\ ((2-\lambda)(2-2\lambda+\lambda^2)(2-4\lambda+6\lambda^2-4\lambda^3+\lambda^4)\\ 1-3\lambda+11\lambda^2-17\lambda^3+14\lambda^4-6\lambda^5+\lambda^6)(-1+(1-\lambda)^{16n}+\ldots))/\\ \left(-16+120\lambda-560\lambda^2+\ldots-16\lambda^{14}+\lambda^{15}\right)\end{aligned} \tag{10}$$

We are interested in the limit for $m \to \infty$ since that forces $g(n) \to 1$. This limit is quite simple:

$$\frac{(-1+\lambda)^2\left(1-3\lambda+10\lambda^2-14\lambda^3+11\lambda^4-5\lambda^5+\lambda^6\right)}{2-8\lambda+28\lambda^2-56\lambda^3+70\lambda^4-56\lambda^5+28\lambda^6-8\lambda^7+\lambda^8}$$

The above showed in some detail the computation that applies to the case $p_1 \equiv_3 1$ and $p_1 \equiv_5 1$. The other seven cases are essentially similar, with the differences arising in the length and zero positions for initial segments, and in the exponents in the repeated parts. A simple computation shows that the averaged same parity frequency is as in (11), again taking the limit as the summation goes to infinity under the assumption that λ is small (recall it is $\mathcal{O}\left(\frac{1}{\log n}\right)$ and we assume $n \gg 1$).

$$\frac{8-36\lambda+128\lambda^2-278\lambda^3+384\lambda^4-340\lambda^5+188\lambda^6-59\lambda^7+8\lambda^8}{8\left(2-8\lambda+28\lambda^2-56\lambda^3+70\lambda^4-56\lambda^5+28\lambda^6-8\lambda^7+\lambda^8\right)} \tag{11}$$

An obvious question is how well this compares to observed frequencies. For the range shown in the examples, the frequency estimate for same parity last digits is around 0.4526. The observed frequency is $\frac{4485163}{9999999}$, or around 0.4485. The relative error is near 1%.

A similar computation was done with n around 10^{20}, using 10^5 consecutive primes. For this range the predicted frequency of same last digit pairs is 0.4797 and the observed frequency is 0.4819, with a relative error now slightly less than a half percent.

The expansion of the limiting frequency, to fourth order in λ is of some interest:

$$\frac{1}{2}-\frac{\lambda}{4}+\frac{\lambda^3}{8}-\frac{\lambda^5}{4}+O[\lambda]^6$$

We have a leading term of $\frac{1}{2}$ as expected, and the main correction term is $\frac{-\lambda}{4}$. A reasonable question is whether this might change were one to consider more small prime modular classes for p_1. Given the fairly close agreement with actual frequencies, it seems plausible that this might be the correct first order correction in accounting for the prime pair last-digit-mod-4 bias. One might be concerned that this has, at first glance, different asymptotic behavior than the conjectured result in [5]. We will address this seeming discrepancy in the next section.

4 Base 5

In this section we work with $p_1 \equiv_5 1$. We perform an asymptotic analytic computation similar to that for base 4, but using residue classes modulo 3, 7, and 11. This larger set of moduli means among other things that the asymptotic expressions become more complicated, but they remain tractable. For the second ten million primes, the frequency expectations for last digit pairs of ((1, 1), (1, 2), (1, 3), (1, 4)) are respectively (0.183, 0.303, 0.301, 0.214). Actual frequencies are quite close, at (0.183, 0.303, 0.299, 0.216). A similar computation shows that for the first million primes larger than 10^{20} the expected frequencies are (0.217, 0.271, 0.274, 0.238), with actual frequencies of (0.215, 0.271, 0.274, 0.240).

The series estimates, shown in (12) to second order in λ, have simple linear terms.

$$\left(\frac{1}{4} - \frac{3\lambda}{8} + \frac{4397\lambda^2}{4800}, \frac{1}{4} + \frac{5\lambda}{24} - \frac{1211\lambda^2}{4800}, \frac{1}{4} + \frac{\lambda}{4} - \frac{829\lambda^2}{4800}, \frac{1}{4} - \frac{\lambda}{12} - \frac{2357\lambda^2}{4800}\right) \tag{12}$$

Of importance is that the first order terms are identical to those in obtained from a similar analytic form, but computed using only residue classes for 3 and 7. This is evidence supporting the conjecture that the correct first order terms to a general analytic estimate for frequencies of last digit pairs in base 5 are as shown in (18).

Perhaps surprising is an observation from [5]: both estimated and actual frequencies for last digit pairs (1, 2) begin larger and eventually become smaller than for pairs (1, 3). We use the asymptotic estimates to study this.

Recall that for primes around n, λ varies as the inverse of $\log n$. If we want to find m such that the range of interest includes the mth prime, then we replace n by $m \log m$. Also there is the factor to account for the fact that all terms of nonzero probability correspond to increments known a priori to be non divisible by 2, 3, 5, 7, and 11. The upshot is we evaluate at $\lambda = \frac{F(5)}{\log(m \log m)}$. Of interest is whether the difference is negative for relatively large λ (corresponding to small n) and becomes positive as λ decreases. A plot from zero to 0.3 in Fig. 1 shows that this is indeed the case.

The crossing where the difference vanishes is at $\lambda \approx 0.215$. This corresponds to $m \approx 2.67\ 10^8$. Numeric tests suggest the actual crossover to be in the vicinity of $5.3\ 10^8$, (around twice the estimated value). A different threshold is reported in [5]

Fig. 1 $\mathcal{P}(1, 3)$–$\mathcal{P}(1, 2)$

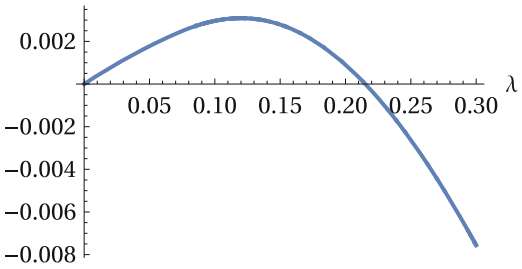
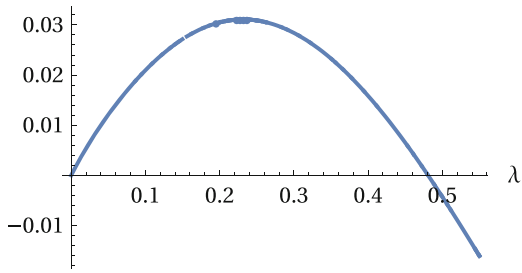


Fig. 2 $\mathcal{P}(1, 4)$–$\mathcal{P}(1, 1)$



for a related but different computation; the salient point is that both actual crossover values are in the ballpark of the estimated one.

A plot for the estimated frequency difference between last digit pairs (1, 4) and (1, 1) in Fig. 2 shows, perhaps more unexpectedly, a similar crossover (so initially the last digit "self-avoidance" is not readily apparent).

This suggests that there will be more last digit pairs of the form (1, 1) than (1, 4) in the "large λ" realm. The expectation crossover in Fig. 2 is around $m \approx 2800$. This also turns out to be a reasonable ballpark figure: from the first 2800 primes there are 98 consecutive pair last digits of the form (1, 4) and 95 of the form (1, 1), and easy tests indicate we are at this point past the crossover. However, if we only consider the first 1300 such pairs, there are 43 of the form (1, 1) and 42 of the form (1, 4), with larger differences appearing as we decrease the range e.g. to the first 900 consecutive prime pairs. So the analytic model has again provided an interesting subtlety, one that to rough approximation is reflected in the actual tallies. One might conclude that the consecutive pair last digits are not quite so "self-avoiding" as had been thought.

Indeed, when one checks last digit pairs in base 11, the (1, 10) frequency only regularly exceeds that of (1, 1) at somewhere between 50,000 and 60,000 such pairs. An analytic approximation again predicts such a crossover, at around 43,000. The analytic form also predicts a later crossover of the (1, 10) pairs with the (1, 6) pairs. Experiment seems to confirm that this indeed happens somewhere in the range between 10^{12} and 10^{13}, and the prediction from the analytic form is also in this ballpark, at 10^{12}.

We return now to a seeming disagreement in the bias asymptotics between the formulas (12) and Conjecture 1.1 in [5], where a bias on the order of $\frac{\log\log n}{\log n}$ is

given. In (12) one must replace λ with $\frac{F(5)}{\log n}$. Recall that a "proper" estimate of the type presented herein would use residue classes of more than just the very first primes, and indeed one would use as many as possible subject to the constraint that the primorial remains less than the range under scrutiny. This implies that when the primes are $O(n)$ we want to use the first $\log n$ primes for residue classes (excluding those that divide the base, but that has no bearing on the asymptotics here). So a better asymptotic expansion would have a first order term of $\frac{F(\log n)}{\log n}$ times a constant. Straightforward analysis moreover shows that $F(n)$ is to first order approximated as a constant times $\log n$. Thus (12) and conjectured behavior in [5] are in agreement in basic form.

5 Symmetry in Reversing Direction

The model presented does not require that one work from first to second prime in a consecutive pair. One could instead start at a prime and use the same frequency model to find its immediate predecessor. For example, if we consider base 4, the nonzero slots in (11) will be reversed when one assumes the larger prime is equal to -1 mod 3 and also -1 mod 5 and assigns frequency probabilities working from larger to smaller prime. A similar reversal happens for the repeated part in (10).

In general we have a given base b and a prime p_1, and we assume p_1 takes on a given value w_1 in that base and moreover a given set of nonzero values $\{v_1, v_2, \ldots\}$ modulo a set of small primes $\{q_1, q_2, \ldots\}$ all relatively prime to b. We later average over all possible values modulo those small primes. We want to model frequencies of equivalence classes for p_2 in base b. We take a specific such case $p_2 \equiv_b w_2$. The important point is that the initial and repeating segments of the model under scrutiny will be exactly reversed if we instead start with a new prime $\tilde{p_2} \equiv_b -w_1$, and $\tilde{p_2} \equiv \{-v_1, -v_2, \ldots\}$ modulo $\{q_1, q_2, \ldots\}$ respectively. Moreover if the next smaller prime $\tilde{p_1}$ satisfies $\tilde{p_1} \equiv_b -w_2$ then the frequency computation for this equivalence class of preceding prime is exactly the same as that for p_2 being the next prime after p_1. Thus contributions to the estimated frequency of p_2 following p_1 are in one to one correspondence with those for $\tilde{p_1}$ preceding $\tilde{p_2}$. This shows the symmetry across the antidiagonal, as was also the case for the model developed in [1].

6 Open Questions and Summary

There are some clear strengths to this frequency model. For base 3, using a set of 6 moduli gives expected frequencies that are remarkably accurate. In the range under scrutiny, for no value of $1 \leq k \leq 20$ did $\mathcal{P}(p_2 = p_1 + 2k)$ depart from the observed frequency by more than a few hundredths in absolute magnitude, and never was it off by more than around 7% in relative error.

To what extent is this frequency model related to, or might it be recast in terms of, the models developed in [1] or [5]? This is very much an open question. Another is to what extent might the general case be approximated by a closed form? A first step in this direction was shown for the special cases of bases 4 and 5, where explicit correction terms were computed. This itself raises more questions. Would these terms, or at least the terms of lowest order, remain correct in a finer-grained model e.g. one that uses more small primes and equivalence classes thereof?

The model presented herein does have some compelling features. It is simple, and computations with it are straightforward in any size range. For the bases under consideration it ranges, loosely speaking, from fairly to remarkably accurate. In addition to the examples shown, experiments in various ranges using bases 5 and 11 gave predicted pair frequencies that were also quite accurate. Using this methodology we are able to obtain closed form approximations for the cases of bases 4, 5, and 11 (that last not covered above, but substantially similar to the case of base 5). These have simple series expansions, thus for example giving explicit estimates for the frequency differences. This in turn allowed for an explicit estimate of the departure of the two parity cases in base 4 from frequencies of $\frac{1}{2}$. Moreover these predicted base 4 frequency differences compare well to actual frequencies in the several ranges that were checked. A similar analytic estimate was made for base 5, and was seen to be quite accurate in predicting the crossover in the frequencies of last digit pairs $(1, 2)$ vs. $(1, 3)$. It was qualitatively on target in predicting the (somewhat unexpected) crossover for last digit pairs $(1, 4)$ vs. $(1, 1)$. Such crossovers are also seen in last digit pairs in base 11, and again the analytic forms predict them. Given how well they seem to correspond to actual computations, these analytic estimates could be a step in the direction of quantifying behavior of lower order terms along the lines noted in [5].

Appendix: Wolfram Language Code

Get the second million primes.

```
max = 10^7;
nexttenmillionprimes = Prime[Range[max + 1, 2 * max]];
```

Tally residue classes for consecutive prime pairs, in bases 3, 4, and 10.

```
talliesmod3 = Sort[Tally[Partition[Mod[nexttenmillionprimes, 3], 2, 1]]];
talliesmod4 = Sort[Tally[Partition[Mod[nexttenmillionprimes, 4], 2, 1]]];
talliesmod10 = Sort[Tally[Partition[Mod[nexttenmillionprimes, 10], 2, 1]]];
```

Base 3 Frequency Estimates

Compute expected frequencies that second prime is first plus 2, 4, . . ., 90 in base 3 case, assuming first is equal to 1 mod 3. We only consider residue classes mod 5 in this next computation.

```
extrafactors3 = {0, 1, 1, 0, 4/3, 1, 0, 1, 1, 0, 1, 1, 0, 1, 4/3, 0, 1, 1, 0, 4/3, 1, 0, 1, 1,
  0, 1, 1, 0, 1, 4/3, 0, 1, 1, 0, 4/3, 1, 0, 1, 1, 0, 1, 1, 0, 1, 4/3};
len = Length[extrafactors3];
lognrecip = Integrate[1/Log[x],
  {x, Prime[max], Prime[2 * max]}]/(Prime[2 * max] - Prime[max]);
prob = 15/4. * lognrecip; mult = 3/4; pNext3[0] = 0;
Do[pNext3[k] = (1 - Sum[pNext3[j], {j, 1, k - 1}])*
  mult * prob * extrafactors3[[k]], {k, len}];
p3tab = Table[pNext3[k], {k, len}];
Sum[pNext3[k], {k, 3, len, 3}]/Sum[pNext3[k], {k, 2, len, 3}]
```

0.801624

Now use more prime residue classes.

```
qq[j_]:=Product[Prime[k], {k, j}]
rr[j_]:=Product[Prime[k] - 1, {k, j - 1}]
ss[j_]:=rr[j]/qq[j]
sTot[j_]:=Sum[ss[k], {k, j}]
frac[j_]:=1/(1 - sTot[j])
nmoduli = 6;
mult3 = Product[(Prime[j] - 2)/(Prime[j] - 1), {j, 3, nmoduli}];
lognrecip = Integrate[1/Log[x], {x, Prime[max], Prime[2 * max]}]/
  (Prime[2 * max] - Prime[max]);
prob3 = frac[nmoduli] * mult3 * N[lognrecip];
end = 8 * Round[Log[N[2 * max]]];
moduli = Table[Prime[j], {j, 3, nmoduli}];
extras3A = ConstantArray[1, end];
Do[extras3A[[j]] = 0, {j, 1, end, 3}];
Do[mod = moduli[[k]];
  Do[extras3A[[j]]* =(mod - 1)/(mod - 2), {j, mod, end, mod}]
  , {k, Length[moduli]}];
pNext3A[0] = 0;
prefactor3A[k_]:=1 - Sum[pNext3A[j], {j, 0, k - 1}]
Do[pNext3A[j] = prefactor3A[j] * extras3A[[j]] * prob3, {j, end}];
ttA = Table[pNext3A[j], {j, end}];
Total[ttA[[3;; - 1;;3]]]/Total[ttA[[2;; - 1;;3]]]
```

0.788925

Base 5 Recurrence and Frequency Estimates

```
biasSum[k_, lam_] =
  RSolveValue[{g[k + 1] == g[k] + (1 - g[k]) * lam, g[0]==0}, g[k], k];
bias5Initial235711[m_, n_, p_, start_, lam_]:=Module[
  {j1 = 3 - m, j2 = 7 - n, jcommon, mults = ExtendedGCD[-3, 7][[2]],
    top, i = 0, res},
  jcommon = (j1 - j2) * mults;
  top = j1 + jcommon[[1]] * 3;
  While[Mod[top + p, 11] ≠ 0, top+ =42];
  top = Mod[top, 2 * 3 * 7 * 11];
  While[Mod[top, 2]=!=0||Mod[top + start, 5]=!=0, top+ =3 * 7 * 11];
  If[top ≤ 5 + start, top+ =2 * 3 * 5 * 7 * 11];
  res = Prepend[Table[
    If[Mod[j + 1, 2] == 0||Mod[j + start, 5] == 0||Mod[j + m, 3] == 0||
      Mod[j + n, 7] == 0||Mod[j + p, 11] == 0, Null, i++];
    biasSum[i, lam], {j, top}], 0];
  {i, res}]
bias235711[skip_, lam_]:=Module[{m = skip},
  Table[If[Mod[j + 1, 2] == 0||Mod[j, 3] == 0||Mod[j, 5] == 0||
    Mod[j, 7] == 0||Mod[j, 11] == 0, Null, m++];
  biasSum[m, lam], {j, 2 * 3 * 5 * 7 * 11}]]
pMain[lm_, term_]:=Module[{bias, diffs, n},
  bias = bias235711[n, lm];
  diffs = Prepend[Differences[bias], 0];
  Cancel[Factor[Total[diffs[[term;; - 1;;5]]]]]/(1 - lm)^n]
repeats = Table[pMain[lm, j], {j, 5}];
biasBase5[m_, n_, p_, start_, len_, lam_]:=Module[
  {init, ideg, diffs, isums},
  {ideg, init} = bias5Initial235711[m, n, p, start, lam];
  diffs = Differences[init];
  isums = Map[If[Length[diffs] < #, 0, Total[diffs[[#;; - 1;;5]]]]&, Range[5]];
  isums = RotateRight[isums, start];
  isums + repeats * Sum[(1 - lam)^
    (ideg + EulerPhi[2 * 3 * 5 * 7 * 11] * (j - 1)), {j, len}]]
probs[start_, n_, lam_]:=
  Sum[biasBase5[i, j, k, start, n, lam], {i, 2}, {j, 6}, {k, 10}]/(2 * 6 * 10)
```

The next takes a couple of minutes or so to find the approximations.

```
pb1 = probs[1, Infinity, lm];   pb1Tog = Together[pb1];
pb2 = probs[2, Infinity, lm];   pb2Tog = Together[pb2];
pb3 = probs[3, Infinity, lm];   pb3Tog = Together[pb3];
pb4 = probs[4, Infinity, lm];   pb4Tog = Together[pb4];
```

```
series = {Series[pb1Tog, {lm, 0, 2}], Series[pb2Tog, {lm, 0, 2}],
  Series[pb3Tog, {lm, 0, 2}], Series[pb4Tog, {lm, 0, 2}]}
```

Evaluate the frequency estimates for the second 10^7 primes.

```
m = 10^7;
estimatedratios = N[N[Map[Most, {pb1Tog, pb2Tog, pb3Tog, pb4Tog}]/.
  lm → frac[5] * lognrecip, 200]]
```

```
{0.1828, 0.3030, 0.3005, 0.2137}, {0.2548, 0.1748, 0.2699, 0.3005},
  {0.2383, 0.2840, 0.1748, 0.3030}, {0.3241, 0.2383, 0.2548, 0.1828}}
```

Evaluate the frequency estimates for primes near 10^{20}.

```
m = 10^20;
estimatedratios = N[N[Map[Most, {pb1Tog, pb2Tog, pb3Tog, pb4Tog}]/.
  lm → frac[5]/Log[m], 200]]
```

```
{{0.2168, 0.2709, 0.2739, 0.2385}, {0.2523, 0.2157, 0.2582, 0.2739},
  {0.2476, 0.2658, 0.2157, 0.2709}, {0.2833, 0.2476, 0.2523, 0.2168}}
```

Check actual frequencies for the first million primes larger than 10^{20}. They are quite close to the estimates above.

```
plistBig20 = NextPrime[10^20, Range[10^6]];
talliesmod5Big = Sort[Tally[Partition[Mod[plistBig20, 5], 2, 1]]]
```

```
{{{1, 1}, 53727}, {{1, 2}, 67806}, {{1, 3}, 68557}, {{1, 4}, 60065},
  {{2, 1}, 63215}, {{2, 2}, 53282}, {{2, 3}, 64931}, {{2, 4}, 68225},
  {{3, 1}, 62140}, {{3, 2}, 66221}, {{3, 3}, 53309}, {{3, 4}, 68341},
  {{4, 1}, 71074}, {{4, 2}, 62344}, {{4, 3}, 63214}, {{4, 4}, 53548}}
```

```
tallies1mod5Big = Cases[talliesmod5Big, {{1, _}, n_} :→ n];
tallies2mod5Big = Cases[talliesmod5Big, {{2, _}, n_} :→ n];
tallies3mod5Big = Cases[talliesmod5Big, {{3, _}, n_} :→ n];
tallies4mod5Big = Cases[talliesmod5Big, {{4, _}, n_} :→ n];
ratios1 = N[tallies1mod5Big/Total[tallies1mod5Big]];
ratios2 = N[tallies2mod5Big/Total[tallies2mod5Big]];
ratios3 = N[tallies3mod5Big/Total[tallies3mod5Big]];
ratios4 = N[tallies4mod5Big/Total[tallies4mod5Big]];
ratios = {ratios1, ratios2, ratios3, ratios4}
```

```
{{0.2148, 0.27106, 0.2741, 0.2401}, {0.2532, 0.2134, 0.2601, 0.2733},
  {0.2485, 0.2649, 0.2132, 0.2734}, {0.2841, 0.2492, 0.2527, 0.2140}}
```

As seen below, the largest relative deviation between estimate and actual is under 1.3%.

```
percentErrors = Abs[100 * (ratios − estimatedratios)/ratios]
```

```
{{0.9512, 0.0736, 0.0687, 0.6893}, {0.3745, 1.0692, 0.7398, 0.2160},
  {0.3839, 0.3663, 1.1629, 0.9129}, {0.2701, 0.6428, 0.16257, 1.2988}}
```

The plots and estimated crossovers can be found as follows. We show this for the difference in counts when the second digits are, respectively, 2 or 3 mod 5.

```
pb32diff = pb1Tog[[3]] − pb1Tog[[2]];
Quiet[Plot[pb32diff, {lm, 0, 3/10}, WorkingPrecision → 100, AxesLabel → {"λ"},
  PlotLabel->"P(1, 3) − P(1, 2)"]]
pb32diff = pb1Tog[[3]] − pb1Tog[[2]];
cross32 = lm/.FindRoot[pb32diff == 0, {lm, 23/100}, WorkingPrecision → 650];
N[ncross32 = Quiet[n/.First[NSolve[frac[5]/Log[n * Log[n]] == cross32, n]]]]
```

2.66512×10^8

References

1. Ash A, Beltis L, Gross R, Sinnott W (2011) Frequencies of successive pairs of prime residues. Exp Math 20:400–411. https://doi.org/10.1080/10586458.2011.565256
2. Hardy GH, Littlewood JE (1923) Some problems of "Partitio numerorum"; iii: On the expression of a number as a sum of primes. Acta Math 44:1–70. https://doi.org/10.1112/plms/s2-22.1.46
3. Klarreich E (2016) Mathematicians discover prime conspiracy: A previously unnoticed property of prime numbers seems to violate a long-standing assumption about how they behave. Quanta Mag (2016)
4. Ko CM (2002) Distribution of the units digit of primes. Chaos Solitons Fractals 13:1295–1302. https://doi.org/10.1016/S0960-0779(01)00135-7
5. Oliver RL, Soundararajan K (2016) Unexpected biases in the distribution of consecutive primes. Proc Natl Acad Sci 113(31):E4446–E4454. https://doi.org/10.1073/pnas.1605366113
6. Wolfram Research, Inc., Mathematica, Version 11.3, Champaign, IL (2018)

Computational Aspects of Hamburger's Theorem

Yuri Matiyasevich

Abstract Riemann's zeta function (defined by a certain Dirichlet series) satisfies an identity known as the functional equation. H. Hamburger established that the function is identified by the equation inside a wide class of functions defined by Dirichlet series.

Riemann's zeta function is a member of a large family of functions with similar properties, in particular, satisfying certain functional equations. Hamburger's theorem can be extended to some (but not to all) of these equations.

The paper addresses the following question: how could we discover the Dirichlet series satisfying given functional equation? Two "rules of thumb" for performing such discoveries via numerical computations are demonstrated for functional equations satisfied by Dirichlet eta function, Ramanujan tau L-function, and Davenport–Heilbronn function.

A conjectured discrete version of Hamburger's theorem is stated.

1 Number-Theoretical Backgrounds

This introductory section presents some well-known definitions and results required for understanding the rest of the paper.

1.1 Riemann's Zeta Function

One of the most important open problems in Number Theory is the celebrated *Riemann Hypothesis*. It is a prediction about positions of the zeros of *Riemann's zeta function*. This function can be defined via *Dirichlet series*

Yu. Matiyasevich (✉)
St. Petersburg Department of V. A. Steklov Mathematical Institute, Saint Petersburg, Russia
e-mail: yumat@pdmi.ras.ru

© Springer Science+Business Media, LLC, part of Springer Nature 2019

N. Fillion et al. (eds.), *Algorithms and Complexity in Mathematics, Epistemology, and Science*, Fields Institute Communications 82,
https://doi.org/10.1007/978-1-4939-9051-1_8

$$\zeta(s) = \sum_{n=1}^{\infty} n^{-s}. \tag{1}$$

This series converges for $\Re(s) > 1$ but the function can be extended to the whole complex plane with the exception of the point $s = 1$ (at this point the zeta function has its only pole).

Riemann [23] conjectured that all non-real zeros of the zeta function lie on the *critical line* $\Re(s) = 1/2$.

While the function is named after Riemann, it was studied (for real values of the argument) already by L. Euler. He also worked with closely related entire function

$$\eta(s) = (1 - 2 \times 2^{-s})\zeta(s) = \sum_{n=1}^{\infty} (-1)^{n+1} n^{-s} \tag{2}$$

named *alternating zeta function* or *Dirichlet eta function*. The alternating series in (2) has the advantage over the series (1) of being convergent in the wider region $\Re(s) > 0$. Respectively, at this half-plane the zeta function can be calculated as

$$\zeta(s) = \frac{\eta(s)}{1 - 2 \times 2^{-s}} = \frac{\sum_{n=1}^{\infty} (-1)^{n+1} n^{-s}}{1 - 2 \times 2^{-s}}. \tag{3}$$

1.2 Euler Product

Euler also gave another, rather different from (1) and (3), definition of the zeta function:

$$\zeta(s) = \prod_{p \text{ prime}} \frac{1}{1 - p^{-s}}. \tag{4}$$

Similar to the series (1), the product in (4) also converges for $\Re(s) > 1$ only. The right hand side of (4) is nowadays known as *Euler product*.

In order to see why (4) is true one can at first observe that

$$\prod_{p \text{ prime}} \frac{1}{1 - p^{-s}} = \prod_{p \text{ prime}} \left(1 + p^{-s} + p^{-2s} + \dots\right), \tag{5}$$

and then apply the *Fundamental Theorem of Arithmetic*. This theorem states that every natural number has a unique factorization into product of powers of primes. This is equivalent to the fact that expanding the right hand side of (5) one gets exactly the right hand side of (1)!

The equivalence of two definitions, (1) and (4), explains why the zeta function is a very important tool in the study of prime numbers.

1.3 The Functional Equation

Euler began his study of the zeta function by determining its values at positive even integers. At first, he computationally discovered an approximate equality

$$\zeta(2) \approx \frac{\pi^2}{6} \tag{6}$$

by calculating (without computer![1]) many decimal digits of the left- and right-hand sides in (6). Later, he proved that the equality is in fact exact, and, more generally, that

$$\zeta(2m) = \frac{(-1)^{m+1}(2\pi)^{2m} B_{2m}}{2(2m)!}, \qquad m = 1, 2, \ldots; \tag{7}$$

here $B_0 = 1$, $B_1 = \frac{1}{2}$, $B_2 = \frac{1}{12}$, $B_3 = 0, \ldots$ are the *Bernoulli numbers*.

Euler also indicated values of the zeta function at negative integers:

$$\zeta(-n) = -\frac{B_{n+1}}{n+1}, \qquad n = 1, 2, \ldots \tag{8}$$

In particular,

$$\zeta(-2m) = 0, \qquad m = 1, 2, \ldots, \tag{9}$$

and today even negative integers are called the *trivial zeros* of the zeta function.

Comparison of (7) and (8) allows one to eliminate Bernoulli numbers and get the equality

$$2(2m-1)!\zeta(2m) = (-1)^m (2\pi)^{2m} \zeta(1-2m), \qquad m = 1, 2, \ldots \tag{10}$$

Euler [8, Sect. 10] claimed that, more generally, for every real s

$$g(s)\zeta(s) = g(1-s)\zeta(1-s) \tag{11}$$

where

$$g(s) = \pi^{-\frac{s}{2}}(s-1)\Gamma(\tfrac{s}{2}+1). \tag{12}$$

[1] In this respect it is interesting to note that A. Turing in [28] used the word "computer" having in mind "a man performing computations"; in this sense a computer (namely, Euler) was involved in the discovery of (6).

The identity (11) is known today as the *functional equation for the zeta function*. Malmstén [16] proved its validity for real s such that $0 < s < 1$. Riemann [23] did it in the full generality, that is, for s being any complex number.[2]

1.4 Hamburger's Theorem

Hamburger [12, 13] established that the functional equation (11) identifies the zeta function inside a wide class of functions defined by Dirichlet series. In particular, the zeta function is the only function $D(s)$ such that

- $D(s)$ can be defined for $\Re(s) > 1$ by convergent Dirichlet series of the form

$$D(s) = 1 + \sum_{n=2}^{\infty} a_n n^{-s}; \tag{13}$$

- $(s-1)D(s)$ is an entire function of finite order;
- $D(s)$ satisfies the functional equation

$$g(s)D(s) = g(1-s)D(1-s) \tag{14}$$

where $g(s)$ is defined by (12).

1.5 Further Generalizations

Riemann's zeta function is (historically first) member of a large family of functions with similar properties. Selberg [25] axiomatically described what is now known as *Selberg class* S. Each function from class S can be defined by a Dirichlet series as well as by (a counterpart of) Euler product, satisfies certain functional equation, and has some other feature akin to the zeta function.

Also all functions from the class S are expected to satisfy corresponding analogs of the Riemann Hypothesis, and for this the existence of Euler products and functional equations is believed to be indispensable.

Original Hamburger's results were extended to other functional equations and improved by weakening certain restrictions on the function; for a recent survey of such *converse theorem* see [22].

However, in general case the linear space of Dirichlet series satisfying certain functional equation has dimension greater than 1.

[2] This form of writing the functional identity is due to Kinkelin [15], Euler and Malmstén worked with an equivalent formula in terms of function $\eta(s)$; Riemann mentioned neither Euler nor Malmstén (for other historical details see, for example, [4]).

2 The Objective

We will look at converse theorems of Hamburger type from computational point of view. In this paper we confine ourselves to consideration of functional equations of the simplest form

$$h(s)D(s) = h(c-s)D(c-s). \tag{15}$$

Here c and $h(s)$ are given number and function respectively, and $D(s)$ is an unknown Dirichlet series with real coefficients,

$$D(s) = \sum_{n=1}^{\infty} a_n n^{-s}. \tag{16}$$

Suppose that we expect that (15) has a unique solution under certain extra restrictions on $D(s)$ (but we may not know these restrictions). The main question is: *how could we discover this series?*

We can distinguish two subquestions:

- *how could we calculate (approximate) values of the initial coefficients*, a_1, a_2, ...;
- *how could we calculate (approximate) value of function* $D(s)$ *and its derivatives for a given* s *(which need not lie in the half-plane of the convergence of the series)?*

In a more general situation we can expect only that the linear space of functions satisfying (15) has a finite dimension. Then we can ask:

- *could we select in a natural way a single "canonical" solution of* (15)*?*
- *how could we discover a basis for the linear space of solution of* (15)*?*

Methods for answering such questions were proposed by D. W. Farmer, S. Koutsoliotas, and S. Lemurell in [10, 11], and by D. W. Farmer and N. C. Ryan in [9].[3] Their main idea is briefly presented in the next section.

The author found (by computer experiments) several rather unexpected ways to answer above stated questions. This paper presents a few of the most interesting discoveries. The technique used in [19] and in Sects. 4–6 below looks somewhat resembling the technique from [9–11]. However, there are essential differences (explained in Sects. 4 and 5.3) between the two approaches, and the one presented here reveals quite another phenomenon in approximations of Dirichlet series.

[3]The author is grateful to the referee for indicating to these papers of which the author was ignorant at the time of writing [21].

3 Previously Proposed Technique

Here we outline the main idea from [10, 11]. To simplify notation we assume that $D(s)$ is an entire function, and the coefficients a_n in (16) are real.

The main tool used in [10, 11] is not the "genuine" functional equation (15) but what is called *smoothed approximate functional equations* (see [24, Sect. 3.2]). In spite of the name, they are exact equalities of the form

$$h(s)D(s) = \sum_{n=1}^{\infty} \big(f_G(s,n)n^{-s} + f_G(c-s,n)n^{c-s}\big)a_n. \tag{17}$$

Here the factors $f_G(s,n)$ and $f_G(c-s,n)$ are defined via certain auxiliary entire function G which should satisfy some mild conditions on its growth. The method exploits our freedom in selection of this function G. Namely, we can take two copies of (17) for suitable different functions G_1 and G_2:

$$h(s)D(s) = \sum_{n=1}^{\infty} \big(f_{G_1}(s,n)n^{-s} + f_{G_1}(c-s,n)n^{c-s}\big)a_n, \tag{18}$$

$$h(s)D(s) = \sum_{n=1}^{\infty} \big(f_{G_2}(s,n)n^{-s} + f_{G_2}(c-s,n)n^{c-s}\big)a_n. \tag{19}$$

Since the left-hand sides in (18)–(19), being independent of functions G_1 and G_2, are equal, we get, by substraction, the equality

$$\sum_{n=1}^{\infty} b_{G_1,G_2}(s,n)a_n = 0 \tag{20}$$

where

$$b_{G_1,G_2}(s,n) = \big((f_{G_1}(s,n) - f_{G_2}(s,n)\big)n^{-s} + \big(f_{G_1}(c-s,n) - f_{G_2}(c-s,n)\big)n^{1-s}. \tag{21}$$

Functions G_1 and G_2 can be chosen in such a way that numbers $b_{G_1,G_2}(s,n)$ in (20) decrease quickly, so fixing some N and taking only the first N summands, we get an approximate equality

$$\sum_{n=1}^{N} b_{G_1,G_2}(s,n)a_n \approx 0. \tag{22}$$

We can consider a formal counterpart of (22)

$$\sum_{n=1}^{N} b_{G_1,G_2}(s,n)a_{N,n} = 0 \tag{23}$$

where $a_{N,1}, \ldots, a_{N,N}$ are treated as unknowns. For different choices of $\langle s, G_1, G_2\rangle$ Eq. (23) are, in general, linear independent. Using sufficiently many such triples, we can determine the values of $a_{N,1}, \ldots, a_{N,N}$ by solving corresponding linear system. Taking into account the fast diminishing of $b_{G_1,G_2}(s,n)$ one might expect that the values of $a_{N,n}$ should be close to a_n. Numerical examples given in [10, 11] show that this indeed can happen.

4 Our First Technique

We will also try to approximate infinite Dirichlet series $D(s)$ (of the form (16) satisfying (15)) by a finite series

$$D_N(s) = \sum_{n=1}^{N} a_{N,n}n^{-s} \tag{24}$$

with some N. To this end we shall use functional equation itself, without any smoothing (thus we do not encounter the problem of selecting proper functions for the role of G). While identity (17) was obtained by *adding* (smoothed versions of) the left- and right-hand sides of (15), we will *subtract* them (non-smoothed); this is the first and the most essential distinction between our approach and the one used in [10, 11]. Namely, (15) implies that

$$\sum_{n=1}^{\infty} \big(h(s)n^{-s} - h(c-s)n^{c-s}\big)a_n = 0, \tag{25}$$

and, instead of (23), we will use the following formal counterpart of (25):

$$\sum_{n=1}^{N} \big(h(s)n^{-s} - h(c-s)n^{c-s}\big)a_{N,n} = 0. \tag{26}$$

The equality (25) holds only for s lying within the vertical strip where both series, (16) and

$$\sum_{n=1}^{\infty} a_n n^{s-c}, \tag{27}$$

converge. However, we are to use large real values for s, for which the series (27) diverges. Thus in our case the factors in front of a_n in (25) exponentially grow up instead of exponentially diminishing as factors $b_{G_1,G_2}(s,n)$ do in (20). This is the second important distinction from [10, 11]: we have no counterpart of (22) and thus we cannot reasonably justify our jump from (25) to (26).

Functional equation (15) can determine $D(s)$ up to a multiplicative constant only; thus we need a kind of normalization condition, and we impose that

$$a_1 = 1. \tag{28}$$

Respectively, we put

$$a_{N,1} = 1, \tag{29}$$

and define the remaining $N-1$ coefficients, $a_{N,2}, \ldots, a_{N,N}$, by solving the linear system

$$h(s)D_N(s) = h(c-s)D_N(c-s), \quad s \in \mathfrak{S}_N \tag{30}$$

where $\mathfrak{S}_N$ is a set containing $N-1$ elements (the equation in (30) is just another notation for (26)).

As it was explained earlier, in our case there is no evident reason to expect that so defined numbers $a_{N,n}$ should be close to a_n, and indeed they can be very different. Nevertheless, numerical data presented in the next two sections demonstrate that useful information about a_n can be gained from $a_{N,n}$, and finite series (24) can give very good approximation to $D(s)$.

5 Calculation of the Eta Function

Within this section we presuppose that in (15)

$$c = 1 \qquad \text{and} \qquad h(s) = \frac{g(s)}{1 - 2 \times 2^{-s}} = \frac{\pi^{-\frac{s}{2}}(s-1)\Gamma(\frac{s}{2}+1)}{1 - 2 \times 2^{-s}}. \tag{31}$$

With this choice, the Eq. (15) is satisfied by the function $\eta(s)$ (according to (2) and (11)) but this fact is used here only as a motivation to consider this particular functional equation.

5.1 Our Specialization

First of all, we need to select the set $\mathfrak{S}_N$ for constructing system (30); this can be done in many ways. Within this section we opt for

$$\mathfrak{S}_N = \{3/2, 5/2, \ldots, N - 1/2\}. \tag{32}$$

The reason for such a choice is as follows.

The gamma function (entering due to $h(s)$ into the both sides of the equation in (30)) satisfies the functional equation

$$\Gamma(s+1) = s\Gamma(s). \tag{33}$$

According to Bohr–Mollerup theorem [5], this equation (together with some other mild restrictions) uniquely determines the gamma function.

Equality (33) can be easily generalized: for a natural number m

$$\Gamma(z+m) = \left(\prod_{k=z}^{z+m-1} k\right) \Gamma(z). \tag{34}$$

The difference of the arguments of the gamma factors in the left- and right-hand sides in (30) is equal to $s - 1/2$, which is a positive integer whenever $s \in \mathfrak{S}_N$. Respectively, applying (34) we can make both arguments of the gamma function equal and hence cancel it. Thus for $s \in \mathfrak{S}_N$ Eq. (30) reduces to

$$h_1(s) D_N(s) = h_2(1-s) D_N(1-s) \tag{35}$$

where

$$h_1(s) = (2s-2)!! \left(1 - 2^{-s}\right), \tag{36}$$

$$h_2(s) = (-1)^{(2s-3)(2s-1)/8} \pi^{\frac{1}{2}-s} \left(1 - 2^{-s}\right). \tag{37}$$

5.2 Explicit Formulas

We can write down an explicit expression for $D_N(s)$. Consider $N \times N$ matrix

$$M_N(s) = \left(\mu_{m,n}(s)\right)\Big|_{m=1}^{N}\Big|_{n=1}^{N} \tag{38}$$

where

$$\mu_{m,n}(s) = \begin{cases} n^{-s}, & \text{if } m = 1, \\ h_1(m - 1/2)n^{1/2-m} - h_2(3/2 - m)n^{m-3/2}, & \text{otherwise.} \end{cases} \tag{39}$$

Let L_N be the $(N-1) \times (N-1)$ matrix resulting from $M_N(s)$ by deleting the first row and the first column. Then

$$D_N(s) = \frac{\det(M_N(s))}{\det(L_N)}. \tag{40}$$

5.3 Approximations

Table 1 shows the coefficients of $D_N(s)$ for $N = 50$. Examining the initial coefficients one can surmise that alternating $a_n = (-1)^{n+1}$ should give a solution of (15), and we know that this is indeed so.

More important is the observation that $D_N(s)$ gives good approximations to $\eta(s)$ for a large range of values of s—see Table 2. Respectively, (approximate) values of the zeta function and its derivatives can be calculated as

$$\frac{\mathrm{d}^k}{\mathrm{d}s^k}\zeta(s) \approx \frac{\mathrm{d}^k}{\mathrm{d}s^k}\frac{D_N(s)}{1 - 2 \times 2^{-s}}; \tag{41}$$

some numerical data are given in Table 3.

The fact that finite Dirichlet series can produce good approximations to $\eta(s)$ and its derivatives by itself is not surprising. Indeed, the coefficients of these finite series are just smooth truncations of the coefficients of the infinite series (2), and this is quite similar to (17). Borwein described in [7] another class of smooth truncations giving exponentially close approximations to $\eta(s)$. What is remarkable in our case is the origin of the smooth truncation. Namely, in [10, 11], in order to get smooth truncation, the authors need to select two functions G_1 and G_2; similar, the truncations in [7] are defined via selection of certain polynomials with special properties. Thus smooth truncations considered in [10, 11] and in [7] are not inherent to the zeta function solely. In contrast to them our definition of finite approximations $D_N(s)$ does not use auxiliary functions and thus is intrinsic to the zeta function.

Another kind of smooth truncation also intrinsic to the zeta function appeared in [17] (for further development see [3, 18, 20]). The coefficients of arising Dirichlet series encode a lot of information about prime numbers. However, from computational point of view that method is very complicated because it requires precalculations of the zeta zeros with high accuracy. In our case, entries to matrices $M_N(s)$ and L_N arise in a natural way from the functional equations (11) and (33)

Table 1 Coefficients of $D_{50}(s)$ defined by (24), (29), (31), (32), and (35)–(37)

n	$a_{50,n}$
1	1.0000000000000000...
3	0.9999999999989788...
5	0.9999999998274060...
7	0.9999999877970622...
9	0.9999995568036840...
11	0.9999906661069766...
13	0.9998765303771978...
15	0.9989149039834088...
17	0.9933942170370522...
19	0.9712077550029767...
21	0.9075744066371679...
23	0.7755520378458773...
25	0.5756614095764715...
27	0.3537541018518465...
29	0.1729474483683150...
31	0.0651146675641168...
33	0.0183498194062661...
35	0.0037635967925978...
37	0.0005445947606385...
39	0.0000535126895000...
41	0.0000033941685251...
43	0.0000001292559287...
45	0.0000000026399504...
47	0.0000000000236649...
49	0.0000000000000588...

n	$a_{50,n}$
2	−0.9999999999999517...
4	−0.9999999999849114...
6	−0.9999999984068907...
8	−0.9999999207764316...
10	−0.9999978351017259...
12	−0.9999641614198762...
14	−0.9996158801266156...
16	−0.9972031802946021...
18	−0.9856470926223216...
20	−0.9464927728130374...
22	−0.8511115240177615...
24	−0.6822017143861291...
26	−0.4632926781019783...
28	−0.2550737531447883...
30	−0.1098431344988656...
32	−0.0359028510830116...
34	−0.0086628349972250...
36	−0.0014987102715588...
38	−0.0001796757789624...
40	−0.0000142901860447...
42	−0.0000007100592081...
44	−0.0000000201600405...
46	−0.0000000002822698...
48	−0.0000000000014588...
50	−0.0000000000000011...

Table 2 Approximation of $\eta(s)$ by $D_{50}(s)$ defined by (24), (29), (31), (32), and (35)–(37)

s	$\left\lvert \frac{D_{50}(s)}{\eta(s)} - 1 \right\rvert$	s	$\left\lvert \frac{D_{50}(s)}{\eta(s)} - 1 \right\rvert$
-35	$1.05021...\cdot10^{-8}$	10i	$4.59065...\cdot10^{-13}$
$-33+1$i	$1.97364...\cdot10^{-9}$	16i	$7.65086...\cdot10^{-11}$
$-31+3$i	$2.06803...\cdot10^{-9}$	22i	$9.21145...\cdot10^{-9}$
$-29+5$i	$2.22572...\cdot10^{-9}$	0.5	$1.26385...\cdot10^{-15}$
$-27+7$i	$2.61390...\cdot10^{-9}$	$0.5+6$i	$9.76838...\cdot10^{-15}$
$-25+8$i	$1.06488...\cdot10^{-9}$	$0.5+8$i	$8.51062...\cdot10^{-14}$
$-23+10$i	$1.62068...\cdot10^{-9}$	$0.5+11$i	$7.51379...\cdot10^{-13}$
$-21+12$i	$2.69873...\cdot10^{-9}$	$0.5+13$i	$5.86538...\cdot10^{-12}$
$-19+13$i	$1.78099...\cdot10^{-9}$	$0.5+15$i	$4.16476...\cdot10^{-11}$
$-17+14$i	$1.37291...\cdot10^{-9}$	$0.5+18$i	$7.69482...\cdot10^{-10}$
$-15+15$i	$1.22591...\cdot10^{-9}$	$0.5+20$i	$2.09350...\cdot10^{-9}$
$-13+16$i	$1.25674...\cdot10^{-9}$	1	$1.24841...\cdot10^{-15}$
$-11+17$i	$1.46323...\cdot10^{-9}$	$1+6$i	$9.42471...\cdot10^{-15}$
$-9+18$i	$1.90802...\cdot10^{-9}$	$1+12$i	$1.28347...\cdot10^{-12}$
$-5+19$i	$1.86971...\cdot10^{-9}$	$1+20$i	$1.20238...\cdot10^{-9}$
0	$1.24841...\cdot10^{-15}$	$3+26$i	$1.07592...\cdot10^{-9}$
5i	$4.81114...\cdot10^{-15}$	$5+38$i	$1.02395...\cdot10^{-9}$

but definitions of these entries (given by (36), (37) and (39)) use only "simple" functions like the exponentiation and the double factorial.

It is interesting to study other properties of matrices $M_N(s)$, in particular, their eigenvalues and singular values (they "feel" zeta zeros).

5.4 Conjectures

Numerical data (presented in this paper and other calculations performed by the author) allow one to state a number of conjectures.

Conjecture A *For every n*

$$\lim_{N\to\infty} a_{N,n} = (-1)^{n+1}. \tag{42}$$

Conjecture B *For every s*

$$\eta(s) = \lim_{N\to\infty} \frac{\det(M_N(s))}{\det(L_N)}. \tag{43}$$

Conjectures A and B say that the coefficients of the Dirichlet series for $\eta(s)$ and values of this function can be calculated from the meager information contained in (29) and (35)–(37) for $s \in \mathfrak{S}_N$. Does it indicate that Bohr–Mollerup's and

Table 3 Approximation of the derivatives of $\zeta(s)$ via $D_{500}(s)$ defined by (24), (29), (31), (32), and (35)–(37)

	$\left\lvert \frac{\mathrm{d}^m}{\mathrm{d}s^m}\zeta(s) \Big/ \frac{\mathrm{d}^m}{\mathrm{d}s^m}\frac{D_{500}(s)}{1-2\times 2^{-s}} - 1 \right\rvert$		
s	$m=1$	$m=5$	$m=30$
-5	$0.101...\cdot 10^{-149}$	$0.136...\cdot 10^{-149}$	$0.784...\cdot 10^{-150}$
-3	$0.717...\cdot 10^{-150}$	$0.102...\cdot 10^{-149}$	$0.784...\cdot 10^{-150}$
-1	$0.771...\cdot 10^{-150}$	$0.780...\cdot 10^{-150}$	$0.784...\cdot 10^{-150}$
-0.5	$0.794...\cdot 10^{-150}$	$0.784...\cdot 10^{-150}$	$0.784...\cdot 10^{-150}$
$-0.5+10\mathrm{i}$	$0.948...\cdot 10^{-148}$	$0.442...\cdot 10^{-147}$	$0.447...\cdot 10^{-127}$
$-0.5+20\mathrm{i}$	$0.485...\cdot 10^{-143}$	$0.121...\cdot 10^{-142}$	$0.285...\cdot 10^{-128}$
$-0.5+40\mathrm{i}$	$0.845...\cdot 10^{-133}$	$0.108...\cdot 10^{-132}$	$0.237...\cdot 10^{-130}$
$-0.5+80\mathrm{i}$	$0.271...\cdot 10^{-114}$	$0.311...\cdot 10^{-114}$	$0.293...\cdot 10^{-105}$
$-0.5+160\mathrm{i}$	$0.126...\cdot 10^{-83}$	$0.163...\cdot 10^{-83}$	$0.343...\cdot 10^{-83}$
$-0.5+320\mathrm{i}$	$0.915...\cdot 10^{-41}$	$0.122...\cdot 10^{-40}$	$0.304...\cdot 10^{-40}$
$-0.5+640\mathrm{i}$	$0.800...\cdot 10^{-4}$	$0.957...\cdot 10^{-4}$	$0.137...\cdot 10^{-3}$
0.5	$0.791...\cdot 10^{-150}$	$0.784...\cdot 10^{-150}$	$0.784...\cdot 10^{-150}$
$0.5+10\mathrm{i}$	$0.138...\cdot 10^{-147}$	$0.178...\cdot 10^{-146}$	$0.115...\cdot 10^{-119}$
$0.5+20\mathrm{i}$	$0.705...\cdot 10^{-143}$	$0.182...\cdot 10^{-142}$	$0.107...\cdot 10^{-124}$
$0.5+40\mathrm{i}$	$0.100...\cdot 10^{-132}$	$0.142...\cdot 10^{-132}$	$0.108...\cdot 10^{-128}$
$0.5+80\mathrm{i}$	$0.273...\cdot 10^{-114}$	$0.372...\cdot 10^{-114}$	$0.180...\cdot 10^{-100}$
$0.5+160\mathrm{i}$	$0.881...\cdot 10^{-84}$	$0.157...\cdot 10^{-83}$	$0.135...\cdot 10^{-81}$
$0.5+320\mathrm{i}$	$0.460...\cdot 10^{-41}$	$0.980...\cdot 10^{-41}$	$0.307...\cdot 10^{-40}$
$0.5+640\mathrm{i}$	$0.461...\cdot 10^{-4}$	$0.796...\cdot 10^{-4}$	$0.129...\cdot 10^{-3}$

Hamburger's theorems could be combined and produce the following discrete version of the latter theorem?

Conjecture C *Riemann's zeta function is the only function $D(s)$ such that*

- *$D(s)$ can be defined for $\Re(s) > 1$ by a convergent Dirichlet series of the form*

$$D(s) = 1 + \sum_{n=2}^{\infty} a_n n^{-s}; \tag{44}$$

- *$(s-1)D(s)$ is an entire function of finite order;*
- *for $m = 1, 2, \ldots$ function $D(s)$ satisfies the numerical equalities*

$$\tilde{h}_1(m+1/2)D(m+1/2) = \tilde{h}_2(1/2-m)D(1/2-m) \tag{45}$$

where

$$\tilde{h}_1(s) = 2^{1-2s}(2s-2)!!, \tag{46}$$

$$\tilde{h}_2(s) = (-1)^{(2s-3)(2s-1)/8}\pi^{\frac{1}{2}-s}. \tag{47}$$

5.5 *Other Options*

The selection of the set (32) is not rigid, it can be replaced by many other sets. For example, for values of s we could use integers greater than 1. In this case Eq. (30) reduces to counterparts of equalities (7) and (9) found already to Euler. Namely, (30) for an odd $s = 2m + 1$ simplifies to equation

$$D_N(-2m) = 0, \tag{48}$$

and to equation

$$2(2m-1)!(1-2\times 2^{2m-1})D_N(2m) = \\ (-1)^m(2\pi)^{2m}(1-2\times 2^{-2m})D_N(1-2m) \tag{49}$$

for an even $s = 2m$.

Integers can be uses for values of s both instead of half-integers or together with them; in the latter case the accuracy of approximation of $\eta(s)$ by $D_N(s)$ is considerably higher.

Naturally, one can extend Conjectures A, B, and C for other choices of the set $\mathfrak{S}_N$.

6 An Equation with Many Solutions

Within this section we presuppose that in (15)

$$c = 1 \qquad \text{and} \qquad h(s) = 5^{s/2}\pi^{-s/2}\Gamma(s/2). \tag{50}$$

For these parameters the functional equation (15) is satisfied by Dirichlet L-function

$$L(\xi_5^{\langle 3\rangle}, s) = 1^{-s} - 2^{-s} - 3^{-s} + 4^{-s} + \\ 6^{-s} - 7^{-s} - 8^{-s} + 9^{-s} + \ldots, \tag{51}$$

and also[4] by the product

$$F(s) = (1+\sqrt{5}\times 5^{-s})\zeta(s) = \\ = 1^{-s} + 2^{-s} + 3^{-s} + 4^{-s} + (1+\sqrt{5})5^{-s} + \\ 6^{-s} + 7^{-s} + 8^{-s} + 9^{-s} + (1+\sqrt{5})10^{-s} + \ldots \tag{52}$$

[4]This example of a pair of functions solving the same functional equation was considered by E. P. Balanzario and J. Sánchez-Ortiz in [1, 2].

Thus there are infinitely many Dirichlet series (for example, any linear combination of (51) and (52)) satisfying (15) for c and $h(s)$ from (50), so it is not evident what will be the behavior of the coefficients of our finite Dirichlet series $D_N(s)$ in this case.

6.1 Numerical Data I

We begin by defining coefficients $a_{N,n}, \ldots, a_{N,N}$ of $D_N(s)$ via (29) and (30) for $N-1$ integral and half-integral values of s, namely, for

$$s \in \mathfrak{S}_N = \left\{ \frac{3}{2},\ 2,\ \frac{5}{2},\ \ldots,\ \frac{N+1}{2} \right\}. \tag{53}$$

Table 4 shows corresponding values of $a_{150,1}, \ldots, a_{150,25}$. They clearly proffer series (51) as a solution of our functional equation.

Table 4 Initial coefficients of $D_{150}(s)$ defined by (24), (29), (30), (50) and (53)

n	$a_{150,n}$
1	1.000000000000000000000000000000000000000000...
2	−1.000000000000000000000000000000000000503818...
3	−0.999999999999999999999999999999999993511570...
4	0.999999999999999999999999999999999920664778...
5	0.000000000000000000000000000000001297531999...
6	0.999999999999999999999999999999971313135966...
7	−0.999999999999999999999999999999219413414513...
8	−1.000000000000000000000000000023824492196236...
9	1.000000000000000000000000000762228891639139...
10	−0.000000000000000000000000024432067893709509...
11	1.000000000000000000000000761946711258008472...
12	−1.000000000000000000000022691776540366859765...
13	−0.999999999999999999993621642650290502711163...
14	0.999999999999999998319846279182551921841 9...
15	0.000000000000000041305010562713988279963 2...
16	0.999999999999999054301295219369613070035 5...
17	−0.999999999999979850687566768172476885047 7...
18	−1.000000000000399549497884122952690241087 4...
19	1.000000000007378760820471220334964556336 9...
20	−0.000000000127046778229379301700927833876 1...
21	1.000000002042117489727262835481085971222 2...
22	−1.000000030687875684522790169984583413440 6...
23	−0.999999568194715174192710449429396517499 4...
24	0.999994301980549516254020670929524272344 4...

It would be interesting to find the "reason" why numbers $a_{N,n}$ "vote" so strongly in favour of (51). One possible explanation is as follows: this series defines an entire function while (52) has a pole.

Another elucidation can be due to the following fact proved (in a greater generality) by J. Kaczorowski, G. Molteni, and A. Perelli in [14]: *among all functions satisfying the functional equation* (15) *for* c *and* $h(s)$ *from* (50)*, which are defined by Dirichlet series and fulfill some other natural conditions, only one (up to a multiplicative constant) function has an Euler product, namely, Dirichlet L-function* (51). Thus we can say that, in a sense, our method of solving the functional equation "is aware of" the existence of the Euler product.

6.2 Numerical Data II

In order to discover another solution, linear independent from (51), we need to work with a different functional equation.

Similar to what was done in Sect. 1, let us consider function

$$\tilde{h}(s) = \frac{h(s)}{1 - 2 \times 2^{-s}} = \frac{5^{s/2}\pi^{-s/2}\Gamma(s/2)}{1 - 2 \times 2^{-s}} \tag{54}$$

and functional equation

$$\tilde{h}(s)\tilde{D}(s) = \tilde{h}(1 - s)\tilde{D}(1 - s) \tag{55}$$

where

$$\tilde{D}(s) = \sum_{n=1}^{\infty} \tilde{a}_n n^{-s}. \tag{56}$$

Clearly, solutions of (15) and (55) are related in the following way:

$$D(s) = \frac{\tilde{D}(s)}{1 - 2 \times 2^{-s}}. \tag{57}$$

Again we introduce finite Dirichlet series

$$\tilde{D}_N(s) = \sum_{n=1}^{N} \tilde{a}_{N,n} n^{-s} \tag{58}$$

and imitate (55) by

$$\tilde{h}(s)\tilde{D}_N(s) = \tilde{h}(1 - s)\tilde{D}_N(1 - s). \tag{59}$$

Table 5 Initial coefficients of $\tilde{D}_{150}(s)$ defined by (53)–(54) and (58)–(60)

n	$\tilde{a}_{150,n}$	n	$\tilde{a}_{150,n}$
1	1	13	$2.11051...\cdot 10^{32}$
2	$2.11051...\cdot 10^{32}$	14	$-4.22102...\cdot 10^{32}$
3	$2.11051...\cdot 10^{32}$	15	$3.41488...\cdot 10^{32}$
4	$-4.22102...\cdot 10^{32}$	16	$-4.22102...\cdot 10^{32}$
5	$3.41488...\cdot 10^{32}$	17	$2.11051...\cdot 10^{32}$
6	$-4.22102...\cdot 10^{32}$	18	$2.11051...\cdot 10^{32}$
7	$2.11051...\cdot 10^{32}$	19	$-1.14278...\cdot 10^{21}$
8	$2.11051...\cdot 10^{32}$	20	$-3.41488...\cdot 10^{32}$
9	$-9.73518...\cdot 10^{6}$	21	$-3.27188...\cdot 10^{23}$
10	$-3.41488...\cdot 10^{32}$	22	$2.11051...\cdot 10^{32}$
11	$-1.01328...\cdot 10^{10}$	23	$2.11051...\cdot 10^{32}$
12	$2.11051...\cdot 10^{32}$	24	$-4.22101...\cdot 10^{32}$

Table 5 shows values of $\tilde{a}_{N,1}, \ldots, \tilde{a}_{N,24}$ obtained by solving the system consisting of Eq. (59) for $s \in \mathfrak{S}_N$ and normalization condition

$$\tilde{a}_{N,1} = 1 \tag{60}$$

for $N = 150$. Extremely large values of all coefficients, different from the default (60), suggest that this normalization was not felicitous. So we perform renormalization via dividing all the coefficients by $\tilde{a}_{N,2}$. Resulting ratios (presented in Table 6) also give a solution to (59) for $s \in \mathfrak{S}_N$.

Examination of the values in Table 6 produces the following surmises about the coefficients of a solution of (55):

- $\tilde{a}_1 = \tilde{a}_9 = \tilde{a}_{11} = \tilde{a}_{19} = \tilde{a}_{21} = 0$;
- $\tilde{a}_2 = \tilde{a}_3 = \tilde{a}_7 = \tilde{a}_8 = \tilde{a}_{12} = \tilde{a}_{13} = \tilde{a}_{17} = \tilde{a}_{18} = \tilde{a}_{22} = \tilde{a}_{23} = 1$;
- $\tilde{a}_4 = \tilde{a}_6 = \tilde{a}_{14} = \tilde{a}_{16} = \tilde{a}_{24} = -2$;
- $\tilde{a}_5 = -\tilde{a}_{10} = \tilde{a}_{15} = -\tilde{a}_{20} = \phi$ where $\phi = 1.618033988\ldots$

Both Wolfram Alpha [31] and The Inverse Symbolic Calculator [30] recognize 1.618033988 as the familiar golden ratio, $\phi = (1 + \sqrt{5})/2$.

Now performing formal division in (57) we get the following values for the 24 initial coefficients of $D(s)$:

- $a_1 = a_4 = a_6 = a_9 = a_{11} = a_{14} = a_{16} = a_{19} = a_{21} = a_{24} = 0$;
- $a_2 = a_3 = a_7 = a_8 = a_{12} = a_{13} = a_{17} = a_{18} = a_{22} = a_{23} = 1$;
- $a_5 = a_{10} = a_{15} = a_{20} = \phi$.

It is quite natural to make a general guess that for all k

- $a_{5k+1} = a_{5k+4} = 0$;
- $a_{5k+2} = a_{5k+3} = 1$;
- $a_{5k} = \phi$.

Table 6 Initial renormalized coefficients of $\tilde{D}_{150}(s)$ defined by (53)–(54) and (58)–(60)

n	$\tilde{a}_{150,n}/\tilde{a}_{150,2}$
1	0.00000000000000000000000000000000047381833...
2	1.00000000000000000000000000000000000000000...
3	1.00000000000000000000000000000000090824864...
4	−2.00000000000000000000000000000000097325743...
5	1.61803398874989484820458683436558728714 25...
6	−1.99999999999999999999999999999983642221007...
7	0.99999999999999999999999999995544680005399...
8	1.00000000000000000000000000014144310868154...
9	−0.00000000000000000000000000461271066102859...
10	−1.61803398874989484820458532591511140361 91...
11	−0.00000000000000000000004801147251082613 23...
12	1.00000000000000000000145925597371797969 15...
13	0.99999999999999999995814850981896174626 55...
14	−1.99999999999999999887557243601807556187 86...
15	1.61803398874989482002214324748487266408 23...
16	−1.99999999999999934245699524107693644677 55...
17	0.99999999999998572984766691805274030657 51...
18	1.00000000000028810003257501330383347363 96...
19	−0.00000000005414733105744837447877268938 8...
20	−1.61803398865505176025536397883861150147 37...
21	−0.00000001550280336901644966939857430466 2...
22	1.00000023682877800892387729240654559783 1...
23	0.99999966134647553554834779826668046542 66...
24	−1.99999545991366161692505669529093792106 37...

But this is equivalent to saying that

$$D(s) = F(s)/2 - L(\xi_5^{\langle 3 \rangle}, s)/2, \tag{61}$$

thus we have discovered a second solution of the functional equation (15) for parameters (50). According to [2], all solutions of this equation with periodic coefficients are linear combination of the two functions (51) and (52).

7 Our Second Technique

Functional equation (15) is equivalent to the statement that the function

$$F(z) = h(c/2 + z)D(c/2 + z) \tag{62}$$

is even. Thus *functional* equation (15) is equivalent to the infinite system of *numerical* equalities

$$\frac{\mathrm{d}^k}{\mathrm{d}z^k} F(z)\bigg|_{z=0} = 0, \qquad k = 1, 3, \ldots, 2m+1, \ldots \tag{63}$$

In terms of $D(s)$ this corresponds to

$$\frac{\mathrm{d}^k}{\mathrm{d}s^k}\Big(h(s)D(s)\Big)\bigg|_{s=c/2} = 0, \qquad k = 1, 3, \ldots, 2m+1, \ldots \tag{64}$$

We again impose normalizing condition (29) and define coefficients $a_{N,2}, \ldots, a_{N,N}$ from (24) by solving the system consisting of $N-1$ analogs of (64) of the form

$$\frac{\mathrm{d}^k}{\mathrm{d}s^k}\Big(h(s)D_N(s)\Big)\bigg|_{s=c/2} = 0 \tag{65}$$

with odd k. This resembles expansion of a function into Taylor series but there are two important distinctions.

First of all, we use odd derivatives only, which for even functions are trivial zeros; all the information about such functions is contained in even derivatives which we ignore.

Second, there is no need to take all consecutive initial derivatives, one can use (65) with k from different sets consisting of $N-1$ odd numbers.

8 Davenport–Heilbronn Function

Within this section we presuppose that in (15)

$$c = 1 \qquad \text{and} \qquad h(s) = \left(\frac{5}{\pi}\right)^{s/2} \Gamma\left(\frac{s}{2} + \frac{1}{2}\right). \tag{66}$$

8.1 Guessing the Coefficients

Table 7 shows, for $N = 30,\ 60,\ 90$, values of $a_{N,2}, \ldots, a_{N,12}$ defined by (29) and (65) for odd $k = 1, \ldots, 2N-1$.

The numerical data suggest the following surmises for $N \to \infty$:

- coefficients $a_{N,2}$, $a_{N,7}$, and $a_{N,12}$ approach certain limiting value

$$\alpha \approx 0.2840790404; \tag{67}$$

Table 7 Initial coefficients of $D_N(s)$ defined by (24), (29) (65) (for $k = 1,\ 3,\ \ldots,\ 2N-1$), and (66)

n	N	$a_{N,n}$
2	30	0.284139345050532242364880 2...
	60	0.284079043840357318902642 4...
	90	0.284079043840412296028291 3...
3	30	−0.284474727238260039908662 2...
	60	−0.284079043840037008264979 2...
	90	−0.284079043840412296028288 8...
4	30	−0.997715705981727718668987 1...
	60	−1.000000000002469240327472 1...
	90	−1.000000000000000000000019 8...
5	30	−0.014268398863186655202364 1...
	60	0.000000000020186909246366 8...
	90	0.000000000000000000000155 3...
6	30	1.092068744923690287798295 7...
	60	0.999999998079004381711738...
	90	0.999999999999999999991173...
7	30	−0.282741386615912827905288 5...
	60	0.284079045686703087258059 1...
	90	0.284079043840412296008016 8...
8	30	2.878473071049254908844632 9...
	60	−0.284079059227310217251188 1...
	90	−0.284079043840412294713983 9...
9	30	−16.56795005298874873675746 8...
	60	−0.999999919209720317580488 8...
	90	−1.000000000000000048747370 1...
10	30	66.742610537951756948237559 2...
	60	0.000000382832663697081393 2...
	90	0.000000000000001472684596 6...
11	30	−246.76040187990681589858620 4...
	60	0.999981555365658330907981 0...
	90	0.999999999999613898393090...
12	30	794.830080537829612219894350 7...
	60	0.284399708340596524171863 3...
	90	0.284079043841308521671857 6...

- coefficients $a_{N,3}$ and $a_{N,8}$ approach $-\alpha$;
- coefficients $a_{N,4}$ and $a_{N,9}$ approach -1;
- coefficients $a_{N,5}$ and $a_{N,10}$ approach 0;
- coefficients $a_{N,6}$ and $a_{N,11}$ approach 1.

The above surmises can be generalized by guessing that for all n coefficients $a_{N,n}$ approach certain limiting quantity a_n which depends only on the value of n modulo 5. Respectively, we can expect that the Dirichlet series

$$\sum_{m=0}^{\infty} (5m+1)^{-s} + \alpha(5m+2)^{-s} - \alpha(5m+3)^{-s} - (5m+4)^{-s} \tag{68}$$

is a solution of (15) for (66).

As for the nature of α, both The Inverse Symbolic Calculator [30] and Wolfram Alpha [31] suggest that (67) is a root of the equation

$$\alpha^4 + 2\alpha^3 - 6\alpha^2 - 2\alpha + 1 = 0, \tag{69}$$

that is

$$\alpha = \frac{-1 - \sqrt{5} + \sqrt{10 + 2\sqrt{5}}}{2}. \tag{70}$$

With this value of α function (68) is the well-known *Davenport–Heilbronn function* $f(s)$. It indeed satisfies functional equation (15) with c and $h(s)$ defined by (66) (see [27, 10.25]) and is the only solution of this equation (see [6, 5.1] or [2, Sect. 8]). However, $f(s)$ cannot be represented by an Euler product, and has non-real zeros outside the critical line.

8.2 Approximation of the Function and Its Derivatives

When calculating coefficients $a_{30,n}$ we imposed restrictions of two kinds:

- normalization $a_{30,1} = 1$;
- vanishing of the *odd* derivatives of the product $h(s)D_{30}(s)$ at $s = 1/2$.

Surprisingly, the values of *even* derivatives of $h(s)D_{30}(s)$ give very good approximations to the values of corresponding derivatives of the product $h(s)f(s)$ at $s = 1/2$—see Table 8.

Thus $D_N(s)$ and $f(s)$ have close initial fragments of Taylor series and respectively $D_N(s)$ gives a good approximations to $f(s)$ whenever $|s - 1/2|$ is not too large—see Table 9. This fact is more peculiar than the good approximations of $\eta(s)$ and $\zeta(s)$ demonstrated in Tables 2 and 3, and the reason why it is so startling is as follows. Table 10 presents all coefficients of $D_{30}(s)$; we see that, except for a few initial, these coefficients differ very much from the coefficients in (68).

Table 8 Comparison of even derivatives of $h(s)f(s)$ and $h(s)D_{30}(s)$ for $D_{30}(s)$ defined by (24), (29) (65) (for $k = 1,\ 3,\ \ldots,\ 59$), and (66)

m	$\frac{\frac{\mathrm{d}^m}{(\mathrm{d}s)^m}(h(s)D_{30}(s))\Big\|_{s=1/2}}{\frac{\mathrm{d}^m}{(\mathrm{d}s)^m}(h(s)f(s))\Big\|_{s=1/2}} - 1$	m	$\frac{\frac{\mathrm{d}^m}{(\mathrm{d}s)^m}(h(s)D_{30}(s))\Big\|_{s=1/2}}{\frac{\mathrm{d}^m}{(\mathrm{d}s)^m}(h(s)f(s))\Big\|_{s=1/2}} - 1$
0	$4.10785\ldots\cdot10^{-6}$	26	$3.38965\ldots\cdot10^{-12}$
2	$8.41657\ldots\cdot10^{-7}$	28	$1.43867\ldots\cdot10^{-12}$
4	$2.13426\ldots\cdot10^{-7}$	30	$6.20969\ldots\cdot10^{-13}$
6	$6.15266\ldots\cdot10^{-8}$	32	$2.64828\ldots\cdot10^{-13}$
8	$1.93507\ldots\cdot10^{-8}$	34	$1.34226\ldots\cdot10^{-13}$
10	$6.48438\ldots\cdot10^{-9}$	36	$-1.05994\ldots\cdot10^{-14}$
12	$2.28126\ldots\cdot10^{-9}$	38	$3.41245\ldots\cdot10^{-13}$
14	$8.34335\ldots\cdot10^{-10}$	40	$-1.99099\ldots\cdot10^{-12}$
16	$3.15036\ldots\cdot10^{-10}$	42	$1.61377\ldots\cdot10^{-11}$
18	$1.22191\ldots\cdot10^{-10}$	44	$-1.70056\ldots\cdot10^{-10}$
20	$4.84975\ldots\cdot10^{-11}$	46	$2.41375\ldots\cdot10^{-9}$
22	$1.96387\ldots\cdot10^{-11}$	48	$-4.79564\ldots\cdot10^{-8}$
24	$8.09412\ldots\cdot10^{-12}$	50	$1.40744\ldots\cdot10^{-6}$

Table 9 Approximation of $f(s)$ by $D_{30}(s)$ defined by (24), (29) (65) (for $k = 1,\ 3,\ \ldots,\ 59$), and (66)

s	$\left\lvert\frac{D_{30}(s)}{f(s)} - 1\right\rvert$	s	$\left\lvert\frac{D_{30}(s)}{f(s)} - 1\right\rvert$
-12	$2.36222\ldots\cdot10^{-7}$	10i	$1.84408\ldots\cdot10^{-4}$
$-10+10\mathrm{i}$	$2.48929\ldots\cdot10^{-8}$	0.5	$4.10785\ldots\cdot10^{-6}$
$-8+15\mathrm{i}$	$1.14415\ldots\cdot10^{-6}$	$0.5+2\mathrm{i}$	$5.69890\ldots\cdot10^{-6}$
$-6+15\mathrm{i}$	$3.15187\ldots\cdot10^{-7}$	$0.5+5\mathrm{i}$	$3.32571\ldots\cdot10^{-4}$
$-4+15\mathrm{i}$	$6.71436\ldots\cdot10^{-6}$	$0.5+10\mathrm{i}$	$2.14447\ldots\cdot10^{-4}$
$-2+15\mathrm{i}$	$1.20440\ldots\cdot10^{-4}$	$1+2\mathrm{i}$	$5.55222\ldots\cdot10^{-6}$
$-1+2\mathrm{i}$	$4.56006\ldots\cdot10^{-6}$	$1+5\mathrm{i}$	$6.03732\ldots\cdot10^{-5}$
$-1+10\mathrm{i}$	$8.22455\ldots\cdot10^{-5}$	$1+10\mathrm{i}$	$1.84408\ldots\cdot10^{-4}$
0	$4.02877\ldots\cdot10^{-6}$	$2+10\mathrm{i}$	$8.22455\ldots\cdot10^{-5}$
2i	$5.55222\ldots\cdot10^{-6}$	$3+10\mathrm{i}$	$3.13852\ldots\cdot10^{-5}$
5i	$6.03732\ldots\cdot10^{-5}$	$5+10\mathrm{i}$	$4.51426\ldots\cdot10^{-6}$

9 Ramanujan Tau L-Function

Within this section we presuppose that in (15)

$$c = 12 \qquad \text{and} \qquad h(s) = (2\pi)^{-s}\,\Gamma(s). \tag{71}$$

We have two methods for "solving" a functional equation—via replicas of the equation itself for particular values of s (as in Sects. 4–6), and via vanishing of the

Table 10 Coefficients of $D_{30}(s)$ defined by (24), (29) (65) (for $k = 1,\ 3,\ \ldots,\ 59$), and (66)

n	$a_{30,n}$	n	$a_{30,n}$
1	1.00000000...	16	19241.29315524...
2	0.28413934...	17	−29407.07560910...
3	−0.28447472...	18	38570.85113607...
4	−0.99771570...	19	−43253.85469735...
5	−0.01426839...	20	41265.28452795...
6	1.09206874...	21	−33287.54237140...
7	−0.28274138...	22	22535.10552513...
8	2.87847307...	23	−12681.87163010...
9	−16.56795005...	24	5858.25583683...
10	66.74261053...	25	−2182.96010798...
11	−246.76040187...	26	639.95094404...
12	794.83008053...	27	−142.11869475...
13	−2199.87496770...	28	22.47824068...
14	5254.16299598...	29	−2.25669057...
15	−10831.19871227...	30	0.10811767...

odd derivatives at one point (as in Sects. 7–8)). In this paper we will use the latter way (the former one was used in [19]); of course, one can combine equations of both types, (30) and (65), in one system .

9.1 Numerical Data

To begin with we define coefficients $a_{N,n}, \ldots, a_{N,N}$ of $D_N(s)$ by (29) and (65) for $k = 1,\ 3,\ \ldots,\ 2N-1$.

Table 11 shows corresponding values of $a_{N,2}, \ldots, a_{N,7}$ for $N = 50,\ \ldots,\ 250$. It does not looks like that the coefficients approach some limiting values. More likely, they behave as partial sums of an asymptotic series—at first approaching "correct" value, but then retreating it.

The values of $a_{N,2}$, especially $a_{100,2}$, are very close to an integer, so we can make a guess that

$$a_2 = -24. \tag{72}$$

Similar but less confident guesses could be made about the values of $a_{N,3}$, $a_{N,4}$, $a_{N,5}$, and $a_{N,6}$. But already for $a_{N,7}$ the data from the table are not sufficient in order to make choice between -16744 and -16745.

At the moment we make only commitment (72), that is, from now on we assume not only (29) but

$$a_{N,2} = -24 \tag{73}$$

Table 11 Initial coefficients of $D_N(s)$ defined by (24), (29) (65) (for $k = 1,\ 3,\ \ldots,\ 2N-1$), and (71)

n	N	$a_{N,n}$
2	50	−24.000000000118497...
	100	−23.999999999999942...
	150	−23.999999999999770...
	200	−23.999999998866933...
	250	−24.000199961334035...
3	50	252.000000057374527...
	100	251.999999999961931...
	150	251.999999999836542...
	200	251.999999165844212...
	250	252.149430632741081...
4	50	−1472.000012515395811...
	100	−1471.999999986251471...
	150	−1471.999999931279260...
	200	−1471.999626780797076...
	250	−1540.912343773167466...
5	50	4830.001582240256756...
	100	4829.999996590998579...
	150	4829.999978579614166...
	200	4829.871536668917347...
	250	29755.868246403074758...
6	50	-6048.129472974338049...
	100	−6047.999374391124392...
	150	-6047.994675315311927...
	200	−6011.297392336898792...
	250	−7657439.816197617182839...
7	50	−16736.650298606985052...
	100	−16744.088289724678448...
	150	−16745.089444954449710...
	200	−25731.482054790443951...
	250	2061626557.103562626814415...

as well; respectively, we reduce the number of other equations by 1, that is, we proceed with the system (65) for $k = 1,\ 3,\ \ldots,\ 2N-3$.

Table 12 shows values of $a_{N,3},\ \ldots,\ a_{N,8}$ recalculated under the two assumptions, (29) and (73). We get greater confidence that

$$a_3 = 252 \tag{74}$$

and from now on we assume also that

$$a_{N,3} = 252. \tag{75}$$

Table 12 Initial coefficients of $D_N(s)$ defined by (24), (29) (65) (for $k = 1,\ 3,\ \ldots,\ 2N-3$), (71), and (73)

n	N	$a_{N,n}$
3	50	252.000000001276967...
	100	251.999999999999973...
	150	252.000000000000027...
	200	251.999999999845893...
	250	252.000030423879946...
4	50	−1472.000000679159479...
	100	−1471.999999999973598...
	150	−1472.000000000034434...
	200	−1471.999999792994376...
	250	−1472.042583958137485...
5	50	4830.000145096068699...
	100	4829.999999987556299...
	150	4830.000000022057612...
	200	4829.999851805405458...
	250	4862.471493291169755...
6	50	−6048.016996351377522...
	100	−6047.999996384861074...
	150	−6048.000009448467037...
	200	−6047.926302953560375...
	250	−23666.925084998437692...
7	50	−16742.744014527296735...
	100	−16744.000722228039628...
	150	−16743.997009796162146...
	200	−16772.110633412127340...
	250	7502100.474170648682726...
8	50	84416.317314370117715...
	100	84480.105533105112501...
	150	84479.263918168986183...
	200	93122.012655507139459...
	250	−2651916509.556645449374102...

Further recalculation (see Table 13) performed under the three assumptions, (29), (73) and (75), suggests that

$$a_4 = -1472 \tag{76}$$

and from now on we assume that

$$a_{N,4} = -1472. \tag{77}$$

The next recalculation with this additional assumption (see Table 14) allows us to guess that

$$a_5 = 4830 \qquad \text{and} \qquad a_6 = -6048. \tag{78}$$

Table 13 Initial coefficients of $D_N(s)$ defined by (24), (29) (65) (for $k = 1, 3, \ldots, 2N-5$), (71), (73), and (75)

n	N	$a_{N,n}$
4	50	−1472.000000014705761...
	100	−1472.000000000000225...
	150	−1472.000000000000278...
	200	−1471.999999999999752...
	250	−1472.000010741048018...
5	50	4830.000006623631354...
	100	4830.000000000255668...
	150	4830.000000000433781...
	200	4829.999999999870874...
	250	4830.020863090813102...
6	50	−6048.001194719567767...
	100	−6048.000000130026647...
	150	−6048.000000326036551...
	200	−6048.000000262414081...
	250	−6068.908783173152161...
7	50	−16743.881211654650012...
	100	−16743.999960465762010...
	150	−16743.999844150211253...
	200	−16743.999559817499552...
	250	−2503.262375742032465...
8	50	84472.490032773918378...
	100	84479.991888819148223...
	150	84479.947056235365133...
	200	84479.656617018282086...
	250	−7254372.111906719899883...
9	50	−113314.578115188801735...
	100	-113641.796404697915159...
	150	−113629.420756134702361...
	200	−113465.380053378057042...
	250	3025890243.971514185540493...

The On-Line Encyclopedia of Integer Sequences [26] recognizes (28), (72), (74), (76), and (78) as the beginning of Sequence A000594 of *tau numbers of Ramanujan*, usually denoted as $\tau(n)$. They can be defined in many ways, in particular, via the formal expansion

$$q\prod_{n=1}^{\infty}(1-q^n)^{24}=\sum_{n=1}^{\infty}\tau(n)q^n. \tag{79}$$

Values $\tau_7 = -17644$ and $\tau_8 = 84480$ are in a sufficiently good agreement with Table 14.

Table 14 Initial coefficients of $D_N(s)$ defined by (24), (29) (65) (for $k = 1,\ 3,\ \ldots, 2N-7$), (71), (73), (75), and (77)

n	N	$a_{N,n}$
5	50	4830.000000079105807...
	100	4830.000000000004008...
	150	4830.000000000001438...
	200	4830.000000000316194...
	250	4830.000000161564014...
6	50	−6048.000027589401480...
	100	−6048.000000004386464...
	150	−6048.000000002441836...
	200	−6048.000000665629448...
	250	−6048.000349947935199...
7	50	−16743.996092066542882...
	100	−16743.999997890350852...
	150	−16743.999998065045899...
	200	−16743.999318063485484...
	250	−16743.633330978509881...
8	50	84479.691889532967686...
	100	84479.999400486350227...
	150	84479.999048286456341...
	200	84479.549703982396722...
	250	84236.978331976094194...
9	50	−113627.478816657760268...
	100	−113642.885788007170829...
	150	−113642.673364788701395...
	200	−113428.463299544266846...
	250	−2200.273147271369263...
10	50	−116460.643498323631911...
	100	−115935.685178325567478...
	150	−116003.489243160027771...
	200	−194313.225722385114160...
	250	−35879210.651607157671389...

The Dirichlet generating function for the tau numbers,

$$L_\tau(s) = \sum_{n=1}^{\infty} \tau_n n^{-s}, \tag{80}$$

is called *Ramanujan tau L-function*. It indeed satisfies the functional equation (15) for parameters (71) as it was shown by J. R. Wilton in [29].

References

1. Balanzario EP (2000) Remark on Dirichlet series satisfying functional equations. Divulg Mat 8(2):169–175
2. Balanzario EP, Sánchez-Ortiz J (2007) Zeros of the Davenport–Heilbronn counterexample. Math Comput 76(260):2045–2049
3. Beliakov G, Matiyasevich Yu (2015) Approximation of Riemann's zeta function by finite Dirichlet series: a multiprecision numerical approach. Exp Math 24(2):150–161. (See also http://arxiv.org/abs/1402.5295)
4. Blagouchine IV (2014) Rediscovery of Malmsten's integrals, their evaluation by contour integration methods and some related results. Ramanujan J 35:21–110; Addendum: Ibid, 42:777–781, 2017
5. Bohr HA, Mollerup J (1922) Lærebog i matematisk analyse af Harald Bohr og Johannes Mollerup, vol III. J. Gjellerups, Copenhagen. https://books.google.ru/books?id=RIpVAAAAYAAJ
6. Bombieri E, Gosh A (2011) On the Davenport–Heilbronn function. Uspekhi Mat Nauk 66:15–66. Translated in: Russian Mathematical Surveys, 66:221–270, 2011
7. Borwein P (2000) An efficient algorithm for the Riemann zeta function. In: Constructive, experimental, and nonlinear analysis (Limoges, 1999). CRC mathematical modelling series, vol 27. CRC, Boca Raton, pp 29–34
8. Euler L (1768) Remarques sur un beau rapport entre les series des puissances tant directes que reciproques. Memoires de l'Academie des sciences de Berlin 17:83–106. Reprinted in Opera omnia. Series prima: Opera mathematica. Vol. XV: Commentationes analyticae ad theoriam seriarum infinitarum pertinentes, G. Faber, ed., pp. 70–90, Leipzig, B. G. Teubner (1911,1980). http://eulerarchive.maa.org/pages/E352.html
9. Farmer DW, Ryan NC (2014) Evaluating L-functions with few known coefficients. LMS J Comput Math 17:245–258. arXiv:1211.4181
10. Farmer DW, Koutsoliotas S, Lemurell S (2014) Maass forms on GL(3) and GL(4). Int Math Res Not 2014(22):6276–6301
11. Farmer DW, Koutsoliotas S, Lemurell S (2015) Varieties via their L-functions. arXive 1502.00850
12. Hamburger H (1921) Über die Riemannsche Funktionalgleichung der ζ-Funktion (Zweite Mitteilung). Math Z 11:224–245
13. Hamburger H (1921) Über die Riemannsche Funktionalgleichung der ζ-Funk-tion (Erste Mitteilung). Math Z 10:240–254
14. Kaczorowski J, Molteni G, Perelli A (2010) A converse theorem for Dirichlet L-functions. Comment Math Helv 85(2):463–483
15. Kinkelin H (1858) Ueber einige unendliche Reihen. Mitteilungen der Naturforschenden Gesellschaft in Bern 419–420:89–104. https://books.google.de/books?id=N3cWAQAAIAAJ&pg=PA89#v=onepage&q&f=false
16. Malmstén CJ (1849) De integralibus quibusdam definitis seriebusque infinitis. J Reine Angew Math 38:1–39
17. Matiyasevich Yu (2012) New conjectures about zeros of Riemann's zeta function. Research Report of the Department of Mathematics of University of Leicester, MA12-03, 44 pp. http://www2.le.ac.uk/departments/mathematics/research/research-reports-2/reports-2012/ma12-03, https://logic.pdmi.ras.ru/~yumat/talks/leicester_2012/MA12_03Matiyasevich.pdf
18. Matiyasevich Yu (2013) Calculation of Riemann's zeta function via interpolating determinants. Preprint of Max Planck Institute for Mathematics in Bonn, 18, 31 pp. http://www.mpim-bonn.mpg.de/preblob/5368, https://logic.pdmi.ras.ru/~yumat/talks/bonn_2013/5368.pdf
19. Matiyasevich Yu (2018) Computational rediscovery of Ramanujan's tau numbers. Integers 18A:1–8. http://math.colgate.edu/~integers/vol18a.html
20. Matiyasevich Yu WWW personal journal. https://logic.pdmi.ras.ru/~yumat/personaljournal/finitedirichlet

21. Matiyasevich Yu (2018) Computational aspects of Hamburger's theorem. Preprint POMI 18-01, 31pp. http://www.pdmi.ras.ru/preprint/2018/18-01.html
22. Perelli A (2017) Converse theorems: from the Riemann zeta function to the Selberg class. Bollettino dell'Unione Matematica Italiana 10(1):29–53. (see also ArXiv 1605.02354)
23. Riemann B (1859) Über die Anzhal der Primzahlen unter einer gegebenen Grösse. Monatsberichter der Berliner Akademie. Included into: Riemann, B. *Gesammelte Werke*. Teubner, Leipzig, 1892; reprinted by Dover Books, New York, 1953. http://www.claymath.org/publications/riemanns-1859-manuscript, English translation: http://www.maths.tcd.ie/pub/HistMath/People/Riemann/Zeta/EZeta.pdf
24. Rubinstein M (2005) Computational methods and experiments in analytic number theory. In: Recent perspectives in random matrix theory and number theory. Proceedings of a school that was part of the programme 'Random matrix approaches in number theory', Cambridge, UK, January 26–July 16, 2004. Cambridge University Press, Cambridge, pp 425–506
25. Selberg A (1992) Old and new conjectures and results about a class of Dirichlet series. In Proceedings of the Amalfi conference on analytic number theory (Maiori, 1989). Univ. Salerno, Salerno, pp 367–385. Reprinted in *Collected papers*, Vol. II, Springer-Verlag, 1991 and 2014
26. Sloane NJA (ed) The on-line encyclopedia of integer sequences. https://oeis.org/A000594
27. Titchmarsh EC (1986) The theory of the Riemann zeta-function, 2nd edn. The Clarendon Press, New York
28. Turing AM (1953) Some calculations of the Riemann zeta-function. Proc Lond Math Soc 3:99–117. Reprinted in: *Collected Works of A. M. Turing: Pure Mathematics* (J. L. Britton, ed.), North-Holland, Amsterdam, (1992); *Alan Turing – His Work and Impact*, S. B. Cooper, J. van Leeuwen, eds., Elsevier Science, 2013. ISBN: 978-0-12-386980-7
29. Wilton JR (1927) A note on Ramanujan's arithmetical function $\tau(n)$. Math Proc Camb Philos Soc 23(6):675–680
30. The inverse symbolic calculator. https://isc.carma.newcastle.edu.au/standard
31. Wolframalpha. https://www.wolframalpha.com

Effective Validity: A Generalized Logic for Stable Approximate Inference

Robert H. C. Moir

Abstract The traditional approach in philosophy of using logic to reconstruct scientific theories and methods operates by presenting or representing a scientific theory or method in a specialized formal language. The logic of such languages is deductive, which makes this approach effective for those aspects of science that use deductive methods or for which deductive inference provides a good idealization. Many theories and methods in science, however, use non-deductive forms of approximation. Approximate inferences, which produce approximately correct conclusions and do so only under restricted conditions before becoming unreliable, behave in a fundamentally different way. In the interest of developing accurate models of the structure of inference methods in scientific practice, the focus of this paper, we need conceptual tools that can faithfully represent the structure and behaviour of inference in scientific practice. To this end I propose a generalization of the traditional notion of logical validity, called *effective validity*, that captures the form of approximate inferences typically used in applied mathematics and computational science. I provide simple examples of approximate inference in mathematical modeling to show how a logic based on effectively valid inference can directly, faithfully represent a wide variety of the forms of inference used in scientific practice. I conclude by discussing how such a generalized logic of scientific inference can provide a richer understanding of problem-solving and mathematical modeling processes.

1 Introduction

There is a long tradition in philosophy of science of using logical tools and methods to gain insight into scientific theories—their structure, concepts, methods and their ability to represent the world. Classical formal logic in particular, specifically

R. H. C. Moir (✉)
Department of Computer Science, The University of Western Ontario, London, ON, Canada
e-mail: robert@moir.net

© Springer Science+Business Media, LLC, part of Springer Nature 2019

N. Fillion et al. (eds.), *Algorithms and Complexity in Mathematics, Epistemology, and Science*, Fields Institute Communications 82,
https://doi.org/10.1007/978-1-4939-9051-1_9

classical first order logic (FOL) paired with set theory (ST), has been enormously influential in philosophy as a model for correct inference in mathematics, science and inference in general. It informs scores of traditional, though still very influential, views in the philosophy of science. Notable examples of this influence are: the Oppenheim-Putnam view of the logical structure of science [14], which has strongly influenced views of reduction and inter-theoretic relations in science; the Hempelian view of scientific explanation [9], which has deeply influenced views of scientific explanation and continues to be influential in its own right; as well as myriad strategies of rational reconstruction of scientific theories, from the early proof-theoretic approaches of the logical empiricists (see Suppe [15] for a good historical discussion of these views) to the model-theoretic approaches of Suppes [16, 17], van Fraassen [19, 20] and others.

A common feature of these uses of formal logic as presentations or representations of science or aspects of science, is a projection of actual scientific theories, concepts and methods into a logical framework, typically classical FOL, so that the products of science are presented in a uniform language. The benefit of such a projection is that imprecise concepts and methods can be reformulated in a precise language, clarifying their structure, content and justification, or lack thereof as the case may be, leading to valuable insights into science. A nice example of how a logical analysis can provide useful insights is Popper's doctrine of falsificationism, contributing to the demarcation problem (between science and pseudoscience) and providing a useful heuristic for many practising scientists who are engaged in the development of testable new theories and models.

This strategy of projecting science into logic also has its limitations. A major limitation is that the logic used is invariably strictly deductive, which is not a problem in itself except for the fact that many theories, concepts and methods in scientific practice use non-strictly-deductive forms of approximation. Not only are approximation methods used in practice, but in most cases in which they appear, they are essential for making problem solving feasible in the restricted inferential contexts of scientific practice; without approximation methods, much of science would be impossible in practice [13]. Thus, the actual processes of description, prediction, explanation and control in science often use approximation, as many philosophers of science, including Wimsatt [24], Harper [8], Batterman [1],Wilson [23], now emphasize.

The basic reason that deductive inference is inadequate to fully represent approximate inference is that it distorts the basic properties of the inferences. Deductive inferences establish sharp and certain relationships between sentences and their consequences, whereas for approximate inferences the relationships are not as sharp and are less certain. As an example, consider the model of a simple pendulum (we consider this example in detail in Sect. 5 below). For small oscillations this model behaves approximately like a simple harmonic oscillator. Thus, it is common to infer that for small oscillations the motion of a simple pendulum is a sinusoid. To represent this as a deductive inference, one may use a conditional of the form:

$$\text{simple pendulum equations} + \text{small oscillations} \Rightarrow \text{motion is sinusoidal} \qquad (1)$$

Setting aside issues with the use of a material conditional here, this sentence does not reflect the fact that what counts as a small oscillation is imprecise (imprecision in the premises), that a sinusoid approximates the motion of the pendulum (imprecision in the consequence), and that the inference is only valid under a certain range of initial conditions of the pendulum before the distortion of the motion away from sinusoidal becomes significant, which is itself an imprecise boundary (this range can be regarded as implicitly defining the needed sense of 'small' here). Thus, the basic character of the inference as one where the conclusion is only approximately correct, and such that this approximate correctness obtains only under restricted conditions before becoming unstable. This is not represented by a deductive conditional.

A defender of hypothetico-deductivism could reply that deductive inference can easily recover the imprecision in the premises and conclusion, as well as the fact that the inference is only reliable under certain conditions, by developing a more sophisticated model of the reasoning involved. For instance, reflecting what is typically done in practice, we could add in an approximation scheme with a certain error tolerance to be able to say that

$$\text{equations} + \text{small oscillations} + \text{approximation scheme} \Rightarrow \text{approximately sinusoidal motion}, \quad (2)$$

and then capture the conditions on the reliability with a statement such as

$$\text{approximate prediction is good} \Leftrightarrow \text{errors are within the tolerance}. \quad (3)$$

When the antecedents of (2) obtain we are then in a position to deduce that the pendulum's motion will be approximately sinusoidal, and, according to (3), that when the errors are within the tolerance this is a good approximation. This move does indeed represent the approximate correctness and conditional stability of the inference using deductive logic. However, it is crucial that it does so by adding in other elements (approximation scheme, definition of good approximation) to make the conditions sufficiently precise to be stated in logical form. This may indeed be valuable, and deductive logic useful for representing approximate inference in similar sorts of ways, but it does not change the fact that we have distorted the basic character of the approximate inference to be able to represent it formally.

To clarify the nature of the distortion involved, notice that the original inference concerning the sinusoidal form of the motion of a simple pendulum for small oscillations (that we tried to represent directly as (1)) contains an inherently indeterminate sense of what counts as a good approximation. All the approximate inference means to express is that the motion of a simple pendulum resembles a sinusoid under certain conditions, namely where the oscillations are small, where what counts as 'resembling' and 'small' is left open. By adding conditions to represent the approximate character of the inference precisely we had to close this openness by choosing a particular way of measuring error and defining what counts as a good approximation. This effects a sharp boundary on what counts as a good approximation, which is not present in the original inference, indicating how the choice to represent the inference in deductive logic distorts the basic character of

the inference by changing its content. It may be a small distortion, and a good and even useful model of the inference, but a distortion nonetheless.[1] As such, the use of deductive logic to represent approximation is itself an approximate representation of the original inference.

But there is a subtler form of distortion here that gets to the heart of the matter. One may see the purpose of this paper as an attempt to demonstrate that approximate inferences are really a different kind of inference than traditional deductive inferences. Rather than seeing the approximate correctness and conditional stability of approximate inferences as something that we add on to deductive inferences, the purpose of this paper is to examine the basic properties of inferences that are inherently approximately correct and conditionally stable. We may then see that deductive inference falls out as a special case of this more general form of reasoning. I argue that this approach not only leads to models of scientific inference that better reflect the forms of reasoning we find in scientific practice, but also to new ways of thinking about scientific reasoning and its products. This argument can only be made by developing an outline of the basic logical features of approximate reasoning and showing how this leads to new ways of viewing science, a task I aim to initiate in this paper.

An important part of the motivation for developing a generalized notion of logical validity is that the general problem I am concerned with is an elucidation of the structure and behaviour of inference as it is observed in scientific practice. Beyond reconstructing episodes of scientific inference the task here is to identify structural and behavioural patterns common to scientific methods and knowledge systems over greater or lesser portions of science. Consequently, the aim is partly philosophical, being interested in the basic structure of theories and inference in science, but also scientific, being concerned with clarifying common structural features of scientific reasoning processes that on the face are enormously diverse. This is rather a different task than that of traditional reconstructions of science and, I argue, requires different methods and strategies as a result. In keeping with this scientific aim we consider a variety of simple (from a scientific perspective) examples of actual scientific reasoning to illustrate the applicability of the concepts we develop, but also to show how they elucidate their basic structure. It will then become evident that the new concepts introduced both accurately describe actual methods and provide fresh perspectives on the nature of scientific reasoning. The scientific and philosophical aims of the paper are therefore complementary.

[1] It should be noted that applied mathematicians commonly introduce approximation schemes and (operational) definitions of good approximations to develop precise conditions under which methods will be reliable. I will draw in part from such approaches in the sections to follow. It is to be emphasized, however, that my approach differs from the approach of the hypothetico-deductivist in that I will be using a generalized valid inference concept that internalizes error and stability. Such inferences, with the *effective equivalence relation* $\sim$ (defined on page 246) left more or less generic, could preserve a sense of openness in the sense of approximation and stability of the original inference. This provides a simple illustration of how generalizing validity can provide direct and more faithful representations of scientific inference.

2 The Concept of *Effective Validity*

At the core of my extended argument in this paper is the idea that we can more faithfully represent approximate inference, and the structure and behaviour of inference in scientific practice as a result, by viewing stable approximate inferences as inferences that are *valid* in a sense more general than that of traditional deductive logic. I call the resulting concept *effective validity*, where the sense of "effective" is that from physics and not computability theory, connoting a capturing of much of the form or functional behaviour of valid inference while producing a concept aware of error and approximation. Although one of the main motivations for this generalization of deductive logic, which I call *effective logic*, is to account for approximate inference, the fundamental concept is not actually approximation. When we reflect on what we require of inferences when approximation is introduced, we may see that we require the inference we wish to make to be *stable* under the *variation* implied by the given kind of approximation. Thus, effective logic is fundamentally about the stability of inferences under different kinds of variation.

Stability is a key property of successful mathematical descriptions of nature. For a mathematical model to successfully describe a phenomenon it must be the case that the description it provides is stable under small changes to the model. If this were not so, then a small change could produce an entirely different description, resulting in a model that no longer describes the phenomenon. A general reason why models capable of describing the world must have this property is that there are always forms of error in the modeling process, so any description we can produce involves error and approximation in relation to the phenomenon we seek to represent.

This idea of stability of mathematical descriptions underlies the technical notion of well-posedness introduced by Hadamard. A problem involving a differential equation is considered to be *well-posed* if there exists a unique solution with the property that the solution varies continuously with small changes to the initial and/or boundary conditions. This continuous variation property is what guarantees that the description (solution) is stable in a mathematical sense, and consequently guarantees its stability as a description of any phenomenon described by a given initial value or boundary value problem.

We may observe that the standard deductive notion of valid inference on its own tells us nothing about whether other nearby inferences, obtained by certain variations of the premises for example, are also valid. Of course if a nearby inference is also an instance of a valid inference form then we know it is also valid, but for inferences in general deductive validity tells us nothing about preservation of validity under changes to propositions. Working in any context in which error and approximation are involved, if our inferences are to be reliable it must be the case, for Hadamard-type reasons, that the inference continues to be valid if the premises are only close to being true. This observation is what underlies the informal definition of effective validity: an inference is *effectively valid* if whenever the

premises are nearly-true the conclusion is nearly-true.[2] Thus, rather than preserving truth, as do valid inferences, effectively valid inferences preserve "near-truth", meaning that nearby inferences continue to be valid.

This informal notion of near-truth is not one that we will work with directly, for similar reasons that the informal notion of truth is avoided in formal logic. Much as for validity in standard formal logic, this informal notion of effective validity can be represented symbolically in a number of different ways. Accordingly, in the sections that follow, we will consider a number of precise formulations of effective validity in terms of stability of inferences under variations in syntax and semantics of sentences. It is important to note, however, that the aim here is not a formal system mirroring those in formal logic. Rather the approach here is to develop precise concepts with the assistance of symbolic notation that accurately model or represent the form of reasoning observed in scientific practice. We are therefore treating inference in scientific practice as a phenomenon to be modeled using the concepts of effective logic, much as physical phenomena are modeled using the concepts of physics and mathematics. As a result, we are engaged in a process of investigation of scientific methods to develop new concepts that accurately capture the structure and behaviour of the inferences involved. I will motivate this approach by showing how this generalization of valid inference in effective logic captures the form of approximation methods typically used in applied mathematics, though I will suggest that it can capture the form of a much wider range of scientific inference.

Effective logic is a generalization of deductive logic as usually conceived in two different ways. First of all, when the variation is reduced to zero, the concept of effective validity reduces to a deductive form of validity. Thus, in this restricted sense, deductive logic is obtained in an appropriate limit of this generalized logic.[3] This is the sense in which effective logic really is a generalization of deductive logic. It is also a generalization in the sense of an expansion in terms of how a logical representation can elucidate the structure and content of science. Rather than working only with uninterpreted languages and their models, as is standard in the metatheoretical treatment of formal logics, effective logic also makes important use of interpreted languages and mappings between them, since this is standard in much of scientific practice, but particularly in applied mathematics and computational science. Indeed, the focus of this paper is primarily on the effective logic of such interpreted languages. We will see how such an approach can provide a faithful representation of inferential structure in scientific practice, helping to account for how and why scientists develop and employ the methods they do.

[2] I avoid the term "approximate truth" here both, because it is a notoriously problematic notion in philosophy and to provide a clearer link to the more precise formulations of effective validity to follow.

[3] Note that this limit is singular in the sense of perturbation theory, because deductive inference (with no internal variation or conditional stability) is qualitatively different.

The aim of effective logic is not only to provide a more refined tool for capturing the structure of scientific inference. It is hoped that it will be useful for many of the purposes for which formal logic has been used in philosophy of science for more than a century. This includes the elucidation of the structure of theories, clarification of scientific concepts and methods, clarification of the processes of description, prediction, explanation and control, as well as accounting for scientific representation. The idea of effective logic, then, is that with a greater ability to make contact with actual scientific reasoning, a logical representation can provide finer grained philosophical insights into scientific practice that are directly applicable to science itself.

Before we develop the precise formulations of the concepts of effective logic, we must first consider the stability of symbolic expressions and mathematical structures under forms of variation. This is the subject of Sect. 3, which examines the stability of syntactic and semantic structures in terms of near-identity transformations. We then move in Sect. 4 into the territory of logic proper by considering interpreted languages, which have truth conditions for propositions, and their stability properties under variation of syntactic or semantic structure. At this point we can discuss in Sect. 5 the nature of effectively valid inferences within a mathematical framework, which we illustrate through the example of the analytic solution of the simple pendulum problem. We then consider in Sect. 6 the effective logic of problem-solving strategies that involve transformations between interpreted languages, taking as an example the numerical integration of the double pendulum problem.

As an indication of the kind of insight that effective logic can make accessible, we will conclude in Sect. 7 by discussing the concept of *inferential structure* identified by effective logic and its implications in particular for computational science, i.e., the science of algorithmic solution to mathematical problems, and our understanding of mathematical modeling processes. It is shown in [13] how a strategy of stable transformation of computational problems underlies strategies of computational complexity reduction in computational science. Notable examples of this sort of strategy are on the one hand numerical methods, which transform difficult continuous problems into rapidly computable ones using forms of discretization, and on the other modular methods, which transform difficult symbolic problems into many smaller problems that can be rapidly computed within a single machine word, i.e., computed effectively within a single processor cycle. Effective logic then shows that a basic requirement of these strategies is the near-preservation of inferential structure. It is discussed how mathematical modeling strategies fulfill this same requirement, so that models can be seen as tools for reducing *inferential complexity*. Although a detailed argument is beyond the scope of this paper, I refer to additional evidence presented elsewhere to suggest there is reason to suspect that near-preservation of inferential structure is a basic requirement for reliable methods very broadly in scientific practice.

3 Stability Under Syntactic and Semantic Variation

The shift from deductive validity to effective validity is primarily about introducing a context of variation for the syntax and semantics of sentences together with a consideration of stability of consequence relations under such variations. In this section we will focus on the general kinds of variation that effective logic involves and what such variation looks like for syntactic forms and semantic structures. In particular we will distinguish the kind of variation particular to approximation from other, weaker forms of variation, which will clarify the structure of inferences involving approximation.

Formal logic deals with two precise kinds of validity, corresponding to two kinds of consequence. The first is syntactic validity, which corresponds to deductive consequence. In this case, for a set Γ of assumptions and a sentence p in some formal language, in a formal system based on that language, $\Gamma \vdash p$ denotes that the inference from Γ to p is valid because there is a proof of p in the formal system with members of Γ as assumptions. The second is semantic validity, which corresponds to semantic consequence, often called logical consequence. In this case, $\Gamma \models p$ denotes that the inference from Γ to p is valid because in any model of the formal language in which each member of Γ is true, p will also be true. Since it is these precise forms of validity that formal logic works with, it is these concepts, or analogues of them, we seek to generalize to inferences in a context of variation, and approximate inferences in particular.

One of the fundamental reasons for considering approximation in inferences is to be able to draw conclusions when certain amounts of error are unavoidable in assumptions and acceptable in conclusions. Typically when error is introduced the idea is that the inferences can continue to go through provided the size of the error is "small". This sense of smallness is not always precise or even clear, but indicates that some measure of size or distance is needed to quantify it. In the general context of effective logic, however, we seek a way of capturing the sense of "smallness" of a variation without the need of such a measure. This will come down in any given context to "smearing" an object into a (typically well-defined) range of effectively equivalent objects, leading to similar ranges of propositions containing such objects, and then to inferential relations between such "smeared out" propositions.[4] For the moment, however, we will analyze the notion of smallness in terms of the allied notion of nearness, and in particular *near-identity*.

In the interest of distilling as much as possible the concept of error to its essence, we may observe that this means that, for the purposes of a stable approximate inference, approximately identical assumptions should lead to approximately identical conclusions. Thinking in terms of meaningful assertions, this means that if we make almost the same assertions, then almost the same conclusions

[4]This is analogous to the mathematical notion of a neighbourhood from topology, but the collections of "smeared out" propositions need not have any topological structure in general.

should be assertible. Thinking in terms of linguistic assumptions, this means that if we make almost the same assumptions, then almost the same consequences should be provable. We may see from this that the stability of near-identity, in the sense of nearly-identical input producing nearly-identical output, captures the basic relation between assertions/assumptions and conclusions/consequences that effective validity aims to make precise.

Before we consider the notion of near-identity more precisely, there are two important basic features to appreciate concerning this concept of stable approximate inference that make it different from deductive inference. The first of these is that we have replaced a concept of exact consequence with one that requires that a special relation of error obtain between input (assertion/assumption) and output (conclusion/consequence). Naturally enough, the stability of this relation will not always obtain. Thus, approximate inferences run the risk of becoming *unstable*. There are a number of ways in which approximate inferences can become unstable, including: small changes to premises lead to large changes in conclusions (analogous to chaotic behaviour in dynamical systems); the premises themselves become unstable (analogous to decay or decoherence of prepared states in quantum mechanics)[5]; or chains of approximate inferences interact in such a way as to no longer be stable (analogous to slippery slope fallacies in probabilistic reasoning).

To be in a position to handle potential instabilities of inferences requires an understanding of the stability of consequence relations. In the context of scientific practice, this can often be handled in terms of support theorems that establish certain kinds of variation over which consequence relations will be stable, such as numerical stability theorems in numerical analysis. This can provide a means of judging when chains of approximate inferences are stable as well. In general, however, stability is judged by some external means, such as agreement with experiment or observation, when proofs of stability are not available, which is typical when modeling complex phenomena such as climate change. In such cases, it is the correctness of the expected consequences over a range of variation that establishes, or provides evidence of, stability externally. In the context of this paper we are concerned only with outlining the structure of approximate reasoning. Though it is of crucial importance, incorporating evidence or proofs of stability into effective logic is a subject for future work.[6]

The second basic feature of approximate inference is something already alluded to, which is that introducing variation forces a move from considering consequence relations between individual sentences to consequence relations between *ranges* of nearly-identical sentences. Thus, in shifting from deductive to approximate

[5]This case covers statements whose truth depends on the location in a state space, which includes statements whose truth can be inherently variable over time. Since the local truth of a statement can be expressed in terms of its effective validity with no premises or assumptions, this is a special case of the notion of effective validity.

[6]When this task is taken up, the initial aim will be to identify the forms of justification of stability that are observed in practice. It is an open question whether this could lead to the ability to develop methods for prediction of the failure of inference methods.

inference we shift to consideration of inferences as correspondences between collections of sentences defined by relations of near-identity. In different terms, this means that to handle approximate inference reliably we need to consider sentences that are nearly-identical, sufficiently close to one another that the same inference form applies to them.

To summarize at this point, we can see that an effectively valid inference will be one that maps a set of ranges of input sentences (corresponding to the set of assertions, assumptions or premises) to a particular range of output sentences (corresponding to the conclusion or consequence), and that joining together effectively valid inferences must be done with care to avoid instability. To become more clear about what this means we need to develop the notion of near-identity responsible for defining these ranges of sentences. The near-identity of sentences will be determined by the near-identity of their components, so that ultimately we need to examine the near-identity of syntactic and semantic entities.

There are two basic ways one can think about near-identity of entities. One is in terms of sets or collections of entities that are nearly-identical to each other, according to some standard; and the other is in terms of the transformations that map an entity to one that is nearly-identical to it. One may see that the near-identity transformations can be understood to *generate* the collections of nearly-identical entities by "smearing out" a given entity. Whichever way we think about ranges of nearly-identical entities, we will always suppose that the near-identity variation is about some fixed centre, corresponding to the particular entity (sentence, equation, expression, structure, object, etc.) that is being "smeared out".

A collection of near-identity transformations always includes the identity operator $\mathbb{I}$, which leaves any object fixed ($\mathbb{I}a = a = a\mathbb{I}$). For our purposes, this fixed object will be the fixed centre of the variation. The case where the identity is the only near-identity transformation corresponds to the case of traditional logic, where all the variation is turned off. For a given non-trivial kind of near-identity variation, we can consider a suitable collection of operators that are in a suitable sense "close" to the identity. In some cases such an operator can be written in a form such as $\mathbb{I} + \varepsilon T$, where T is some operator that acts on the given kind of entity and ε is a "small" parameter. When such a near-identity operator acts on an entity it produces a nearby nearly-identical entity ($(\mathbb{I}+\varepsilon T)a = a+\varepsilon T a = a+\varepsilon b$) that is only slightly different (here by εb) from the centre of variation (here a). A suitably well-defined collection of near-identity operators then generates a range of entities nearly-identical to the given entity we are varying around, i.e., the entity we are "smearing out".

To gain a sense of the variety of such transformations, we may observe that such transformations can be continuous or discrete (in the sense of classical analysis), and can exactly or approximately preserve structural features of the objects they act upon. Consider, for example, the context of continuous groups of transformations, such as the rotations of an object in space. Here the transformations are continuous in the sense that small changes of the input to the transformation yield small changes in the output. For example, for a pair nearly-identical vectors $\mathbf{x}$ and $\mathbf{x} + \boldsymbol{\varepsilon}$ in the plane, rotation by θ about the origin yields another pair of nearly-identical vectors. When such groups also form a manifold, they are called Lie groups, and

can be studied in terms of the near-identity transformations of the group. The collection of near-identity transformations determine a linear space called the Lie algebra of the group, which can be thought of as specifying all of the possible infinitesimal (near-identity) transformations when the group acts on a collection of objects. Since Lie groups naturally describe continuous symmetry operations (operations that produce another entity with identical structure), they generally yield exactly structure-preserving transformations. Thus, the Lie algebras of Lie groups give us an example of continuous, exactly structure-preserving near-identity transformations.

For a contrary case of discrete, approximately structure-preserving transformations, consider the case of stable numerical methods for differential equations. Differential equations generally specify how entities in some state space evolve over some continuous transformation. Indeed, if we know the operator that transforms the state space in accordance with the differential equation, we have solved the differential equation.[7] In general it is very hard, however, to compute the operator that solves the equation. Consequently, it is very useful to approximate this operator by considering small discrete changes that correspond to the continuous changes specified by the differential equation. This is the strategy followed by numerical methods, which consist of a collection of (difference) equations that determine a discrete near-identity state transition map on the state space. These difference equations can then be evaluated or approximated by a computer. Since the discrete operations break the structure of the differential equation, they can only produce approximate solutions to the equation. For a stable method, the smaller the discrete change (e.g., time step or interval of a regular spatial mesh), the more accurate the approximation the numerical method yields. Thus, the state transition maps of stable numerical methods give us examples of discrete, approximately structure-preserving near-identity transformations.

Leaving aside for the moment the kind of transformation involved in near-identity variation, let us turn to consider the *inferential* relation articulated by an effectively valid consequence relation. The idea is that nearly-identical inputs (premises, assertions, assumptions) yield nearly-identical outputs (conclusions, consequences). In other words, a "small" change to the input results in a "small" change to the output. This is thus very similar to the relation that a continuous transformation or map must have, where infinitesimal changes to the input result in infinitesimal changes to the output. Thus, an effectively valid consequence relation can be understood as being analogous to considering "continuous" maps between premises and their conclusions. The word 'continuous' is in double-quotes because effectively valid inferences need not be continuous in the mathematical sense, since the concepts involved may not admit a precise or unique mathematical formulation,[8]

[7]Given this, it should not be surprising that the theory of Lie groups and Lie algebras is useful in the theory of differential equations.

[8]This is so even for the sense of continuous map from general topology, despite apparent similarities, since the kind of relation specified by an effectively valid inference is intended to

but they do exhibit a generalized kind of continuity based on the nature of the near-identity relation that is preserved by the inference.

Given the analogy between effectively valid inference and continuous maps, which are stable under infinitesimal variations of the input, we can consider effectively valid inferences to be stable under "micro-local" variations of their premises. All that is meant by the phrase "micro-local variation" is a near-identity variation, which is what effective validity guarantees stability for. In such a case we fix some particular entity as centre and smear it out with near-identity transformations. Though micro-local variations have a fixed centre, in general we are also interested in the stability of inferences for "large" changes to the input (premises, assertions, assumptions, etc.). For such "large" changes, the centre of the variation itself moves. When we move the centre of variation of the premises of an effectively valid inference there is no guarantee that we move to another effectively valid inference. This is to say that generically effectively valid inference forms are only *locally stable*, i.e., they are stable only locally to a certain scope of "macro-local" variation. Thus, outside of this scope of variation, effectively valid inference forms will become unstable, in one of the ways described above.

As such, effective logic is in general a local logic, where the range of (macro)-local validity is determined by the range variations over which the inference is stable. This introduces the question of boundaries of variation of the centres of entities, or propositions containing them, over which inferences are effectively valid. I have decided not to consider such boundaries here, focusing instead on the stability relations that are central in scientific inference, though boundaries of validity must be considered in the further development of effective logic. There are a number of reasons for this decision, two of which I will mention here. One is that the stability relations are fundamental in effective logic. Introducing variation or approximation *causes* inferences to have boundaries of validity, which is why boundaries of validity are so often not known in practice. Thus, a first step is to clarify the conditions that must obtain for reasoning to be reliable; the question of when and where failure happens is a secondary consideration, albeit a matter of central concern in practice, as it will be for effective logic. Another reason to avoid considering boundaries in this paper is that it forces decisions to be made about how such boundaries are to be represented, a challenge because such boundaries are typically vague, because it is not always clear how much error is too much, or ambiguous, because boundaries depend on modeling choices or epistemic interest. This is why such boundaries are variously represented (e.g., as open sets, probability distributions, towers of sets with different confidence or error, etc.) in practice. To avoid making choices that lack generality and adding too much clutter to the definitions, we leave boundaries

be considerably more general, applying to inferences outside of mathematics. Suitably precise inferences, however, even if they are discrete in the sense of classical analysis, could still be regarded as continuous in the discrete topology.

out for the purposes of this paper.[9] It is important to keep in mind, however, that *all of the stability relations defined in this paper are generally local to certain variations*, even though this fact is not represented explicitly in the notation or definition.

In certain special cases an inference form will remain stable for all possible variations (of a given type), in which case the inference form will be called *globally stable* (relative to the type of variation). One of the natural places to look for examples of globally stable inferences is abstract algebra. Since algebraic theories pertain to a given class of structures that are defined (or definable) axiomatically, a standard way of studying a given kind of structure is to study the structure-preserving maps between structures satisfying the axioms. Such maps are called *homomorphisms*, connoting "same structure". If a homomorphism between objects is invertible, then the objects have identical or "equal" structure, and the map is called an *isomorphism*. Any inference that is valid purely in virtue of the structure of the objects appearing in the premises will continue to be valid if those objects (and corresponding ones in the conclusion) are replaced by isomorphic ones. Accordingly, the inference in question is globally stable under isomorphism transformations of the objects in the premises. From the perspective of effective logic, however, such a case is one where micro-stability expands to macro-stability, since *all* isomorphic transformations of the objects in the premises count as near-identity transformations, in which case the inference is really only a single isolated inference that maps between isomorphism classes of objects.[10] It is essentially for this reason that theorems in abstract algebra tend to focus on so-called *universal* properties, i.e., those properties that hold for all objects of a given structure or type.[11]

Examples of this phenomenon also obtain in analysis, however, as the following example shows. Consider a meromorphic function $f(z)$ on the complex plane that has a single isolated simple pole. Then consider a two distinct points c_1 and c_2 in the vicinity of the pole. It is a well-known theorem from complex analysis that contour integrals exhibit a form of path-independence. According to this, the truth of the sentence

$$\int_{\mathcal{C}} f(z)dz = c, \tag{4}$$

where $c \in \mathbb{C}$ and the endpoints of $\mathcal{C}$ are c_1 and c_2, is invariant under which $\mathcal{C}$ we choose *provided* that for any two contours $\mathcal{C}_1$ and $\mathcal{C}_2$ with endpoints fixed at

[9] It may be possible to represent boundaries in terms of a notion of an indeterminate boundary structure, but we do not explore this possibility here.

[10] Note that even with isomorphism classes we do have a notion of *near*-identity, not identity *simpliciter*, because isomorphism is always identity *relative to a type*. Saying two objects are isomorphic is only non-trivially meaningful if the two objects have some other structure that distinguishes them.

[11] Note that this extends to theorems with exceptions where the exceptions determine a substructure or subtype over which the result applies universally.

c_1 and c_2 are homotopic to each other in the sense of a continuous deformation that does not go through the pole. Thus, the truth of (4) is stable (invariant) under micro-local change of the contour, and the local stability (invariance) expands to the global within the equivalence classes defined by homotopic contours. Thus, the continuous variety (varying $\mathcal{C}$ while keeping c, c_1 and c_2 fixed) of true sentences of form (4) reduces to a discrete set of isolated propositions corresponding to the equivalence classes. Indeed, if we turn the set of isolated propositions into a group with the law of composition determined by the addition or subtraction of loops around the pole, then group of propositions is isomorphic to the fundamental group of the punctured plane.[12]

Another place to look for examples of globally stable inferences is formal logic. In the case of categorical theories, such as the second order theory of the real numbers as the order-complete totally ordered field, all of the models of the theory are isomorphic. If we consider (logical) structure-preserving transformations between models of a categorical theory, then any theorem will be globally stable under this class of transformations. On the other hand, from the perspective of effective logic, all of the models are accessible by a near-identity transformation (since all of the transformations are isomorphisms and thus strictly identity preserving), in which case there is really only a single isolated model of the theory. For a non-categorical theory, however, such as the first order theory of the real numbers as a real closed field, not all of the models of the theory are isomorphic. In this case, only certain (logical) structure-preserving transformations between models will preserve stability, i.e., truth in the model. If one considers theorems of the formal theory, they will be globally stable by definition, but other inferences that are valid in some models will only be stable under some transformations between models. Thus, non-categorical theories have statements whose truth is only locally stable.

We see from these last two examples, that where the near-identity variation is generated by isomorphisms, the ranges of nearly-identical entities tend to have sharp boundaries. Indeed, they must in some manner because being related by an isomorphism is an equivalence relation, i.e., a reflexive, symmetric and transitive relation. This shows that interpreting near-identity as isomorphism leads to a strong sense of near-identity and to inferences that are trivial from the perspective of micro-

[12]There is a clear connection here to the ideas of homotopy type theory (HoTT), which is an extension of Martin Löf type theory that adds identity types for collections of isomorphic objects, allowing isomorphic objects to be treated as formally identical. HoTT is based on Voevodsky's discovery of a model of type theory in abstract homotopy theory. See [21] for a general introduction to the foundations of the theory. The concept of identity in HoTT has a clear relation to near-identity transformations that exactly preserve the structure of sentences, and exactly preserve truth. Thus, the form of inference is still deductive, as we expect from a logical foundation for pure mathematics. The notion of identity in HoTT does not, therefore, natively capture approximate near-identity or preservation of near-truth, as is required for effective logic. The similarity of the approaches, however, makes the relationship of HoTT and effective logic an interesting subject for future research.

local stability (tend to reduce to a single isolated proposition).[13] Approximations, on the other hand, do not have such strict requirements on structure-preservation, leading to a weaker notion of near-identity transformation involving near-preservation of structure, in a contextually relevant sense of 'near'. To fix language, and to distinguish such maps from those of algebra, we could call a near-structure-preserving map a *continomorphism*, connoting "near structure".[14] We here adapt the modern Greek word *κοντινός*, the adjectival form of 'near' or 'close'. Such a map need not be invertible, much like a homomorphism need not be invertible. Thus, we can introduce the term *contisomorphism*, connoting "near equal structure", for an invertible near-structure-preserving map. It is (micro-local) transformations of this latter kind that generate the near-identity transformations we have been discussing.[15]

A key property of contisomorphisms is that as relations between entities they are reflexive and symmetric but not transitive. Thus, unlike isomorphisms, they do not give rise to equivalence classes. This is what allows inferences that are micro-locally stable under contisomorphisms about the centre of variation to become unstable under contisomorphic motions of the centre itself. The idea is that reflexivity and symmetry are necessary for micro-local stability, since the inference must hold for the centre of variation and must continue to hold for any inference accessible by a near-identity variation from the centre.[16] Transitivity, on the other hand, would imply that the inference continue to hold for any near-identity transformation applied to any inference away from the centre, which if iterated would imply global stability, which does not hold for approximate inferences in general.

To illustrate these ideas we will consider an example of near-identity variations of symbolic expressions and mathematical objects, leaving the case of inference for the next section. For a simple syntactic example, consider the case of bivariate polynomials over the real numbers,[17] e.g., $P = xy^2 + 2x - y$. Traditionally

[13] Once again, we see the connection between near-identity as isomorphism and the concept of identity types HoTT.

[14] This terminology has the nice property of being bringing to mind the word 'continue', which associates a meaning of "continuing" a structure, a natural idea of "continuous" variation. This is appropriate given that effective validity is articulating a generalized concept of continuity for inferences.

[15] Note that effective validity itself can be understood in terms of continomorphisms. Though the ranges of nearly-identical sentences must be generated by cont*iso*morphisms, the relation between the premises and conclusion is only that of a continomorphism. This is so because effective validity tells us that for any near-identical transformation of the premises, *some* near-identical conclusion will follow, but it does not say that *any* near-identical conclusion follows from a near-identical premise.

[16] For nearness relations in general, symmetry may not be required, but it is required for a notion of near-*identity*, since entities connected by such a relation must in some sense be interchangeable.

[17] The syntax here is intended to include polynomials with real numbers or real-valued parameters as coefficients, as well as meta-variables, such as P, that range over polynomials. This is in contrast to a semantics where the polynomial variables (x and y here) can take values in some domain, say a number field such as $\mathbb{R}$ or $\mathbb{C}$.

such expressions are only identical if they are mathematically equivalent, in the sense that they are interchangeable by changing the order of the terms and the order of x and y. We can introduce a syntactic form of variation centred at P by allowing small changes to the coefficients of the monomial terms, so that, e.g., each $P_\delta = (1+\delta)xy^2 + 2x - y$, $\delta > 0$, will be nearly-identical to P for $|\delta|$ sufficiently small. This is often done by introducing some *tolerance* $\varepsilon > 0$ such that for changes to the coefficients less than (or equal to) ε in size, the resulting expressions would be effectively equivalent. Thus, a single expression is expanded to a continuous range of nearly-identical expressions defined by changes of coefficients within the tolerance.

An example of a near-identity contisomorphism (NIC), or a micro-local change in expression, in this context is a map from P to an element of the set $\{P_\delta \mid |\delta| \leq \varepsilon\}$. Notice that the identity map is included in this since $P \mapsto P_0$ is included in the NICs, so being related by a NIC is a reflexive relation. It is also symmetric, since for any $0 < |\delta| \leq \varepsilon$, there is a NIC from P to P_δ, and a map from P_δ to P is included among the NICs for near-identity variation with P_δ as the centre. It is not, however, the case that being related by a NIC is transitive, since there are many NICs centred at P_δ that are not accessible to NICs centred at P, such as $P_{\delta+\varepsilon}$ if $\delta > 0$.

This shows both how approximation can be introduced in an essentially syntactic context and illustrates the non-transitivity of near-identity transformations. Though this example is continuous, there are many other kinds of near-identity transformations that are discontinuous. For example, consider nearly synonymous sentences. We could specify a range of sentences nearly-identical to a given sentence as centre based on nearly-identical meaning, say as judged by some well-trained deep learning algorithm. Once defined, the range is just a collection of sentences, thereby purely syntactic, and the NICs are just the maps from the centre sentence to the members of the range. For a given centre sentence s, it may be synonymous to s', but not synonymous to all s'' synonymous to s', so once again transitivity fails.

There is nothing about the general concept of near-identity transformation that requires the changes to be small in a sense that we might find intuitive. Keep in mind that the basic theme is stability under variation, and near-identity maps allow us to track what is judged in the context to be "nearby". Another kind of example that fits into this frame is the stability of mathematical theorems under change in mathematical context. In this case, what counts as a near-identity transformation may be highly discrete, such as a change in dimension or kind of domain. This allows us to think of the stability of a theorem in terms of how much or how easily (in the sense of how much its formulation has to change) it generalizes. An example we might consider is the fundamental theorem of calculus, which, starting from functions of a single real variable, is stable under a wide range of variations that preserve its basic content, extending to complex variables, vector variables, different kinds of integration (line, surface, etc.) and to curved spaces (manifolds).

This shows a variety of ways in which approximation can be introduced in purely syntactic contexts, or essentially syntactic in the sense that some syntactic structure stands in contrast to its interpretation. It is perhaps more obvious how approximation may be introduced in purely semantic contexts. Thinking of semantics in terms of objects in contrast to a symbolic syntax, then it is natural to think of near-

identity in terms of indistinugishability. Just as before, we have two distinct kinds of variation. The first kind is exactly indistinguishable objects, in the sense of having no distinguishing properties according to some standard.[18] A nice example of this is identical particles in quantum mechanics. This case corresponds to isomorphism, since the indistinguishability relation is an equivalence relation. But we can also have approximately indistinguishable objects, which are micro-locally indistinguishable but over wider variations can be distinguishable. Though not necessarily picking out "objects", a nice example of this is colours, for which close-by colours can be indistinguishable but are distinguishable over larger changes. This case corresponds to contisomorphism, since the near-indistinguishability relation is reflexive and symmetric but not transitive.

It should be natural enough to see how small changes to an object treated as nearly-indistinguishable give rise to near-identity transformations of objects. For a mathematical example, we can consider the geometric objects corresponding to bivariate polynomials, namely one-dimensional algebraic varieties.[19] These algebraic varieties are curves in the plane that correspond to the zero-sets of bivariate polynomials, i.e., for a polynomial $P(x, y)$, the locus of points in the plane that satisfies the equation $P(x, y) = 0$. We will restrict ourselves to real varieties, i.e., where we only consider real valued solutions to the polynomial equation. Traditionally such geometric objects are uniquely given and do not admit of variation. We may suppose a context, however, where small variations of the curve are acceptable and can be treated as effectively the same curve. For a formal condition, we could require that the maximum separation of the two curves is less than a tolerance ε in order to be considered nearly-identical. Thus, a single curve is expanded into a continuous family of effectively equivalent curves defined by variations of the curve within the tolerance. This kind of approximate identity is essentially similar to the colours example above, since sufficiently close curves are nearly-indistinguishable but beyond the tolerance differences are noticeable and near-identity fails.

One may see that allowing entities to vary by near-identity transformations can lead to near-identity variations of properties or relations of such objects. In the case of polynomial varieties, an example of a relation is $P(x, y) = 0$. To make the assumption that nearby curves (within the tolerance ε) are nearly-identical coherent with the relation $P(x, y) = 0$, we must allow this relation to be satisfied only approximately for each point on a nearly-identical curve. This leads to the relation $P(x, y) = 0$ being smeared out into a range of nearly-identical relations, e.g., $S = \{P(x, y) - \delta = 0 \mid |\delta| \leq \varepsilon\}$. According to the standard specified by the tolerance

[18] Note the connection between exact indistinguishability and the general concept of identity from type theory, viz., identity is always relativized to a particular type, where from a structural perspective all instances of the type have identical structure.

[19] Generally, the algebraic varieties of bivariate polynomials are geometric objects in what is called in algebraic geometry the *complex plane*, meaning $\mathbb{C}^2$, not the Argand plane of complex analysis, which is a geometric representation of $\mathbb{C}$. Since the algebraic complex plane is difficult to visualize, we restrict to real algebraic varieties, where the variables x and y can only take values in $\mathbb{R}$. Real algebraic varieties of bivariate polynomials are therefore curves in the real plane $\mathbb{R}^2$.

ε, then, having a curve that satisfies any of the equations in the set S would allow us to say in a precise sense that the relation specified by $P(x, y) = 0$ is approximately satisfied. An effectively valid inference based on the assumption of $P(x, y) = 0$, then, would yield some property or relation specified by q such that any relation in S yields a property or relation nearly-identical to q.

We now have a sense of how the notion of effectively valid inferences, as inferences producing stable outputs (conclusions, consequences) under small variations of their inputs (premises, assertions, assumptions), can be made precise in terms of near-identity transformations of sentences, both for purely or essentially syntactic sentences and for properties or relations of objects. We have not yet seen proper instances of effectively valid inferences. After the following section, in which we clarify the basic requirements for making inferences in interpreted languages and introduce some notation, we will see in Sect. 5 how effectively valid inferences arise naturally when approximation methods are used in scientific practice.

4 Interpreted Languages and Synto-Semantic Stability

In scientific practice it is common to work with interpreted languages, i.e., where we have a particular formal language or symbolic system and a well-defined intended interpretation. So when approximation is introduced in such cases, we need in general to simultaneously track variations of the syntax and semantics. Since here we have interpreted linguistic forms, we are in a context where we can talk properly about near-truth. A natural way to think about this is the following. Suppose that we begin with an individual sentence p, which has a unique fixed meaning m. We may suppose that, with this fixed meaning m, p is true *simpliciter*. We then introduce approximation in the semantics, so that the unique fixed meaning m is smeared out micro-locally into some collection of nearly-identical meanings $\mathcal{M}$. Then, if any $m' \in \mathcal{M}$ is picked out by p then p is *nearly-true*. It should be evident that the basic structure of introducing approximation has nothing to do with sentences, and that a more generalized kind of synto-semantic approximation is possible for any syntactic form and corresponding fixed reference.

What we have so far is only half of the story, however, since we can also have approximation at the syntactic level. Thus, the individual sentence p can also be smeared out into a collection of nearly-identical sentences $\mathcal{P}_p$. The idea is that if p is nearly-true, then so should be many of the nearly-identical sentences in $\mathcal{P}_p$. The only way for all sentences of $\mathcal{P}_p$ to be nearly-true would be for m to obtain, so that p is true *simpliciter* and p is the centre of micro-local variation determining all the nearly-true sentences. But given that $m' \in \mathcal{M}_p$, $m' \neq m$, is the meaning that witnesses the near-truth of p, then some sentence $p' \in \mathcal{P}_p$ will be true *simpliciter* given that m' obtains, and p' the centre of micro-local variation. We can therefore consider the collection $\mathcal{P}_{p'}$ of nearly-identical sentences to p'. This must include p because of the symmetry of near-truth, but it need not include all $p'' \in \mathcal{P}_p$ that are nearly-identical to p, because near-truth is not transitive. The following example will illustrate this situation.

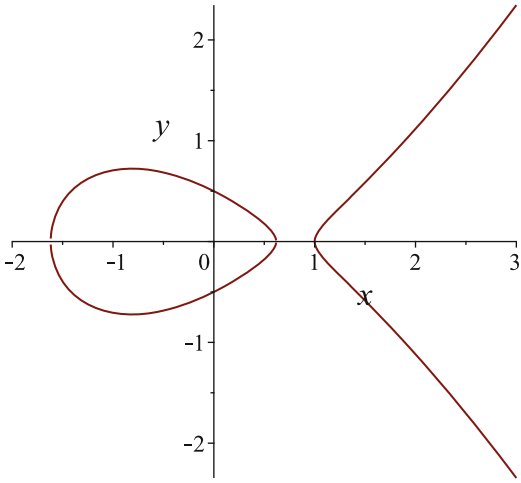


Fig. 1 An elliptic curve, the real algebraic variety of the polynomial $2y^2 - x^3 + 2x - 1$

Consider once again the algebraic varieties of bivariate polynomials. Suppose in this case that a given polynomial equation $P(x, y) = 0$ is *true simpliciter* if we find a point (x^*, y^*) such that $P(x^*, y^*) = 0$ exactly, but that the same equation is *nearly-true* if we find a point (x^*, y^*) such that $|P(x^*, y^*)| \leq \varepsilon$ for some small $\varepsilon > 0$. For concreteness, suppose we have a polynomial $P(x, y) = 2y^2 - x^3 + 2x - 1$, whose locus of real zeros picks out an elliptic curve in the plane (see Fig. 1), and that our tolerance is $\varepsilon = 0.001$. Thus, any point in the plane that satisfies the equation $P(x, y) = 0$ with an error within 0.001 is a witness to the near-truth of $P(x, y) = 0$.

Consider the point $(x^*, y^*) = (0.618, 0)$. When we substitute this into the expression for P we find that $P(0.618, 0) = -0.000029032$, so that since $|-0.000029032| < 0.001$, picking this point makes $P = 0$ nearly-true. It turns out that this point is so close to the curve because the golden ratio $\frac{1}{2}(\sqrt{5} - 1)$ picks out one of the points on the x-axis.

A natural way to introduce nearly-equivalent expressions compatible with the approximate semantics in this case is to say that any polynomial P' otherwise identical to P but with a constant term that is within ε of -1 (inclusive) is nearly-equivalent to P. Then, the polynomial equation $P'(x, y) = 2y^2 - x^3 + 2x - 0.999970968 = 0$ is nearly-equivalent to that of $P = 0$, which the point $(0.618, 0)$ now satisfies exactly. Thus, $P' = 0$ is the sentence that is true *simpliciter* given the witness point $(0.618, 0)$, which determines the centre of a micro-local range of nearly-equivalent sentences around it, determined now by the condition that the constant term is within ε of -0.999970968 (inclusive). Notice, then, that the polynomial equation $P''(x, y) = 2y^2 - x^3 + 2x - 1.001 = 0$ is nearly-equivalent to that of $P = 0$, given the definition, but it is not nearly-equivalent to $P' = 0$. If we substitute the point $(0.618, 0)$ into the equation $P''(x, y) = 0$ we find that it does not witness the near-truth of $P''(x, y) = 0$ according to our standard. This is because the point (0.618.0) witnesses the truth *simpliciter* of $P'(x, y) = 0$ and the near-truth of $Q(x, y) = 0$ for any Q that differs from P' within the tolerance. This

includes the original $P = 0$ but does not include $P'' = 0$ because P'' differs from P' by $1.001029032 > \varepsilon$. Thus, $P'' = 0$ is not nearly-equivalent to $P' = 0$, thus witnesses to the near-truth of $P' = 0$ need not witness the near-truth of $P'' = 0$. This shows how witnessing near-truth is not transitive.

This matter of syntactic and semantic approximation being compatible is essential for being able to move back and forth between syntactic and semantic reasoning in an interpreted language. When approximations are introduced in the semantics, the syntax should be flexible enough to allow a compatible form of approximation in the syntax. Similarly, if approximations are introduced purely syntactically, i.e., without it being immediately obvious that the meaning of the syntactic approximation is nearly-true, then it should be possible to find a variation of the semantics that makes the nearly-equivalent syntactic form nearly-true. This flexibility of the syntax and semantics under variation commonly obtains for languages in applied mathematics, particularly in well-developed areas where the semantics is well-understood.

Such compatibility of syntax and semantics is actually another kind of stability property, this time relating syntax and semantics. We may see here two natural stability properties relating syntax and semantics of an interpreted language. One is the property that approximations in the syntax lead to approximations in the semantics, which we may call *semantic stability* of the syntax. The other is the property that approximations in the semantics lead to approximation in the syntax, which we may call *syntactic stability*. Both of these two properties holding together, which we will assume for interpreted languages, may be called *synto-semantic stability* of an interpreted language.

That the two conditions (semantic and syntactic stability) are not equivalent can be made clear in terms of the concept of a continomorphism. In fact, each condition implies that its respective map is a continomorphism, one (semantic stability) from syntax to semantics and the other (syntactic stability) from semantics to syntax. Each condition on its own does not, however, imply that the map is also a cont*iso*morphism, which would indicate invertibility. Thus, it is possible to have a semantically stable language, meaning that small changes to the syntax produce small changes in the semantics, have *some* semantic changes require large changes in the syntax. This is to say, that the condition of semantic stability does not imply that small changes of the syntax lead to *all* possible small changes of the semantics, only certain changes of the semantics. If the language is also syntactically stable, however, then any small change in the semantics leads to a small change in the syntax. Thus, the two conditions holding together, i.e., when the language is synto-semantically stable, imply that the intepretation map for the language is a cont*iso*morphism.

Since the notion a synactic approximation is an unfamiliar one, as might be a distinction between syntactic and semantic reasoning, it will help to understand how these stability conditions can fail in realistic inferential circumstances. Suppose that our interpreted language is the mathematics used in standard, or what Wallace [22] calls Lagrangian, quantum field theory. In this context situations are sometimes encountered where needed mathematical results are assumed to hold in some

approximate domain, even though there is no clear semantics to support the theorems and they cannot be proved. For example, building theories for interacting fields by analogy to theories for non-interacting fields that have a rigorous theory. Here one makes syntactic transformations that are not supported by a well-defined interpretation, a kind of breakdown of semantic stability.[20] This is a case where the rigorous semantics is too rigid for the syntactic moves the scientist needs to make.

Although an example of failure of syntactic stability is more tricky to provide, consider the following example that can illustrate the idea. Consider a theory of relational databases that requires that all relations in tables be strictly consistent with the database constraints, and suppose that this theory is the basis of the design of a database management system (DMS). If we roughly identify the DMS and the theory (for illustrative purposes), we can regard the DMS as the syntactic level of an interpreted language, such that actual databases provide its semantics. Then suppose that it is recognized that approximately consistent relations are required in some application, and the databases are modified/hacked to allow them, but the management system, based on strict consistency, is forced to view them as strictly consistent, when they are not, or as inconsistent (depending on how the modified databases are structured). This could be seen as an example of a breakdown of syntactic stability, because the system design cannot adapt to the semantic variation, being based in a theory that does not allow approximately consistent constraints. This is a case, then, where the syntactic level is too rigid for the semantic variations the data scientist needs to make.

It may appear that the relationship articulated by semantic and syntactic stability is related to soundness and completeness. There is certainly a meaningful analogy, but soundness and completeness as notions only make sense for *un*interpreted languages. Since our focus is on interpreted languages here, we do not consider effective versions of soundness and completeness. To explore the analogy a bit, we can say that semantic stability is analogous to soundness since it pertains to variations in the syntax of the language that result in meaningful variations in the semantics, or variations to syntax that can be tracked in the semantics. Viewed in terms of inferences, semantic stability requires that inferences that can be made in the syntax can be made in the interpretation of the language. Thus, it is in essence a point-wise version of soundness, in that rather than requiring provable statements to be valid in all interpretations, they just need to be valid in the specified interpretation of an interpreted language. On the other hand, syntactic stability is analogous to completeness since it pertains to variations in the semantics that can be tracked or recovered in the syntax. Viewed in terms of inferences, syntactic stability requires that inferences that can be made in the interpretation of the language are also provable in the syntax of the language. It is therefore essentially a point-wise

[20] One can see the empirical validation of computed results as providing evidence that it will be mathematically possible to develop a more general semantics to provide rigorous proofs, and this has happened in many areas of this field over time. The fact that new mathematics has to be developed in such cases, however, is a reflection of the rigidity of the existing semantics and why this is a failure of semantic stability.

version of completeness, in that rather than all logical truths being required to be provable, the truths in the specified interpretation of an interpreted language need to be provable. An essential difference to classical soundness and completeness, however, is that the inferences in question need only be effectively valid, and the relations of semantic and syntactic stability need only obtain locally to a relevant range of variations.

To specify more precisely how the concept of synto-semantic stability pertains to stability of inference, we will introduce some notation. To bring out the (effective) logical character of the inferences involved, we will introduce a special notation based on an extension of the single turnstile notation in standard logic. For strict *syntactic* consequence relations we will use the notation

$$\text{assumed sentences} \;\Big|\!\!\frac{}{\text{globally imposed sentences}}\; \text{valid consequence},$$

so that the (local) assumptions appear on the left of the turnstile, valid consequences appear on the right, and any system-wide (structural) sentences, functioning as constraints, can be displayed underneath the horizontal stroke of the turnstile. For strict *semantic* consequence relations in an interpreted language we will use the special notation

$$\text{assumed relations} \;\Big\|\!\!\frac{}{\text{globally imposed relations}}\; \text{valid consequence},$$

where the second vertical stroke indicates we are dealing with inferences from interpreted relations to other interpreted relations.[21]

The notation as it stands indicates an exact consequence relation (syntactic or semantic). Since our consideration of micro-local variation involves treating nearly-identical sentences and relations as effectively equivalent, we can consider being related by a near-identity contisomorphism (micro-locally) as an *effective equivalence relation*. An effective equivalence relation requires a fixed centre of variation according to which equivalence is judged, so that *relative to that centre* all of the nearly-identical entities can be considered equivalent. As such, the relation of being nearly-identical to a *fixed* entity is symmetric, reflexive *and* transitive. The relation among this collection of entities should not be considered an equivalence relation *simpliciter*, however, since it remains the case that for two nearly-identical entities a and b some other entity c may be nearly-identical to b but not to a. Thus, near-identity transformations give rise to micro-local effective equivalence relations.[22]

[21] This may seem like an odd notation, but this notation has been used before in the interest of specifying a notion of semantic consequence for dynamical systems by van Fraassen [18] for what he calls "semi-interpreted languages". Such notation therefore has a precedent in philosophy of science. Though the notation overlaps with that for forcing, it should be clear from the context what is the intended meaning.

[22] This can be seen as analogous to the local flatness of smooth spaces.

Let us denote the effective equivalence relation connecting nearly-identical entities by $\sim$. Thus, for a set of sentences Γ or their interpretations $[\![\Gamma]\!]$, with respective consequences, p or $[\![p]\!]$, and framework constraints, $\mathcal{F}$ or $[\![\mathcal{F}]\!]$, the effectively valid consequence relations can be denoted by

$$\Gamma \,\big|\!\!\frac{\sim}{\mathcal{F}}\, p, \qquad [\![\Gamma]\!] \,\big\|\!\!\frac{\sim}{[\![\mathcal{F}]\!]}\, [\![p]\!].$$

Since we are working with a fixed interpretation for an interpreted language and the double vertical stroke makes it clear that we are dealing with semantic consequence relations, we will generally omit the interpretation notation $[\![\cdot]\!]$ and simply write

$$\Gamma \,\big\|\!\!\frac{\sim}{\mathcal{F}}\, p$$

for an effectively valid semantic consequence.

The idea behind the condition of synto-semantic stability for an interpreted language is that the syntax and semantics *covary*, so that changes to one are reflected in the other. This property is essential so that proof or derivation methods that can be applied purely syntactically lead to stable results in the interpretation. Thus, syntactic variations are carried forward to semantic variations for a synto-semantically stable language. Though some care is needed to judge the scope of stability of approximations, we expect interpreted languages to have the property that stable syntactic inferences result in stable semantic ones, i.e., that

$$\Gamma \,\big|\!\!\frac{\sim}{\mathcal{F}}\, p \Longrightarrow [\![\Gamma]\!] \,\big\|\!\!\frac{\sim}{[\![\mathcal{F}]\!]}\, [\![p]\!].$$

We may call this property of an effectively valid syntactic inference a *semantically stable consequence*. Conversely, we expect that stable semantic variations will result in stable syntactic ones, so that they can be proved in the language, in which case

$$[\![\Gamma]\!] \,\big\|\!\!\frac{\sim}{[\![\mathcal{F}]\!]}\, [\![p]\!] \Longrightarrow \Gamma \,\big|\!\!\frac{\sim}{\mathcal{F}}\, p.$$

This property of an effectively valid semantic inference may be called a *syntactically stable consequence*.

The advantage of having a language with both of these properties, where we would say that consequence relations are *synto-semantically stable*, is that we can reason stably using both syntactic and semantic arguments, which is common in mathematical practice. As indicated above, it is this stability property that takes the place of soundness and completeness for interpreted languages. Since it is a property that is possessed or assumed in many scientific languages, we will use a special notation for synto-semantically stable consequences in interpreted languages, namely

$$\Gamma \,\big\|\!\big\|\!\!\frac{\sim}{\mathcal{F}}\, p,$$

where we have introduced a third vertical stroke (one for syntax, two for semantics) to indicate that the consequence is both syntactically and semantically effectively valid in the interpreted language. This can be understood as asserting the (effective) commutativity of interpretation and derivation. We will consider our first explicit examples of such inferences in the following section.

5 Effective Validity in Scientific Inference

For this and the following section we will assume that we are working in synto-semantically stable (interpreted) languages, and will use the triple-turnstile notation to reflect this. When considering the nature of effectively valid inference in the context of scientific inference-making, we must appreciate that the mathematical framework within which inferences are being made may consist of a number of inter-related subframeworks. In particular, it is common for inferences to consist of finding a solution to a mathematical problem, where the problem is typically merely one in a space of similar problems picked out by different data (parameter values, functions). The solution also lives in a space of similar solutions, related by some form of variation to the solution to a specified problem. As such, solving problems can be viewed as computing a map from a space of problems (problem space) to a space of solutions (solution space), since we are typically interested in solving problems with parameters in them. Often in practice an approximate solution provides us with all of the information we need, so we are happy to obtain nearby solutions to problems. Moreover, when we are modeling a target system, so that the specified problem itself involves approximations, we are happy to be able to solve nearby problems approximately, provided they provide just as much information about a target system that the exact solution to the specified problem would provide.[23] For these and related reasons effectively valid inference becomes important in scientific inference-making.

We will consider some basic features of effectively valid inference in an interpreted language in terms of a simple example. To this end, let us consider a simple mechanical system, the simple pendulum, composed of a single massive weight connected to a massless rod frictionlessly connected to a pivot (see Fig. 2). For a rod of unit length, this system specifies a simple differential equation of motion

$$\ddot{\theta} + g \sin\theta = 0, \tag{5}$$

where dots denote time derivatives, θ denotes the angle made by the rod relative to the vertical, and g is the gravitational acceleration in units of inverse time squared. It is deceptively simple, however, because it is nonlinear (due to the presence of

[23] See [11, Ch. 5] for a discussion of this matter in the context of mathematical modeling using ordinary differential equations (dynamical systems).

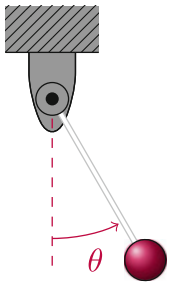


Fig. 2 The simple pendulum model. It has one configurational degree of freedom specified by the angle θ. Together with the corresponding momentum degree of freedom ℓ_θ of the weight, the system has a two dimensional phase space

$\sin\theta$) and so cannot be solved by the usual methods presented in undergraduate differential equations courses, which (for second- and higher-order) work for linear equations. For this reason, one often seeks an approximate solution to this equation by considering the case of small oscillations. In the case that we assume θ is sufficiently small, we can conclude on essentially syntactic grounds that the $\sin\theta$ term can be replaced with its linear approximation θ. This is an instance of the standard technique of linearizing the equation, which replaces a function with its linear Taylor approximation. The result of doing this is that we obtain the equation for the simple harmonic oscillator

$$\ddot{\theta} + g\theta = 0, \tag{6}$$

which has circular (sine and cosine) functions as solutions. Treating the original differential equation as an assumption, we add the assumption that the angle measured in radians is small, specifically $\theta \ll 1$, which corresponds to the condition that the angle remain much smaller than 57.3°. Under these conditions, the simple harmonic oscillator equation of motion (6) is effectively equivalent to the equation for the simple pendulum (5).

This provides our first concrete example of an effectively valid scientific inference:

$$\ddot{\theta} + g\sin\theta = 0, \theta \ll 1 \;\Big\|\!\!\sim_{\mathcal{P}}\; \ddot{\theta} + g\theta = 0,$$

where the subscript $\mathcal{P}$ on the effective equivalence relation $\sim$ indicates that the near-identity relation is given in terms of the problem space. Note that this inference stays entirely within the problem space here, which for concreteness could be specified to be the space of second order differential equations. As such, we have defined conditions under which a range of problems are to be considered nearly-identical, and derived a nearly-identical problem that is easier for us to solve.

Given the potential for instability of effectively valid inferences, we could not necessarily expect approximate solutions to (6) to provide a good approximate solution to (5), which is to say it is not immediately clear that we have chosen an appropriate standard of near-identity given our modeling aims. But if we obtain an *exact* solution to the simple harmonic oscillator, given we are interested in small angles, then we should expect a good approximation to the simple pendulum

for small angles over a limited (local) time scale, which is part of what this statement says. Indeed, this is the reason why we linearize, to obtain a locally valid approximation to the original equation.

Consider, then, the simpler equation (6). To produce a unique solution we must supply some other condition, such as an initial condition. We may take $\theta(0) = \theta_0$ as an initial angle and take $\dot{\theta}(0) = 0$ as the initial angular velocity. In this case, we can solve (6) to obtain $\theta(t) = \theta_0 \cos(\omega t)$, where $\omega = \sqrt{g}$, which is straightforward to verify by substitution into (6). We can also express this in our logical notation. Suppose that we now work in a framework where the differential equation (6) is imposed as a global constraint, then we can express the exact solution to the initial value problem as

$$\theta(0) = \theta_0, \dot{\theta}(0) = 0 \;|\!|\!|\!\!\frac{\quad}{\ddot{\theta}+g\theta=0}\; \theta(t) = \theta_0 \cos(\omega t),$$

where we now drop the $\sim_{\mathcal{P}}$ to indicate the fact that the consequence relation is exact. Thus, this expresses that the solution curve $\theta(t) = \theta_0 \cos(\omega t)$ is a valid consequence of the initial conditions $\theta(0) = \theta_0$ and $\dot{\theta}(0) = 0$. Thus, in this case we have computed an exact map from the given initial conditions (premises) in the problem space to a solution curve (conclusion) in the solution space.

Combining this with our conclusion that the simple harmonic oscillator is an effectively valid approximation of the simple pendulum for small angles, we can shift to a framework in which the differential equation (5) is imposed as a global constraint, in which case we can express the approximate solution to the corresponding initial value problem as

$$\theta(0) = \theta_0, \dot{\theta}(0) = 0, \theta_0 \ll 1 \;|\!|\!|\!\!\frac{\sim_{\mathcal{F}}}{\ddot{\theta}+g\sin\theta=0}\; \theta(t) = \theta_0 \cos(\omega t), \tag{7}$$

where the subscript $\mathcal{F}$ now indicates that the effective equivalence is in terms of the framework constraint $\ddot{\theta} + g \sin\theta = 0$, i.e., this says that there is an effectively equivalent framework within which the conclusion is valid. This inference follows (again for a limited (local) time scale) since if the initial angle $\theta_0 \ll 1$ then we know $\theta(t) \ll 1$ holds for all times t. Note once again that there is an implicit boundary of validity for this consequence relation, as there is for all of the relations we consider in this paper, since it only holds locally to a certain time range $t \in [0, T(\theta_0)]$, for some function T, outside of which the inference becomes unstable. This shows how effective logic captures the standard approach of linearizing the simple pendulum.

The solution strategy here was to approximate the nonlinear problem by a linear problem we could easily solve, that would still be able to describe the simple pendulum for small oscillations and short times. We made a very strong assumption about the nature of the problem, however, since for this strategy to be viable another kind of stability property is required, viz., that small changes to the problem must result in small changes to the solution.[24] This properly holds in this case, but it

[24] In dynamical systems this property is called *dynamical stability* and in the context of numerical methods it is called *well-conditioning*. A weaker concept called *well-enough conditioning* was

does not in all cases, such as for chaotic systems for which we are interested in the behaviour of the system over time. Considering the stability of solutions, or relevant properties of solutions (see footnote 24), under changes to the problem (or data) is therefore very important in the context of approximate inference to ensure the stability of inferences.

If our solution strategy was successful, it must be that the smaller the initial angle, the closer the exact solution of the simple pendulum approaches the corresponding simple harmonic oscillator solution. In fact this can be proved explicitly in this case by solving the nonlinear equation directly, and then varying the solution into the range of small oscillations.[25] Before we examine how this approximation argument works, let us first consider how to treat approximations in the solution space.

The configuration space is particularly simple for this system, since the mass of the pendulum is restricted to move on a circle of radius equal to the length of the rod, which we have assumed is length 1 for simplicity. Thus, for a description of the motion of the system it is enough to specify the angle $\theta(t)$ as a function of time. For definiteness, we can picture the state space as a cylinder of radius one, where the height is the time and the angular position is θ. The motion of the pendulum then traces out a unique curve on the cylinder over time. The scientific problem to be solved is to specify this curve for any given initial state the pendulum is in.

In this case, considering Eq. (5) as a model of a real pendulum, it is clear why approximation is acceptable, since there is error involved in the construction of the model, so error in its solution can be acceptable provided it does not interfere with the applicability of the result. Thus, we are interested in any curves that stay sufficiently close, say within any experimental error, to the model solution curve for a reasonable period of time. This introduces a near-identity condition very similar to the one introduced in the previous section for one-dimensional algebraic varieties. More specifically, we may require that any acceptable solution $\varphi(t)$ differ from the exact solution by no more than some tolerance ε, i.e., so that $|\theta(t) - \varphi(t)| \leq \varepsilon$. It is such a condition that can only be satisfied for a limited period of time when approximation is allowed.

Consider now the nonlinear equation (5) for the simple pendulum. What makes this equation different from most nonlinear equations encountered in practice is that it can be solved exactly. For the same initial condition we considered above this equation can be solved in terms of the Jacobi elliptic function $\mathrm{sn}(z, k)$, where z is a complex variable and $k \in [0, 1]$ is a parameter. For our purposes we do not need to know much about this function, the following two properties are enough: (1) it is periodic along the real line, i.e., for z real; and (2) in the limit $k \to 0$, $\mathrm{sn}(z, k) \to \sin z$, so it asymptotically approaches the sine function as the second variable k goes

introduced by Corless [2] to describe the situation where some *relevant* properties of the solution are stable under perturbations of the problem. A generalized version of this latter notion, called *partial-well-conditioning*, is discussed in [12, Chs. 5,6].

[25] Note that the stability of problems can be studied without solving the problem in question. For details of how this conditioning analysis can be done for various mathematical problems see [3].

to zero. The solution of (5) can be expressed in the form $\sin(\theta(t)/2) = A\,\mathrm{sn}(\omega t + K, A)$,[26] where K is the quarter period and $A = \sin(\theta_0/2)$ contains the initial angle [10]. We can write this in symbols as

$$\theta(0) = \theta_0, \dot{\theta}(0) = 0 \,\|\!|\!|_{\ddot{\theta}+g\sin\theta=0}\, \sin(\theta(t)/2) = A\,\mathrm{sn}(\omega t + K, A),$$

noting that this is an exact consequence relation and (5) is imposed as a global constraint.

To get a sense for the behaviour of this solution, Fig. 3 compares the behaviour of the simple pendulum and simple harmonic oscillator solutions for a moderate initial angle $\theta_0 = 1$. It is evident for such a large initial angle the solutions are very different, and significantly so after 6 s. We see that the amplitudes of the two solutions are the same, as they should be since they both describe a conservative pendulum system dropped from rest at the same initial angle. The frequencies are quite different, however, and a noticeable difference is observable on this plot after about 0.3 s.

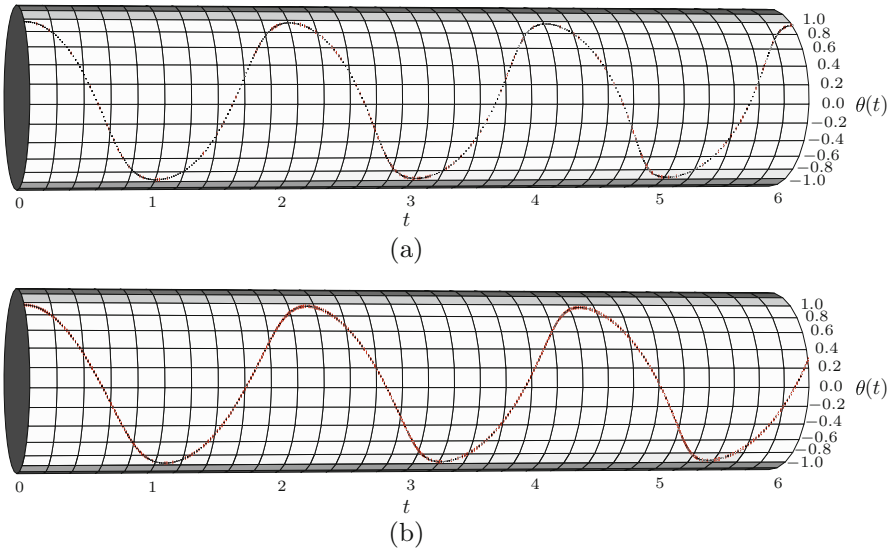


Fig. 3 The simple harmonic oscillator compared to the simple pendulum for a moderate initial angle. Grid lines are separated by 0.2 radians around the cylinder and 0.2 s along its length. The amplitudes match, but the frequencies are distinct, and the solutions diverge noticeably after only about 0.3 s. (**a**) $\theta(t)$ for the simple harmonic oscillator with $\theta_0 = 1$. (**b**) $\theta(t)$ for the simple pendulum with $\theta_0 = 1$

[26]The solution can be written as a function of θ simply by rewriting it in the form $\theta(t) = 2\sin^{-1}(A\,\mathrm{sn}(\omega t + K, A))$.

Now, if we restrict ourselves to small angles, so that $\theta \ll 1$ rad ($\theta \ll 57.3°$) as before, then $A \approx \theta_0/2$ and $\sin\theta \approx \theta$, so that the expression

$$\sin(\theta/2) = A\,\mathrm{sn}(\omega t + K, A)$$

reduces to

$$\theta(t) = \theta_0\,\mathrm{sn}(\omega t + K, \theta_0/2).$$

Furthermore, since θ_0 is very small, corresponding to the regime where the parameter $k = \theta_0/2 \to 0$, the elliptic function $\mathrm{sn}(x+K, k)$ approaches $\sin(x+\pi/2)$ (sine advanced by a quarter period).[27] Since $\sin(x + \pi/2) = \cos(x)$, we therefore obtain in the limit of small initial angles,

$$\theta(t) = \theta_0 \cos(\omega t),$$

the solution of the simple harmonic oscillator.

Thus, for small initial angles, and sufficiently short times, the simple harmonic oscillator solution is a good approximation to the motion of the pendulum. Recall our condition for an acceptable solution that it differ from the exact solution by less than some tolerance ε. We can then consider any two solutions $\theta(t)$ and $\varphi(t)$ to be effectively equivalent, written $\theta(t) \sim_{\mathcal{S}} \varphi(t)$ provided $|\theta(t) - \varphi(t)| \leq \varepsilon$, i.e., provided they are within ε of each other on the solution cylinder. We therefore have established a solution-approximation version of the consequence relation we established earlier using a linearization argument:

$$\theta(0) = \theta_0, \dot{\theta}(0) = 0, \theta \ll 1 \;\Big|\!\Big|\!\Big|_{\ddot{\theta}+g\sin\theta=0}^{\sim_{\mathcal{S}}} \theta(t) = \theta_0 \cos(\omega t).$$

This is actually a much stronger result, since the tolerance ε now refers to a definite bound on the error in the solution, rather than a definite bound on the error in the differential equation model. We have therefore shown rigorously that the simple harmonic oscillator solution is an effectively valid solution to the simple pendulum for small angles, a statement that is true for sufficiently short time scales. This result is illustrated in Fig. 4, which shows how for an initial angle of $\theta_0 = 0.2$ the solutions of the simple pendulum and simple harmonic oscillator are nearly-indistinguishable through 6 s.

We have seen in this section that the nature of effectively valid inferences in a problem-solving context depends very much on where approximations are made (problem/data or solutions) and on the stability properties of the specified problem of interest. There is an important connection between these concerns of effectively valid inference and a branch of modern error theory called *backward error analysis*, which distinguishes between error viewed as variation of the problem

[27] This follows because $\mathrm{sn}(x, k) \to \sin(x)$ as $k \to 0$, and $K \to \pi/2$ because $\pi/2$ is the quarter period of the sine function.

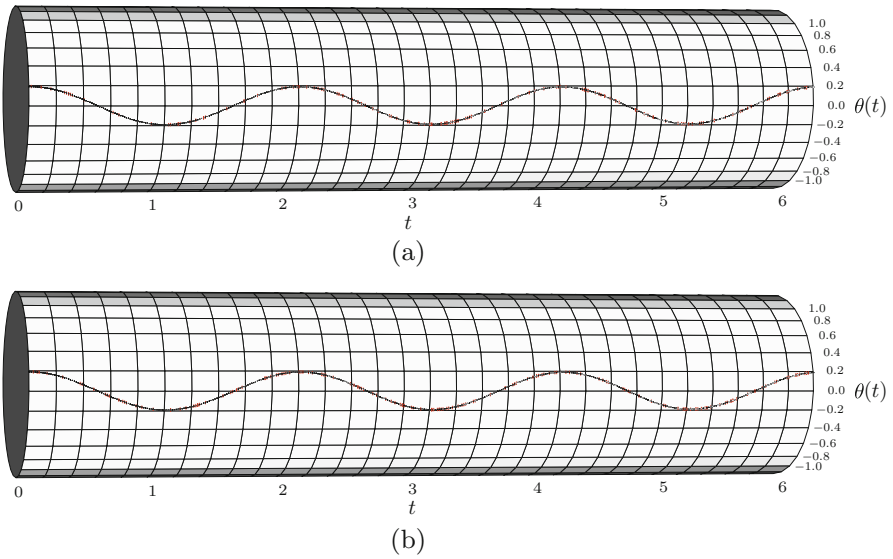


Fig. 4 The simple harmonic oscillator compared to the simple pendulum for a small initial angle over the time interval [0, 6]. Grid lines are separated by 0.2 radians around the cylinder and 0.2 s along its length. The phase and amplitude match closely, with only a small error discernable after 6 s. (**a**) $\theta(t)$ for the simple harmonic oscillator with $\theta_0 = 0.2$. (**b**) $\theta(t)$ for the simple pendulum with $\theta_0 = 0.2$

or data (backward error) and error viewed as variation of the solution (forward error), and the stability of the solutions under variations of the problem/data (conditioning). Though central to effectively valid inference in problem-solving in applied mathematics, a consideration of backward error analysis here is beyond the current scope. We refer the interested reader to Fillion and Corless [4] and Fillion and Moir [5] for discussions of the philosophical significance of backward error analysis and to Corless and Fillion [3] for a treatment of the foundations of numerical methods from the point-of-view of backward error.

Though we used a simple differential equations example to illustrate the basic features of effectively valid inference, the applicability of the concepts of effective validity does not rely on anything specific to differential equations. Any problem that can be expressed in terms of some input assumptions and an output solution can be treated in essentially the same way. The case of deductive validity captures any instances where exact solutions are sought, and effective validity captures cases where an appropriate standard for approximate problems and solutions is introduced. The analogue of an inference rule in this case is any operation that produces a solution to a problem given appropriate input. Thus, the nature of scientific inference in a problem solving context is seen to be highly analogous to logical deduction. Indeed, in the case of exact solutions, it is exactly equivalent to deduction. When we introduce approximation, as is very common in scientific practice, the analogue in effective logic is then *nearly deductive* effectively valid inferences.

6 Effective Logic and Scientific Problem-Solving

We have now seen how basic scientific problem-solving using approximation can be treated in terms of effectively valid inference. We will see in this section how more complex problem-solving methods can be treated in terms of effective logic. It is very common in science to have a problem that cannot be solved exactly in the originally-posed form. As was mentioned in the previous section, this is generally the case for nonlinear differential equations, which is the typical case for models in applied mathematics that are accurate for a wide range of conditions. In such situations, one generally resorts to some form of approximation. This typically requires some kind of modification of the problem itself, a shift to an analogous problem that is easier to solve. Indeed, this is actually what was done with the use of linearization to solve the simple pendulum problem approximately, we shifted from a nonlinear problem to a linear one, which was easier to solve. This strategy of transforming a difficult problem into an easier one in the search for an approximate solution is a very common one in scientific practice, and underlies strategies of computational complexity reduction in computational science, as was pointed out above. We will see in this section that this method can be understood in terms of the stability of inferential relations for mappings between interpreted languages.

Moving to the next more complicated problem from the simple pendulum takes us to the double pendulum, where we simply add another rod and weight to the simple pendulum. The result is a simple looking system that exhibits surprisingly complex behaviour; indeed, the double pendulum is chaotic for some initial conditions. The double pendulum is a simple example of a nonlinear differential equation for which we do not know a class of special functions that solve it analytically, unlike the simple pendulum that can be solved with Jacobi elliptic functions. This means we need to use other means to describe the behaviour of solutions of the equation. A standard approach here is to use numerical methods and computation to solve the equations approximately. We will see in this section how the approach of solving equations by numerics can be understood in terms of effective logic.

We begin in this case with the differential equation of motion for the double pendulum, which can be given in terms of Hamilton's equations

$$\dot{\mathbf{q}} = \frac{\partial H}{\partial \mathbf{p}}, \quad \dot{\mathbf{p}} = -\frac{\partial H}{\partial \mathbf{q}}, \tag{8}$$

with the generalized position $\mathbf{q} = (\alpha, \beta)$ composed of the two angles describing the state of the system (see Fig. 5) and the generalized momentum $\mathbf{p} = (\ell_\alpha, \ell_\beta)$ composed of the angular momenta of the two weights. It can be shown, according to a standard algorithm [see 6], that the Hamiltonian for the system is

$$H(\mathbf{q}, \mathbf{p}) = -2\cos\alpha - \cos(\alpha+\beta) + \frac{l_\alpha^2 - 2(1+\cos\beta)l_\alpha l_\beta + (3+2\cos\beta)l_\beta^2}{3-\cos 2\beta}. \tag{9}$$

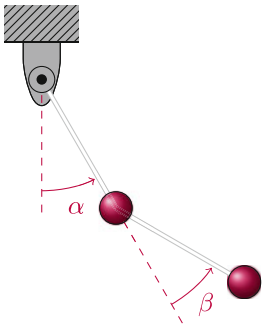


Fig. 5 The double pendulum model. It has two configurational degrees of freedom, specified by the angles α and β. Together with the corresponding momentum degrees of freedom ℓ_α and ℓ_β of the two weights, the system has a four dimensional phase space

If we suppose that we are interested in the case where the first weight is held at an initial angle of $\alpha_0 = \pi/2$ (90°), and the second weight is left hanging by gravity, corresponding to $\beta_0 = -\pi/2$, then the inferential problem we are faced with, analogous to the simple pendulum, is

$$\alpha_0 = \pi/2, \beta_0 = -\pi/2 \,\Big\|\!\!\frac{\sim}{\mathcal{H}}\, \alpha(t) =?, \beta(t) =?,$$

where $\mathcal{H} = \left\{\dot{\mathbf{q}} = \frac{\partial H}{\partial \mathbf{p}}, \dot{\mathbf{p}} = -\frac{\partial H}{\partial \mathbf{q}}, H(\mathbf{q}, \mathbf{p}) = (9)\right\}$ specifies the framework constraints for our Hamiltonian system. This notation is intended to express that we seek an effectively valid solution for the given initial conditions that specifies the evolution of the angles α and β over time. Unlike for the simple pendulum, however, we have no way of obtaining a valid (or effectively valid) solution directly, so we must search for an approximate solution by *transforming* the problem.

Since we seek an approximate solution and the system is chaotic for these initial values, we cannot expect fidelity for a very long time given that small errors grow exponentially. We can control the error quite well, however, by using a specialized numerical method that preserves very closely the geometric structure of the problem, in this case the symplectic form on phase space defined by Hamilton's equations. Numerical methods that accomplish such near-preservation of geometric structure are called *geometric numerical methods*, and specifically *symplectic methods* in the case of the symplectic structure of Hamiltonian systems [7].

A simple example of a symplectic method that we can use for this problem is the Störmer–Verlet method, which replaces Hamilton's continuous-time differential equation (8) with a pair of discrete-time difference equations[28]

$$\mathbf{q}_{n+1} = \mathbf{q}_n + \frac{h}{2}(\mathbf{k}_1 + \mathbf{k}_2), \quad \mathbf{p}_{n+1} = \mathbf{p}_n - \frac{h}{2}(\mathbf{m}_1 + \mathbf{m}_2), \tag{10}$$

[28] Using the first equation to illustrate, we can see that writing these equations in a slightly different form, $\frac{\mathbf{q}_{n+1} - \mathbf{q}_n}{h} = \frac{1}{2}(\mathbf{k}_1 + \mathbf{k}_2)$, shows how the time derivative, $\frac{d\mathbf{q}}{dt}$, of Hamilton's equations is approximated by a finite difference and the partial derivative of the Hamiltonian, $\frac{\partial H}{\partial \mathbf{p}}$, is approximated by the average of its value at two special points.

where $\mathbf{k}_1$, $\mathbf{k}_2$, $\mathbf{m}_1$ and $\mathbf{m}_2$ are given by the (semi-implicit) equations

$$\mathbf{k}_1 = \frac{\partial H}{\partial \mathbf{p}}\left(\mathbf{q}_n, \mathbf{p}_n + \frac{h}{2}\mathbf{m}_1\right),\ \mathbf{k}_2 = \frac{\partial H}{\partial \mathbf{p}}\left(\mathbf{q}_n + \frac{h}{2}(\mathbf{k}_1 + \mathbf{k}_2),\ \mathbf{p}_n + \frac{h}{2}\mathbf{m}_1\right),$$

$$\mathbf{m}_1 = \frac{\partial H}{\partial \mathbf{q}}\left(\mathbf{q}_n, \mathbf{p}_n + \frac{h}{2}\mathbf{m}_1\right),\ \mathbf{m}_2 = \frac{\partial H}{\partial \mathbf{q}}\left(\mathbf{q}_n + \frac{h}{2}(\mathbf{k}_1 + \mathbf{k}_2), \mathbf{p}_n + \frac{h}{2}\mathbf{m}_1\right), \quad (11)$$

where h is the time-step and H, for the double pendulum, is given by (9) as before. Rather than being continuous curves, $\mathbf{q}(t) = (\alpha(t), \beta(t))$, the solutions of the Störmer–Verlet equation are time series $\mathbf{q}_n = (\alpha_n(t_n), \beta_n(t_n))$, where $t_n = nh$. The idea is that the map $\varphi : (\mathbf{q}_n, \mathbf{p}_n) \mapsto (\mathbf{q}_{n+1}, \mathbf{p}_{n+1})$ on phase space that advances the system forward in time is very nearly a symplectic map, meaning that among other things energy is very nearly conserved over time. This gives the numerical method the ability to adequately control the error over extremely long times, which will give us decent performance on this chaotic problem.

To clarify the logic of this situation, observe that we are seeking to find an approximate solution to our problem by mapping the problem to a different problem in a different framework (difference equations) in a way that *nearly preserves* the structure of the original problem; this way, solutions to the new problem can give us approximate solutions to the original problem. Thus, rather than a near-identity transformation of entities or sentences, we are considering a near-identity transformation of a framework (we saw a less extreme version of this in the previous section in the near-identity transformation from (5) to (6) expressed by $\sim_{\mathcal{F}}$). If we obtain a solution to the transformed problem, we obtain it in the synto-semantics of the alternative framework, which here means that we get a discrete solution not a continuous one. Nevertheless, the nature of the near-structure-preservation guarantees that if we carry a solution to the Störmer–Verlet equation back to the framework of Hamilton's equations, we will obtain a sequence of solution points very close to the corresponding solution points of the original equation, i.e., we will have an effectively valid solution to Hamilton's equations for a sequence of times t_n. We can recover an approximate solution for the intervening times using some form of interpolation, which can convert the discrete solution $\mathbf{q}_n = (\alpha_n(t_n), \beta_n(t_n))$ into an approximate continuous solution $\mathbf{q}(t) = (\alpha(t), \beta(t))$.

Since we are now considering mappings between interpreted languages, we need a notation to indicate this. Since we may regard such a mapping as an external effective interpretation of the synto-semantics of interpreted language, we can regard the mapping operation as being somewhat analogous to an (external) interpretation of an uninterpreted language in a model, which uses the double horizontal "models" notation $\models$. Accordingly, we introduce the notation

$$\alpha_0 = \pi/2, \beta_0 = -\pi/2 \,|||\!\overset{\sim}{\underset{\mathcal{H}\to\mathcal{N}}{=\!=}}\, \alpha(t_n) =?, \beta(t_n) =?,$$

where $\mathcal{N} = \left\{\mathbf{q}_{n+1} = \mathbf{q}_n + \frac{h}{2}(\mathbf{k}_1 + \mathbf{k}_2), \mathbf{p}_{n+1} = \mathbf{p}_n - \frac{h}{2}(\mathbf{m}_1 + \mathbf{m}_2), H(\mathbf{q}, \mathbf{p}) = (9)\right\}$ indicates the constraints now imposed, to denote the mapping of the problem to the framework of the numerical method. The notation $\mathcal{H} \rightarrow \mathcal{N}$ indicates that the *source framework*, where the problem was originally posed, is $\mathcal{H}$ and the *target framework*, where the problem we mapped to is posed, is $\mathcal{N}$. Since we are working with interpreted languages it is important to keep track of which languages are being mapped to. Notice that in the mapping we have had to substitute a continuous solution $(\alpha(t), \beta(t))$ with a discrete one $(\alpha(t_n), \beta(t_n))$, since this is what the numerical method can provide.

Now, in our search for an effectively valid solution all would be well here provided we could solve the difference equations of the Störmer–Verlet method, but we do not know functions that solve these equations either. Given that the system exhibits chaotic behaviour this is not surprising. Adding to the difficulty is the fact that two of the equations (those for $\mathbf{m}_1$ and $\mathbf{k}_2$) are implicit, meaning that the variable we are solving for appears on both sides of the equation. For this reason, the typical strategy is to transform the problem again so that approximate solutions to the Störmer–Verlet equations can be found, making the solution of the problem fully algorithmic in the process. This means writing computer code to implement the Störmer–Verlet method, solving the implicit equations approximately, including code for the Hamiltonian (9). This is generally done in some high-level programming language, such as C, C++, Fortran or Python, or in a numerical mathematics system such as MATLAB or OCTAVE.

To map the problem from the mathematical framework of the numerical method into the framework of a programming language, we must interpret the constraints $\mathcal{N}$ of the numerical method in the synto-semantics of the programming language. This means that the real-valued quantities of the numerical method are interpreted as floating point quantities, meaning finite precision rational numbers with a well-defined error model. This also means that approximate satisfaction of the constraints is judged in terms of floating point arithmetic. For concreteness, let us suppose that we choose C as our programming language, and we have interpreted the equations in $\mathcal{N}$ in C, giving us a programming framework $\mathcal{P}$. Then we need to write C code to solve our problem algorithmically, which amounts to the construction of code for an inference rule in $\mathcal{P}$, which we suppose we store in a file `stover.c`. For simplicity, we will assume that this program takes initial conditions (`a0` and `b0`) and a time interval (`[0,t]`) as input and outputs vectors (`a` and `b`) of solution values (`a[i]` and `b[i]`) over the given time interval. If we suppose that the software when run returns vectors `v` and `w` for the angles of the two weights, then internally to $\mathcal{P}$, we could write

$$\texttt{a0 = pi/2,b0 = -pi/2} \Big\|\!\Big|_{\mathcal{P}}^{\sim} \texttt{a = v,b = w},$$

where `pi` is a machine approximation of π, to express that the software computes what it is supposed to compute, namely that `v` and `w` contain effectively valid values of the state of the double pendulum according to the software version of the dynamics specified in $\mathcal{P}$.

An alert reader will recognize that there is something missing in the story as presented, since the code `stover.c` does not provide us with solutions either. It provides *code* for an inference rule, but not an inference rule itself. Thus, to obtain an inference rule we need to compile it into machine code so that a processor can compute solutions in binary, which are then converted back into floating point numbers to fill the vectors `v` and `w`. Thus, there is actually another synto-semantic transformation required to solve the problem fully. Since it does not serve us to consider this in detail here, we will just treat the compilation and running of the code as happening within the software framework.

Considered as an implementation of our original problem in a software environment, we then we can express the same core content as the above displayed expression in terms of the language of our original problem with[29]

$$\alpha_0 = \pi/2, \beta_0 = -\pi/2 \Big|\Big|\!\!\overset{\sim}{\underset{\mathcal{H}\to\mathcal{N}\to\mathcal{P}}{}} \alpha(t_i) = v_i, \beta(t_i) = w_i,$$

where v_i and w_i are respectively the i-th components of the vectors $\mathbf{v}$ and $\mathbf{w}$, and $\mathbf{v}$ is mapped to `v` and $\mathbf{w}$ is mapped to `w` in the external synto-semantics. The truth of this statement expresses that we can obtain an effectively valid solution to our problem *in terms of the synto-semantics of the software system*, or in other words an effectively valid solution to the constraint system $\mathcal{P}$. But this is not what we really care about, since we are interested in a solution to our original problem $\mathcal{H}$. Thus, what we really want to know is whether when we *back-interpret* this solution into the original framework we obtain an effectively valid solution to the original problem.

Since the software implementation is designed to solve the Störmer–Verlet equations accurately, if we wrote the code properly, then the statement

$$\alpha_0 = \pi/2, \beta_0 = -\pi/2 \Big|\Big|\!\!\overset{\sim}{\underset{\mathcal{H}\to\mathcal{N}\leftrightarrows\mathcal{P}}{}} \alpha(t_i) = v_i, \beta(t_i) = w_i,$$

which is interpreted in $\mathcal{N}$ will be true, where the $\mathbf{v}$ and $\mathbf{w}$ generated by our software are now interpreted as vectors of real numbers (each floating point value is mapped to its corresponding rational number). The standard of effective validity here is that of approximate solution to the Störmer–Verlet equations for the Hamiltonian H, i.e., $\mathcal{N}$. But what we really want is to be able to back-interpret this solution into the framework of Hamilton's equations and have that statement

$$\alpha_0 = \pi/2, \beta_0 = -\pi/2 \Big|\Big|\!\!\overset{\sim}{\underset{\mathcal{H}\leftrightarrows\mathcal{N}\leftrightarrows\mathcal{P}}{}} \alpha(t_i) = v_i, \beta(t_i) = w_i,$$

[29] Note that an alternative notation would be to simply write $\mathcal{H} \to \mathcal{P}$ to indicate the original source framework and the framework of synto-semantic interpretation. We are being fully explicit here for reasons of clarity, but since in general we could end up with graphs of synto-semantic mappings, some simplified notation will eventually be required, and the notation $\mathcal{H} \to \mathcal{P}$ need not lead to confusion when the sequence of mappings is clear.

now interpreted in $\mathcal{H}$, come out as true, in which case the approximate solution of the Störmer–Verlet equations also furnishes us with approximate values along a solution curve to Hamilton's equations. This is determined by the properties of the numerical method, which if we have chosen our method well should be the case. Indeed, to have this come out true for as wide a range of initial conditions as possible was why we chose the symplectic Störmer–Verlet method in the first place. Notice the single horizontal stroke indicating we are no longer using an external synto-semantics and external standard of effective equivalence, instead we are back in the original framework. If this statement holds, then we have essentially solved our problem, since we can recover an effectively valid solution to Hamilton's equations, over some time interval, by interpolating the discrete set of values v_i, w_i returned by the software. Suppose we wrote code for an appropriate interpolant, yielding curves $v(t)$ and $w(t)$, then our ultimate solution would be expressed as

$$\alpha_0 = \pi/2, \beta_0 = -\pi/2 \,\|\!\!\frac{\sim}{\mathcal{H}}\, \alpha(t) = v(t), \beta(t) = w(t),$$

where we have now dropped the notation indicating the sequence of mappings that led to the effective solution. Recall that this statement will only ever be locally true, locally to some time interval, since eventually the overall error will accumulate and the error will exceed whatever tolerance we choose.

In this example of scientific problem-solving we have seen how the search for approximate solutions leads to mappings between synto-semantic frameworks of scientific problems *in a way that nearly preserves the structure of the problem*, in the sense of nearly-identical assumptions should yield nearly-identical solutions. Although much of the mathematical (geometric) structure of the problem is preserved in the mapping from the Hamiltonian framework $\mathcal{H}$ to the numerical framework $\mathcal{N}$, this structure is only preserved in a coded manner in the translation to the programming framework $\mathcal{P}$, even more so when the code is compiled to machine language.

If we abstract away from this particular example, we may see that what is really essential in this strategy of transforming the problem to find solutions is the preservation of the *inferential structure*, in the sense the transformation provides an image of the graph of effectively valid inferences in the source framework as a nearly-identical graph of effectively valid inferences in the target, at least locally to some portion of the graph of inferences in the source. The reason this is so is that we only require that following the graph of effectively valid inferences in the target allows us to make effectively valid inferences in the source, the particular content of the target language becomes immaterial. The transformation must make generating solutions easier to be useful, but near-preservation of inferential structure is nevertheless essential for the transformation process to produce approximate solutions to the original problem. Of course, in the case we just considered, all of the transformations are near-content-preserving in a clear way, we point now to a basic structural feature underlying reliable scientific inference that is independent of content.

We can make this notion of inferential structure-preservation precise in terms of stability properties of the transformation between interpreted languages. As was just pointed out, part of what we require in approximate problem-solving is a transformation of the problem that preserves, at least locally, the inferential structure of the source framework, which ensures that solutions in the target solve a problem with effectively the same structure. Thus, assuming we have an effectively valid inference in the source framework $\mathcal{S}$, $\Gamma \Vert\!\!\sim_{\mathcal{S}} p$, then it must be the case that when Γ and p are mapped into the synto-semantics of a target framework $\mathcal{T}$ that p is still an effectively valid consequence of Γ, i.e., it must be the case that

$$\Gamma \Vert\!\!\sim_{\mathcal{S}} p \Rightarrow \Gamma \Vert\!\!\sim_{\mathcal{S}\rightarrow\mathcal{T}} p.$$

Notice that this condition looks formally much like an effective version of the soundness condition in traditional logic. It is not a soundness condition, effective or otherwise, however, since the source language is already interpreted. Thus, the condition on a mapping really has to with the preservation of effective consequence relations. Since this is another kind of stability condition and one that deals with preservation of inferential structure in a mapping to an external synto-semantics, we will call this condition *forward inferential stability*. The term "forward" here indicates the forward direction of the mapping to the target language. Thus, forward inferential stability of a mapping assures us that we land in a target language on a problem having an nearly-identical structure to the problem we had in the source.

Since the reason we have been considering mapping into a target language is to facilitate making effectively valid inferences in the source, a successful mapping between languages for this purpose requires that the mapping be invertible, so that we can import solutions from the target back to the source. To be able to do this, it must be the case that an effectively valid inference made in the target maps back to an effectively valid inference in the source, at least locally to those problems in the source framework of interest to us. In our logical notation this is expressed by

$$\Gamma \Vert\!\!\sim_{\mathcal{S}\rightarrow\mathcal{T}} p \Rightarrow \Gamma \Vert\!\!\sim_{\mathcal{S}} p,$$

which, in logical terms, expresses that an inference that is externally effectively valid is also internally effectively valid. Though this is akin to the condition of completeness in standard logic, it is not a completeness condition for interpreted languages. Since it does imply the ability to map effectively valid inferences *back* along the original mapping to the target language, we will call this condition *backward inferential stability*. A mapping between interpreted languages that is both forward and backward inferentially stable will be called *inferentially stable*, in which case the relation

$$\Gamma \Vert\!\!\sim_{\mathcal{S}} p \Leftrightarrow \Gamma \Vert\!\!\sim_{\mathcal{S}\rightarrow\mathcal{T}} p$$

holds, at least locally.

The condition of inferential stability implies that a target framework presents not only a problem with effectively the same inferential structure, but essentially the *same* problem, so that solving the target problem is essentially the same thing as solving the source problem. Stated another way, the two conditions imply that making inferences in the target framework is essentially the same thing as making inferences in the source, and *vice versa*. Thought of in another way, a forward inferentially stable map is like having an inferential continomorphism (generalized homomorphism) from the source to target,[30] and a backward inferentially stable map is like having an inferential continomorphism from the target to source. Having both together is like having an inferential contisomorphism (generalized isomorphism) from source to target,[31] telling us that inferences in the two languages are effectively equivalent where this condition holds. Thus, when this condition holds, effectively valid reasoning can be done in either language, so that inferences that are easier to make in one language can be mapped over to the other language. Thus, the condition of inferential stability is what allows mapping between languages to be used as a strategy to solve problems approximately. It is important to recognize that, just as for the condition of effective validity, it will only hold locally to some inferential scope, outside of which the mapping will become unstable and the inferences in the two languages will no longer correspond.

What is particularly interesting about the conditions of (forward and backward) inferential stability is that they correspond to conditions on the reliability of approximate reasoning. These conditions must hold in order for approximation methods to yield scientifically meaningful results. Moreover, in many contexts in applied mathematics, particularly in the context of numerical methods, applied mathematicians prove theorems to articulate the conditions under which an approximation method will generate solutions that are nearly-identical to the solutions of the original problem. In numerical methods these are numerical stability theorems, which essentially give conditions under which solutions of the numerical method provide approximate solutions to the original problem.[32] The proof of such a theorem is then actually a backward inferential stability proof. Forward inferential stability for numerical methods is ensured by generating them in terms of some method of approximation of the original problem. In certain cases there is a technical

[30] In categorical terms, this would correspond to some kind of generalization of a functor between categories.

[31] In categorical terms, this would correspond to some kind of generalization of a pair of adjoint functors between categories or a categorical equivalence.

[32] This kind of numerical stability is called *forward stability*, which assures that a method provides an approximate solution. Typically, however, theorems establish that the numerical method provides an exact solution to a slightly modified problem, rather than an approximate solution to the original problem. This alternative stability concept is called *backward stability*. Backward stability results can easily be accommodated by effective logic by adding an equivalence condition to the imposed constraints of a framework, so that the framework is expanded into a family of fixed frameworks, or by leaving any constraints that can be modified as assumptions rather than imposing them globally.

condition corresponding to forward inferential stability that must be met for any potential method, such as the consistency condition for numerical methods for ordinary differential equations.

We pause briefly before moving on to re-emphasize that effective logic is a local logic without locality being built in to the system, in contrast to other alternative logical systems that explicitly introduce contextuality or local truth (e.g., contextual languages, modal logic, topos theory). The basic notion in effective logic is near-structure-preserving variation, which has the effect of converting properties that are traditionally exact and making them sensitive to error and reliable only under certain conditions, validity being a paradigm example. Rather than being something we build into the structure of the system, then, local properties become a direct consequence of approximation. This matches the approach used in science, since it is the need to solve problems efficiently that often leads to seeking out approximations and it is the nature of approximation to make methods only locally valid, applicable or stable. Nevertheless, an consideration of boundaries of validity is crucial for reliable scientific inference, and will be an important subject of consideration in future developments of effective logic.

7 Modeling, Complexity Reduction and Inferential Structure-Preservation

We have now seen how we can understand problem solving methods in science in terms of moving effectively valid inferences between frameworks via near-structure-preserving maps. Moreover, we have seen how this process can be understood in terms of forward inferentially stable maps allowing the movement of inferences from a source framework to a target, and backward inferentially stable maps allowing movement of inferences from the target back to the source. The strict form of validity in traditional logic cannot account for these processes in a *direct* way, i.e., without adding conditions or dressing up valid inference as something else, because effectively valid inferences, being approximate by nature, are generally not strictly valid and the maps between frameworks are also only structure-preserving in an approximate sense.[33] Effective logic accounts for problem solving strategies in science by showing how a precise (generalized) logical structure can nevertheless obtain in scientific methods that involve approximations.

As we have noted, the approach of seeking inferentially stable mappings between problem-solving frameworks is very common in applied mathematics, where solving a problem in the framework in which it is originally posed often

[33] As indicated in the introduction, this is not to say that traditional logic cannot provide approximate representations of effectively valid inferences, or even useful models of effectively valid inferences, just that traditional logic cannot regard valid inference as fundamentally approximate and conditionally stable.

proves to be very difficult, prompting a search for effectively equivalent problems that are easier to solve. Examples of these methods include asymptotic analysis, perturbation theory and numerical methods, which covers quite a large range of mathematical methods. This kind of method also underlies the ubiquitous approach in pure mathematics, e.g., in algebraic geometry and algebraic topology, of solving problems in one framework or category by transforming to equivalent problems in another framework or category, though in this case near-identity is based upon exact structure-preservation. It was also mentioned above that this kind of method underlies the modular methods used to accelerate symbolic computations. Thus, effective logic stands to be able to capture the natural reasoning processes of a large portion of science, by capturing the *structure* of the reasoning that scientists use in their own languages. This contrasts distinctly with the traditional approach of reconstructing theories by casting them in a uniform formal language or system, and suits the scientific purposes for which we are developing the concepts of effective logic.

Moreover, as is shown in [13], the methods of computational science, including both numerical methods and symbolic computation, rely on transformations between problems to reduce the computational complexity of mathematical problems sufficiently so as to make them rapidly computable in practice. The epistemological drive underlying these methods is the need to overcome obstacles to making efficient, reliable inferences given the contextual constraints of scientific practice. We can call such efficient, reliable inference *feasible inference*. These feasible inference methods must preserve the inferential structure of the problem to be able to generate solutions that can potentially correspond to solutions of the original problem, and they must be invertible so that the computed solutions can actually produce a solution to the original problem. Thus, it is seen that strategies of complexity reduction in computational science rely on inferentially stable transformations of mathematical problems in a way that makes their solution computable rapidly.

A consequence of this observation is that if all that is required for an accelerated algorithm is that it reduce computational complexity and preserve inferential structure, then little or none of the mathematical content of the problem *needs* to be preserved in the transformation, provided that solutions to corresponding problems correspond. Thus, despite the fact that it is natural and standard to look for solutions to problems by transformations that nearly-preserve their mathematical structure in some way, there may nevertheless exist transformations with even lower complexity that preserve little or no mathematical structure at all, yet nevertheless deliver efficient, reliable solutions. It was argued above that the successful reduction of problems to machine language is an illustration of this kind of idea, but the increasingly popular problem-solving methods based on machine learning algorithms perhaps provide a more compelling kind of case. Here there is no explicit near-structure preservation at all, rather the algorithm learns how to solve a problem in some iterative fashion through a more or less opaque set of transformations. Inferential structure is nevertheless being nearly-preserved by successfully trained algorithms over some range of cases or conditions, as is evidenced by the many achievements of machine learning algorithms, such as their well-known applications to image classification and language processing problems.

The notion of inferential stability also provides a new way of thinking about the abstraction processes involved in mathematical modeling. From the perspective of effective logic, a mathematical modeling problem begins with the desire to make effectively valid inferences about the behaviour, or the reasons underlying the behaviour, of some phenomenon. Such inferences are usually formulated in some scientifically-augmented form of natural language.[34] We wish to be able to make inferences about properties or states of a system or phenomenon, sometimes with very tight tolerances on error. Since we typically cannot make these inferences in our scientifically-augmented natural language, we resort to mapping to some other, usually mathematical, language that we expect will assist us in making the desired inferences. We do this so that by making corresponding inferences in the scientific language, we can map the conclusions back to our (scientifically-augmented) natural language to yield descriptions, predictions and explanations of behaviour of the phenomena. This is to say that we require the mapping from the natural language to the scientific language to be inferentially stable. When the mapping has this property, we can rely on inferences made in the target scientific language to be informative about the world.

We can therefore understand the mathematical modeling process as overcoming an inferential obstacle to drawing conclusions about the structure or behaviour of some natural phenomenon, conclusions that are not accessible without the use of scientific theories or mathematics. We use models, then, to facilitate reasoning processes that are not feasible directly. Thus, mathematical modeling can also be seen as a strategy of problem transformation that makes inference feasible. Moreover, we have seen that for this process to be successful and reliable, the mapping from the description of the phenomenon using natural or operational/experimental language to the language of the mathematical model must be inferentially stable, so that the conclusions drawn in the model give reliable conclusions about the phenomenon. Consequently, we may see mathematical modeling procedures as tools for reducing the *inferential complexity*, i.e., the cost of drawing inferences, for description, prediction and control of natural phenomena by transforming between languages. Furthermore, just as for computational complexity reduction strategies in computational science, a key requirement is inferential stability.[35]

[34] Specifying the semantics for such expressions is a notoriously difficult problem, but one faced by any attempt to account for our descriptions of the world. Accordingly, I will not consider this matter here except to point out that a semantics often relies on experiential states, states of experimental apparatus or some canonical, and maximally scientifically neutral, physical model of phenomena.

[35] We note here that the observation above concerning the fact that only the inferential structure, not the content, needs to be preserved in transformations, can be applied the modeling case. This has interesting philosophical consequences for how we might understand scientific representation, since the effective logic model is consistent with a plurality of aims among scientists, some of whom will be interested in direct descriptions of the structure of some part of the world and others content with empirical adequacy. Any further consideration of these issues is beyond the current scope.

With effective logic, therefore, we obtain a picture of the mathematical modeling process that accounts for the kinds of methods used throughout the entire process, including computation, and a picture that is, or can be, fully compatible with the actual methods that practicing scientists use. This is the advantage of having a form of description that can map on to scientific language, capturing its basic structure, rather than requiring a mapping of scientific language into a logical language in order to reconstruct it. The task when using effective logic for philosophical purposes, then, is to ensure that it does indeed capture the structure of inference in scientific languages. I have only presented a limited amount of evidence for the representational capacity of effective logic in this paper. Though a more fully developed argument is reserved for future work, further evidence that effective logic captures the structure of scientific inference in practice is provided in [12, 13], which together show a common inferential pattern across parts of pure mathematics, applied mathematics, computational science and data handling in applied science (specifically optical astrometry for the orbit determination problem). It is important to notice, however, that though effective validity can only capture nearly deductive inference within a language, the ability to use inferentially stable mappings between languages allows one to make inferences that do not remotely resemble deductive inference by appealing to radically different languages that nevertheless allow one to complete effectively valid inferences by mapping back to the original, source framework. Effective logic therefore stands to capture some very general structural patterns in scientific inference.

8 Conclusion

In summary, I have presented a generalized logic based on the concept of effective validity that stands to account for the basic structure of inference in scientific practice, to clarify the structure of reliable methods of computational complexity reduction in computational science, and to provide an account of the mathematical modeling process that views modeling methods as tools of reliable inferential complexity reduction. Such an account emerges naturally from regarding scientific reasoning procedures in terms of inferentially stable mappings between languages.

As it has been presented, this generalized logic functions to capture the basic form of scientific inference as it occurs in real scientific languages. It is opposite in approach to the traditional strategy of representing scientific inference through reconstructions of scientific languages in some formal language. Rather, effective logic employs an (epistemological) modeling approach in the sense that it captures the structure of inference in particular interpreted languages rather than requiring treatment in some specialized uninterpreted formal language or large class of models. At the same time, it is complementary to traditional rational reconstruction because it is well-suited to a very different problem, viz., mapping the inferential structure of scientific practice. With a more flexible, error sensitive notion of validity, it becomes possible to faithfully and directly capture a wider range of scientific

inference. Because effective logic is concerned with developing conceptual tools that reveal the underlying structure of methods in scientific practice, I believe that through further development it has the potential to produce insights into the reliability of scientific languages and to cope with the potential for instability in reasoning involving error.

By extending the traditional notion of valid inference into a context of variation, we open up logic to a treatment of the forms of inference typical in the approximation methods used in mathematical analysis, in contrast to traditional logic which more closely suited to the forms of inference typical in the exact methods of abstract algebra. At the same time, it opens up logic to an accurate treatment of the forms of inference in the mathematical modeling process and potentially to scientific inference more broadly. We have seen how the introduction of a context of variation can lead to different kinds of mathematical questions, such as the stability of consequence relations, or even mathematical proofs, under certain (near-identity) transformations of the syntax. Nothing here is strictly new, since there already exist forms of each of these things within traditional logic, and traditional logic can surely illuminate all of these things in its own terms. The difference with effective logic is that we move toward a natural language for approximate inference, which stands to introduce fresh and illuminating perspectives on old problems while also suggesting new kinds of questions and directions of inquiry.

Acknowledgements The author would like to thank David Stoutemyer, Chris Smeenk, Erik Curiel and an anonymous reviewer for their valuable comments.

References

1. Batterman RW (2002) The devil in the details: asymptotic reasoning in explanation, reduction, and emergence. Oxford University Press, Oxford
2. Corless RM (1994) Error backward. In: Kloeden PE, Palmer KT (eds) Chaotic numerics: an international workshop on the approximation and computation of complicated dynamical behavior, Deakin University, Geelong, Australia, July 12–16, 1993, vol 172. American Mathematical Society, Providence
3. Corless RM, Fillion N (2013) A graduate introduction to numerical methods. Springer, Berlin
4. Fillion N, Corless RM (2014) On the epistemological analysis of modeling and computational error in the mathematical sciences. Synthese 191(7):1451–1467
5. Fillion N, Moir RHC (2018) Explanation and abstraction from a backward-error analytic perspective. Eur J Philos Sci 8(3):735–759. https://doi.org/10.1007/s13194-018-0208-6
6. Goldstein H, Poole CP, Safko JL (2002) Classical mechanics. Addison-Wesley, Boston
7. Hairer E, Lubich C, Wanner G (2006) Geometric numerical integration: structure-preserving algorithms for ordinary differential equations. Springer, Berlin
8. Harper WL (2011) Isaac Newton's scientific method. Oxford University Press, New York
9. Hempel C (1965) Aspects of scientific explanation, and other essays in the philosophy of science. Free Press, New York
10. Lawden DF (1989) Elliptic functions and applications. Springer, Berlin
11. Moir RHC (2010) Reconsidering backward error analysis for ordinary differential equations. MSc Thesis, The University of Western Ontario

12. Moir RHC (2013) Structures in real theory application: a study in feasible epistemology. PhD Thesis, The University of Western Ontario
13. Moir RHC (2018) Feasible computation: methodological contributions of computational science. In: Cuffaro ME, Fletcher SC (eds) Physical perspectives on computation, computational perspectives on physics. Cambridge University Press, Cambridge, pp 172–194
14. Oppenheim P, Putnam H (1958) Unity of science as a working hypothesis. Minn Stud Philos Sci 2:3–36
15. Suppe F (1974) The structure of scientific theories. University of Illinois Press, Chicago
16. Suppes P (1969) Models of data. Springer, Berlin
17. Suppes P (2002) Representation and invariance of scientific structures. CSLI Publications, Stanford
18. van Fraassen BC (1970) On the extension of Beth's semantics of physical theories. Philos Sci 37(3):325–339
19. van Fraassen BC (1980) The scientific image. Oxford University Press, Oxford
20. van Fraassen BC (2006) Representation: the problem for structuralism. Philos Sci 73(5):536–547
21. Voevodsky V et al (2013) Homotopy type theory: Univalent foundations of mathematics. Institute for Advanced Study (Princeton), The Univalent Foundations Program, Princeton
22. Wallace D (2006) In defence of naiveté: the conceptual status of Lagrangian quantum field theory. Synthese 151(1):33–80
23. Wilson M (2006) Wandering significance: an essay on conceptual behavior. Oxford University Press, Oxford
24. Wimsatt WC (2007) Re-engineering philosophy for limited beings: piecewise approximations to reality. Harvard University Press, Cambridge

Counterfactuals in the Real World

James Woodward and Mark Wilson

Abstract Following Jacques Hadamard, applied mathematicians typically investigate their models in the form of *well-set problems*, which actually consist of a family of applicational circumstances that vary in specific ways with respect to their initial and boundary values (and other forms of "side condition"). The chief motive for investigating models in this wider manner is to avoid the improper behavioral conclusions one might reach from the consideration of a more restricted range of cases. Suitable specifications of the required initial and boundary variability typically appeal to previously established experimental conclusions as to how the target system will behave under a range of eternally applied manipulations of the form "If the conditions pertaining to S were altered in manner M, internal features X would/would not alter" (such claims are called *manipulation counterfactuals* in the essay and arise in a variety of distinct forms). In his investigations of causal reasoning within other parts of science, our first author (Woodward) has emphasized the conceptual importance of counterfactuals of this nature, for which he was been often criticized by authors of a self-styled "metaphysical" inclination. The purpose of this note is to argue, *pace* these objections, that closely analogous considerations have long been part of the practice of investigating differential equation models in a sensible way.

J. Woodward · M. Wilson (✉)
University of Pittsburgh, Pittsburgh, PA, USA
e-mail: mawilson@pitt.edu

© Springer Science+Business Media, LLC, part of Springer Nature 2019

N. Fillion et al. (eds.), *Algorithms and Complexity in Mathematics, Epistemology, and Science*, Fields Institute Communications 82,
https://doi.org/10.1007/978-1-4939-9051-1_10

1 Counterfactuals in Scientific Use

The pioneers of analytic philosophy, such as Gottlob Frege and Ernst Mach, viewed any form of modal concern as a blemish upon the progress of science and studiously sought to avoid any appeal of a seemingly fictive character. They didn't want the progress of science to become retarded by complaints such as "Your Minkowskian approach to space time must be inadequate because faster-than-light signals are clearly possible counterfactually." The doctrine that Carnap and Quine later dubbed "the thesis of extensionality" partially stemmed from these legitimate liberation-from-irrelevant-possibility concerns. However, it was eventually recognized that these militant strictures are too harsh, and that certain forms of counterfactual construction (such as "if this material were placed in water, it would dissolve") are plainly wanted in science. Indeed, Goodman motivates his celebrated study of counterfactuals [1] by citing the importance of dispositional claims of this general character. However, attention to dispositions alone can prove misleading, for such limited exemplars tacitly suggest that the project of tolerating modal appeals within science should lie entirely in the direction of reducing such claims to some substratum of occurrent fact, such as the target system's actualist properties, and the laws of nature, also conceived along actualist lines. This was the project on which Goodman himself embarked in *Fact, Fiction and Forecast*, and subsequent philosophical research on the "modality in science" problem has largely adopted this paradigm.

But there are subtler forms of modal appeal within science that do not point in a "reductive" direction at all, and suggest that other collections of counterfactuals can facilitate an opposing scientific goal, that of *insulating* productive science from undue reliance upon unverified lower scale speculations of a "reductive" flavor. Our objective in this paper is to bring some of the appeal of not attempting to treat counterfactuals in a reductive fashion into sharper focus. Our specific examples will highlight the manners in which carefully culled collections of manipulationist counterfactuals can heighten the reliability of a modeling scheme (here, by a "manipulationist counterfactual" we mean formula of the sort, "If a controlling variable A were experimentally set to a value α, behavior B would be the result"). The non-reductive utilities of such appeals are rarely discussed within the philosophical literature, despite the fact that these opportunities are frequently exploited within working science. Here we bring these considerations to more overt attention.

Before we turn to our examples, let us offer a few prefatory remarks on our purposes here. Following Goodman's original discussion, the majority of writings on counterfactuals have presumed that philosophers must offer a "general theory of counterfactuals" adequate to every intelligible instance of such conditionals, including unconstrained claims such as "if Caesar had been in charge of UN forces during the Korean War, he would have used catapults."[1] We see no reason why this should be the case. In itself, the employment of the subjunctive mood merely alerts an audience to the fact that something fictive is or may be afoot. But the byways of pretense are incredibly varied, and any discussion of their focal purpose requires a delicate discussion of the specific authorial intentions behind the counterfactual act (if they can be pinned down at all). Assertions like "if Sherlock Holmes hadn't smoked a pipe, he could have still deduced that a snake was involved in the Case of the Speckled Band" can sometimes be reasonably adjudicated and sometimes not. But conditionals like the ones just described do not display any intimate connection with the specific counterfactual utilities we shall highlight here, whose truth-values can be frequently established through simple induction from experiment. Immersing our specific classes of isolating counterfactuals within a wider ocean of unrelated fictive projects has the unwanted effect of obscuring the characteristic scientific purposes that our specific types of counterfactuals serve to advance. It is only the latter we care to explicate, not the wider musings about Caesar and Sherlock Holmes.

The bulk of this essay will focus upon an important notion that applied mathematicians now call, following the guidance of Hadamard's ground-breaking

[1]How do we determine whether this claim is true or whether Caesar might have used atomic weapons instead?

work ([2], originally published in 1923), "well-set problems."[2] Contrary to its singular denomination, a "well-set problem" actually consists of a carefully selected *family* of problems: a collection of mathematical claims that differ from one another through alterations in their attached "side conditions" (a notion that we'll explicate fully in a moment). Most of these "claims with altered side conditions" are, in fact, counterfactual in character: they consist in assertions of the form "If system conditions S were altered in manner A, condition B would result." Nowadays, no modern text in engineering or physics goes forward without specifying the proper "setting" of its modeling proposals (where a "proper setting" consists in a precisely specified set (or "function space," in the usual lingo) of associated counterfactual variations). But why do these modelers believe that fleshing out their models in a counterfactually amplified way is vitally important? The baseline answer we shall develop: the supplementary data helps insure that the model in question has been soundly formulated and will not be subject to deceptive artifacts. Later we'll find that further methodological virtues can be added to this list: associating a problem with a cannily selected counterfactual family can direct a practitioner to very significant forms of data simplification. These are not "exercises in reduction" in any plausible sense; they typically represent assurances that one doesn't need to worry about superfluous lower scale details.

As it happens, one of us (Woodward) has investigated the tacit requirements that experimenters employ in determining whether their data indicates the presence of a causal mechanism within the system S under investigation. Woodward has determined that faith in a causal model often directly depends upon the experimenter's trust that her data supports a family of counterfactual variations rather like the ones we shall investigate in this essay. And the underlying motivations are much the same: the counterfactual appeals guarantee that the model is not prey to spurious defects (such as confounding by "common causes"). But this analysis has been criticized as "circular" by the modern metaphysicians who seek a "general account of counterfactuals" in the all-embracing mode that we have spurned. "tis circular," these critics contend, "because counterfactual claims must be *grounded* in underlying laws of nature, and these laws need to be accepted before Woodward's counterfactuals can be credited with appropriate truth-values. From a metaphysical point of view, therefore, we should look directly to the underlying laws and bypass the irrelevant counterfactual go-betweens."

[2]Sometimes the phrase "well-posed" is employed as an alternative, but it invites an ambiguity that we'd prefer to skirt. In the case of standard initial-boundary value problems, Hadamard lays down three basic conditions: (1) solutions will exist for a certain span of time; (2) they will prove unique; and (3) they will demonstrate behavioral stability under suitable norms, selected according to various further criteria. Often the term "well-posed" focuses upon Hadamard's third criterion, which we will not discuss further due to its associated technicalities. But similar morals pertaining to counterfactuals will apply here as well.

We scarcely know what to make of these peculiar claims, for they do not map onto actual scientific practice in any evident way. Nor do we adequately understand what the demand for a "grounding" entails. However, it is evident that distractions of this amorphous character have steered philosophical attention away from the practical modes in which carefully monitored forms of counterfactual appeal assist productive science, in manners that cannot be reduced to simple dispositional analysis. In our opinion, Woodward's metaphysical critics have counseled the unwary, "Look away, for there's nothing interesting to be seen here." But it would be rash to accept such advice.

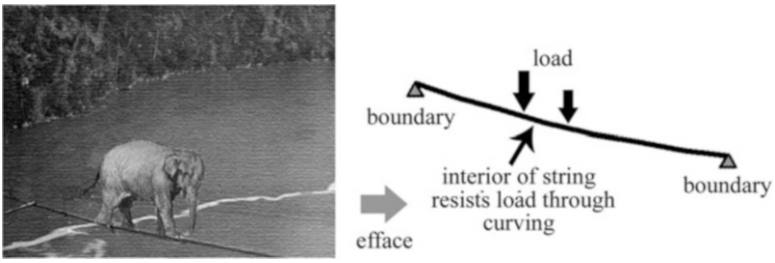
A significant source of these distractive policies stems from the fact that many philosophers employ the standard terminologies of applied mathematics in extremely loose ways. Such descriptive practices represent the unfortunate logicist heritage of Carnap, Hempel, Quine and their many "possible world" successors. In point of origin, the proper mathematical parsing of central notions such as "initial condition" and "boundary condition" trace to Hadamard's original work on well-set problems, and a significant part of the diagnostic labor we shall undertake simply returns these vocabularies-gone-astray to their properly focused contours, as well as articulating Hadamard's motivations in framing these discriminations. Our second author (Wilson) has complained fulsomely of other woes afflicted upon modern philosophy through the persistent misuse of the standard classificatory terminologies of applied mathematics (he calls such abuses "theory T thinking"—see [3]). Excessively simplified pictures of scientific methodology thrive upon fuzzy categorizations, and Wilson believes that the "metaphysical" undervaluing of manipulationist counterfactual data qualifies as a case in point.

2 Well-Set Problems

As just noted, many philosophers currently misapply the mathematician's terms "initial conditions" and "boundary conditions" in loose and unconstrained ways, often embracing modeling ingredients that fall under neither category. So let us now return these notions to their original, as-explicated-by-Hadamard contours. Doing so immediately opens the doors to a better appreciation of our restricted classes of associated counterfactuals.

To consider what a "well-set problem" demands, let's adapt an old example of Eddington's [4]. An elephant walks across a tightrope over Niagara Falls. For the purposes of understanding how the rope internally responds to this duress, a modeler only needs to boil down the elaborate details of this elephantine loading into an upper surface distribution of downward force. Pachyderm and scenery vanish from view, replaced by a simple schedule of downward arrows upon the string. For simplicity, let us suppose that the elephant remains in the same position on the rope over the time interval we are interested in and that the rope is perfectly straight at time t_0 (of course, it will immediately start sagging because of the elephant's load). We shall also presume that the two far endpoints of the rope remain completely

immobile. Properly applied, the rope's straight condition at time t_0 represents the problem's assigned *initial condition* and the requirement that the endpoints remain fixed at all times later than t_0 represents its assigned *boundary conditions*.[3]



At first glance, these two forms of stipulation may look very similar except that initial conditions articulate facts that obtain on a specific time slice t_0 (the rope is entirely straight at t_0), whereas the boundary conditions dictate what happens on a spatial-slice (i.e, the two end point locations remain fixed for all $t \geq t_0$). But when we turn to the kinds of demands we want to place upon our modelings, we find that they characteristically assume significantly distinct characters. Distinct classes of associated counterfactuals correlate closely with these sharply differentiated requirements, so that it is important to distinguish between the very different roles played by initial and boundary conditions. Later in the paper we'll discuss a third class of specialized counterfactuals that emerge in connection with the variety of "side condition" that mathematicians label as *constraints*.

In particular, a successful modeling should allow us to settle a wide range of questions about what will happen if some small perturbation is initially introduced somewhere, e.g., by a gremlin hitting the cord with a brisk hammer blow far away? Will significant waves then travel to the feet of our unfortunate elephant? Will it lose its balance in that happenstance? And so forth. In the usual jargon, the "initial conditions" encountered within a well-set problem should be *freely assignable*—we want to consider what might happen to our elephant + rope system over a generous set of potential starting conditions at time t_0. This is why we claimed earlier that a "well-set problem" really represents a large collection of related individual problems, differing in the initial conditions with which they begin. Why do we do this? We'll later find that we haven't *understood* the inner workings of our model

[3]Technical remark: due to the collapsed one-dimensionality of this reduced modeling, the elephant loading itself is usually classified as a *forcing condition*, rather than a "boundary condition" per se. When we instead model our string as a two or three-dimensional solid, the loading converts to a straightforward boundary condition. For most conventional solids and liquids, their exterior bounding surfaces supply suitable opportunities upon which a worthy policy of what we later call "E versus I effacement" can be reasonably effected.

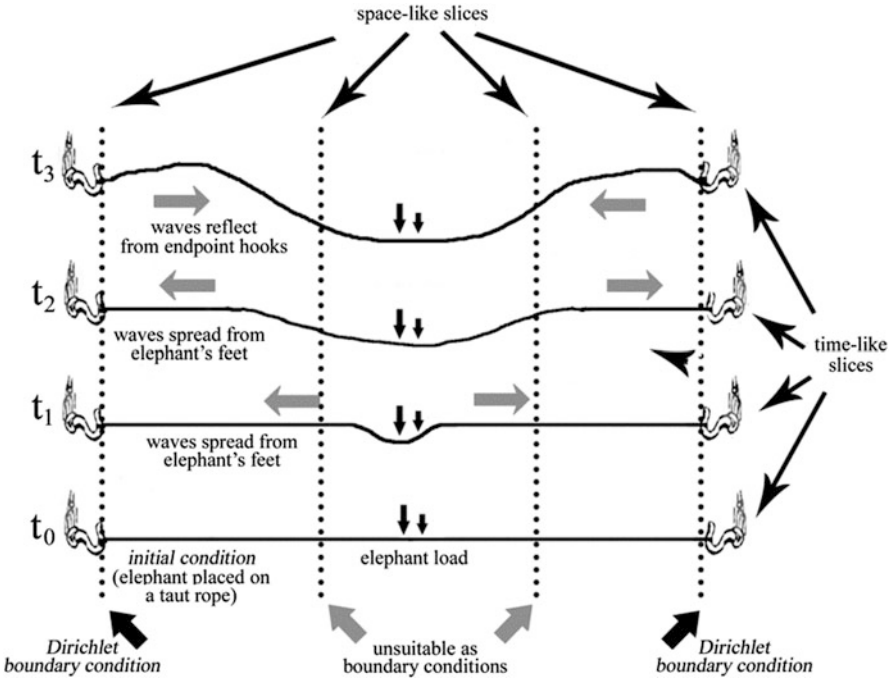
properly until we can address such concerns effectively.[4] Note that we are typically interested only in the restricted family of counterfactuals that feature freely varied conditions on the time slice t_0, not at later times t_1.

In contrast, when we turn to *boundary conditions*, we find that we are rarely concerned with "freely assignable" conditions in the foregoing manner, but instead are concerned with the special collection of models in which those end conditions remain exactly the same as originally articulated (the stipulation that endpoints remain fixed is officially designated as a *Dirichlet boundary condition*, and applied mathematicians typically investigate such problems first before they turn to more elaborate boundary region behaviors). The formal analog to "freely assignable initial conditions at time t_0" would be "freely assignable boundary behaviors occurring at the spatial location l_i for all times $\geq t_0$", and we can rarely anticipate ahead of time what behaviors may occur at the location l_i if we select an l_i *inside* the rope. So why can we do better with respect to the two special end point locations l_L and l_R? Because we know ahead of time that they will remain fixed (unless the elephant somehow breaks the rope). Accordingly, suitable choices of boundary condition seek out descriptive stipulations that we can safely presume will continue to hold over a suitable span of future times. We presently know that these ropes are firmly attached to the hooks and rock and won't wiggle much at l_L and l_R no matter how wildly the interior rope lashes about (mathematicians say that we know this "on an *a priori* basis"). In contrast, we can't usually know ahead of time what will happen at on an arbitrary interior location l_i until we have actually *solved* our modeling problem. To assign suitable boundary conditions to a modeling, we must seek the privileged locales that offer opportunities for what [3] calls "*system versus environmental effacement*" (we'll sometime abbreviate this mouthful as "S versus E effacement"). In certifying that our rope remains fixed at l_L and l_R for all future times, we can guarantee that any wave traveling along the rope will be forced to reflect back into the cord's interior to conserve energy.[5] If so, the fact that we lack the details of what transpires inside the rock + hook environment to which our rope is attached (= our system's immediate environment E) won't greatly hamper our ability to augur the future behaviors of the rope + elephant interior. This is

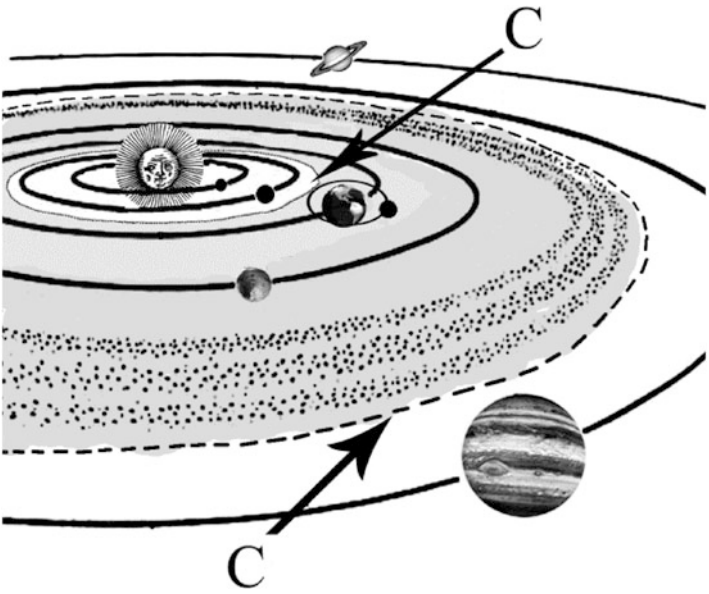
[4]This "free variability" is closely dependent upon a suitable collection of parallel boundary conditions choices. If under most starting conditions, the rope breaks shortly thereafter, Dirichlet-style counterfactuals that begin "if the rope were to stay straight at time t_1, then ... would happen" will no longer seem germane to the modeling situation before us and needn't display firmly established truth-values.

[5]If we instead know that such energy will leak out of the rope at a particular rate, we encounter a different category of boundary condition called a *Neumann condition*. The critical feature remains that we can ascertain the leakage rate properties of the endpoints on an *a priori* basis. Other forms of *a priori* assurance are also considered in applied mathematics (they are collectively labeled as "side conditions" to the interior differential equations), including interfacial stipulations and the constraints we shall later survey.

the sense in which our *a priori* trust in a fixed end point boundary specification allows us to *efface* the interior behaviors of our modeling away from the greater complexities of the external environment beyond. Analytic metaphysicians are fond of declaring that they aim at "carving nature at the joints." Insofar as these ambitions can be correlated with sound scientific practice, the locales of effective S versus E effacement represent the "joints" that mathematical modelers attempt to capture within their well-set problems. To be sure, these opportunistic "cuts" are not be as absolutist as the "joints" that the metaphysicians seek, but they represent objective facts about nature of central importance to descriptive science [3]. classifies these as "descriptive opportunities": trustworthy facts that we should exploit in rendering a mathematical modeling tractable.



A related term that modelers sometimes employ to emphasize allied consideration of S versus E effacement is "*cut*." In modeling the complex behaviors of the faster inner planets S, workers in celestial mechanics often freeze the sun and outer planets into fixed positions and ask how the smaller bodies will whisk about within the environmental "field" E generated by their slower neighbors. Once these answers are reached, the modeler can then "turn back on the dynamics" of E and ascertain how changes in E will affect their previous interior system S answers. This "fast versus slow times" technique is commonly described as "cutting the S behaviors away from the greater complications of its surrounding environment."



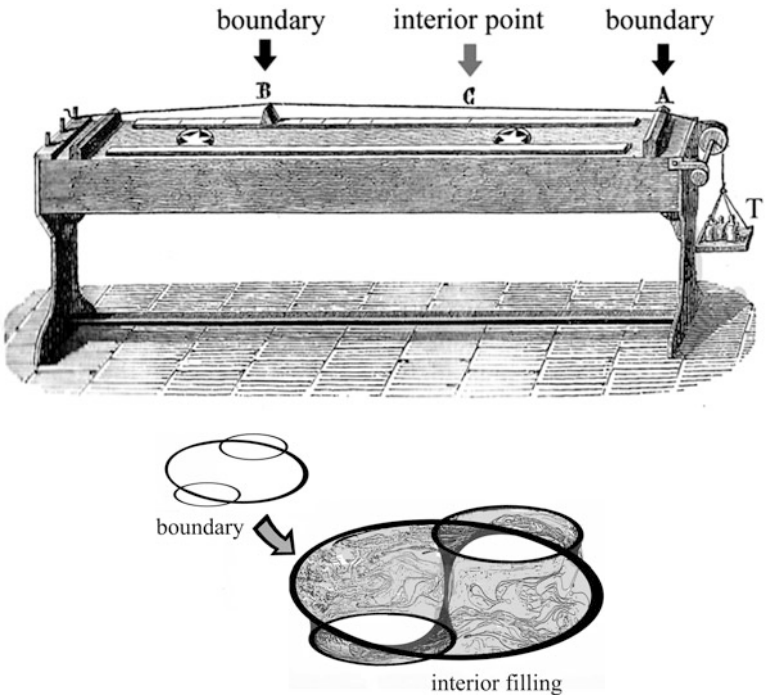
Commonly our faith in these pre-established, time-forward conditions stems from direct experiment and/or everyday experience: we've loaded a lot of ropes with elephants and found that, under mild conditions, the endpoints remain fixed. These are matters that we can readily verify in a well-planned experiment: twiddle with the rope's exterior circumstances and verify that "in all the manipulations we have tested, the endpoints of this rope remain fixed."[6] To ward off unwanted artifacts, we shall want to extend these conclusions to the counterfactually extended family: "If the rope were subjected to any mild form of external manipulation, its endpoints would remain fixed." In a moment, we'll explicate this need "to ward off unwanted artifacts" requirement further.

Before plowing ahead, let us observe all of the preceding discussion pertains equally to a violin string attached firmly at bridge and nut (Dirichlet boundary conditions) obeying the standard wave equation $\partial y^2/\partial t^2 = c^2 \partial y^2/\partial x^2$ within the string interior. We have fabricated our rope + elephant example to render more vivid the policies of S versus E effacement that are always involved in articulating well-set modelings of this general type.

Summarizing our discussion to date, we have found that a "well-set modeling problem" in applied mathematics consists of a family of problems that are intimately

[6]As remarked earlier, many contemporary metaphysicians maintain that a "grounding in laws of nature" is required to support these assertions, but this rhetoric suggests an intellectual quest that rarely takes place. To be sure, there are probably lots of "laws" that govern the internal behavior of rocks and hooks in ways that can further explicate in great detail why ropes can be more firmly fastened to rocks than jello. But we rarely delve into such ancillary concerns when we worry about elephants on tightropes. Merely knowing from experiment that the rope will hold on both ends usually suffices for our purposes.

connected with two distinct collections of attached counterfactuals: (1) the freely assignable collections of allowable *initial conditions* and (2) the contrasting set of counterfactual guarantees that assure us that we needn't worry about unwanted variations occurring in the future in the special locales where *boundary conditions* can be confidently laid down. We maintain that any adequate account of "modality's uses within science" should explain why we approach these two classes of modal requirement in such markedly different ways. In the next section, we will review the insightful answers that Hadamard provided within his original studies of well-set problems.



Before we do so, let us mention another significant consideration that Hadamard has brought to our attention. Nature supplies a wide variety of well-set problems that differ from one another significantly in terms of their underlying strategic architecture. Consider the task of determining how a child's soap film will distribute itself across the interior of a twisted wire frame, possibly broken into several disjoint pieces. Mathematicians call this a "pure boundary value problem" and do not assign any initial conditions in the proper "time slice at t_0" sense to the problem at all. Why? Because we are usually interested in the final state configurations that the soap may reach once it stops jiggling about and settles into equilibrium (several end states may fulfill this condition). In these circumstances, the question of how the film was initially applied to the wire rim becomes irrelevant; we only require a guarantee that the shape of the boundary doesn't alter over the "relaxation time" of the soap film. Formulated as "well-set problems" only boundary conditions appear as required

side conditions. Specification of initial conditions is completely unwanted in such a setting, and acceptable answers no longer need to be unique. This is why explanatory architectures of this complexion are standardly labeled as "pure boundary value problems."[7]

But what accounts for the discrepancies in side condition requirement between our ropes and strings and the soap bubble? These features trace to the fact that the *equilibrium states* forthcoming in the soap case substantially improve our capacities to control the target system through exterior manipulations. As long as we can patiently wait until our soap settles down, we can control its final configuration by simply fixing its bounding wire appropriately. In contrast, we can freely manipulate our elephant and rope circumstances only at the initial time t_0, leaving everything else to the ways in which the interior physics of the situation will autonomously unfold over time. In exerting this more limited species of initial control, we will also need to be very careful about the *initial velocity* we impart to our elephant: careless loadings will send myriads of wild waves streaming along the rope. If an equilibrium state is forthcoming, we can be far more cavalier about how we initially deposit the soap upon its surrounding frame. The centrality of "do statics first" policies within standard mechanical pedagogy stem from the improved reliability of experimental tests that only establish controlled equilibria without attending to the messier details of how the system winds up in that state.[8]

We mention these alternative varieties of well-set problem now, because we will later argue that conventional philosophical thinking about counterfactuals do not adequately explain why the attached classes of salient counterfactuals shift rather dramatically when we move into an adjusted form of explanatory format.

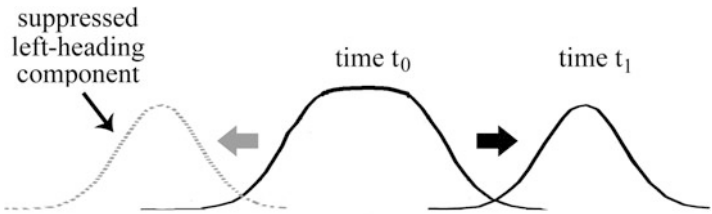
3 Motives for Counterfactual Family Enlargement: Initial Conditions

Why should applied mathematicians want to examine their models within widened classes of freely assignable initial conditions? *Answer*: to filter away misleading conclusions that trace to artifacts of the particular initial conditions assigned, rather than reflecting true features of the target system S internal responses. Here's a typical, if somewhat artificial, Hadamard-like illustration involving the wave

[7]In the jargon of the mathematicians, the side conditions appropriate to a well-set *elliptic problem* (the soap film) differ from those appropriate to *hyperbolic circumstances* (the elephant on a rope). For more details on these distinctions, see any standard text on partial differential equations and/or [3]. It should be remarked that engineers commonly approach mild elephant-on- a-rope problems in an equilibrium-centered manner in which they only attempt to ascertain the final shape of the rope after the elephant has settled into quiescence, not attempting to ascertain how it will bounce around beforehand. In such circumstances, the associated well-set problem becomes elliptic.

[8]Consult the essay on Pierre Duhem in [3] for more on these experimental advantages.

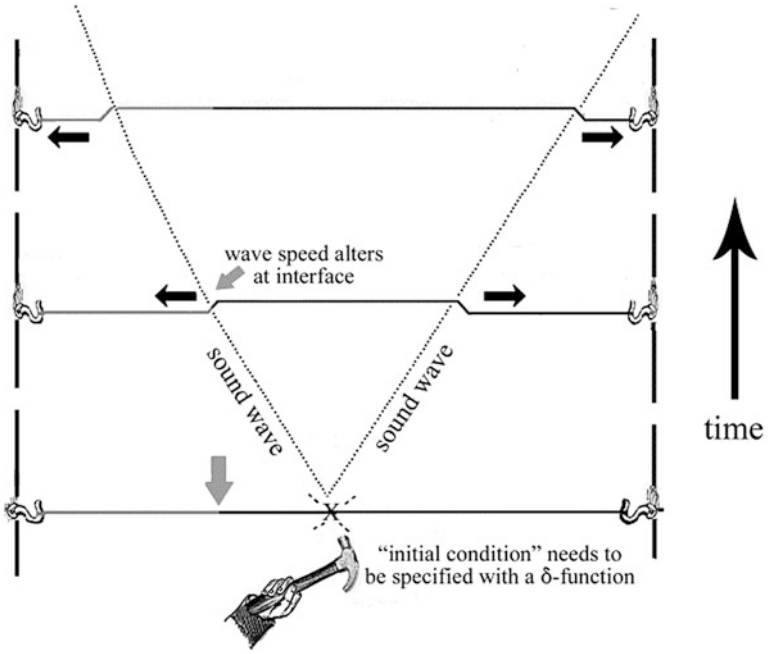
equation.[9] Suppose that we happen to consider the restricted initial conditions <P(x, t_0), V(x, t_0) > in which the position specification P(x, t_0) is required to equal $1/\sqrt{c}$ $\int_L{}^x$ V(x, t_0) dx of the velocity specification V(x, t_0). Call this restriction **R,** and let's suppose that we have failed to notice this special feature of our chosen starting conditions. Within all **R**-delimited circumstances, the induced waves will travel in a resolutely right-handed direction, until they collide with the nut at the far end. Within this limited range of **R** variation, a simpler form of governing law becomes viable: the uni-directional wave equation ($\partial y/\partial t = c\partial y/\partial x$). What's wrong with this modeling? Physically, a proper registration of the restorative processes active within the interior of a string S should reveal that a pure disturbance in initial position P(x, t_0) will normally split into two left and right heading waves, as a natural outcome of S's attempts to straighten itself out. If we only consider initial conditions that are **R**–obedient, the go-in-both-directions natural response of our string gets masked by a special restriction on initial velocity that completely suppresses the normal left-heading reaction. In other words, our model's apparent propensities in favor of right-heading waves do not reflect characteristics that are genuinely *internal to* S, but instead qualify as ersatz projections into S's interior of special features that largely reflect the restricted manner in which we unwittingly released the cord at time t_0. The proper corrective, Hadamard advises, is to examine our modeling of S under a wider range of potential starting conditions capable of erasing externalist biases of a **R**–projected character. Avoiding faulty internalist projections of this type supplies the primary reason why mathematicians build freely assignable initial conditions into the formal requirements of a well-set problem (if the modeling calls for initial conditions at all).



Similar requirements on modeling variability apply to other aspects of modeling that do not qualify as initial conditions in a proper sense. Suppose that the region in which our string is located contains an ambient 60 cycle hum. Its coupling with our string may induce wave patterns to appear within the latter that, once again, are not characteristic of the string's intrinsic propensities, and we way wish to examine our core modeling under a wider range of variations that can filter away these unrepresentative behaviors. Situations of this ilk are standardly labeled as control

[9]His chief illustration [1952] is quite substantive, for he shows how a parallel limitation to analytic initial data fails to reveal the underlying processes within a hyperbolic modeling. In our toy substitute, the restriction on initial conditions turns off the leftward heading component within d'Alembert's general solution for the wave equation A(x − at) + B(x + at).

variable problems, where amplitude of extraneous hum qualifies as an externalist "control variable."[10]



Of course, if we merely examine a selected set of varied initial conditions for our rope, they may all accidently embody the unwanted characteristic **R**. Accordingly, mathematicians refuse to certify a modeling as properly well-set if they haven't managed to filter away these unwanted artifacts through considering a sufficiently widened family of starting conditions. At root, these same considerations lie behind Woodward's requirements on counterfactual dependency when we attribute "causal processes" to a target system based upon experimental manipulation: he hopes to filter off the unwanted projection of externalist correlations due to "common causes" into the interior characteristics of the system under examination. To be sure, in real life we can't concretely test for every possible variation of this sort, but we want our understanding of "causal process" to reflect the fact that we view any remaining externalist projections as unwanted.

A second set of allied considerations, likewise emphasized by Hadamard in his influential studies of wave motion ([5], originally published in 1903), is that we may wish to enlarge our family of "initial conditions" to encompass starting

[10]For allied reasons, we might want to examine our rope + elephant model over a wider arena, in which we vary the weight of the elephant as a "control variable." As remarked in an earlier note, forcing conditions sometimes become true boundary conditions when the dimensionality of an example is enlarged. For such reasons, control variable problems concerning boundary condition assignments are often important. Nonetheless, we must distinguish these motives for considering wider assignments of boundary condition values from the more central requirements that reflect the effacement-from-environment concerns we shall detail in the next section. Mathematicians distinguish "control problems" from regular "well-set problems" for exactly this reason.

states that cannot be put into conventional <P(x, t_0), V(x, t_0)> format, because the internal causal factors operating within the system will become more vividly revealed under this enlargement. Consider our cord + elephant once again. The best way to understand how causal processes operate within a linear rope is to consider how newly-formed waves will travel through the rope after we give it a sharp rap at t_0 with a hammer. "Initial conditions" of this sort cannot be properly captured in <P(x, t_0), V(x, t_0)> terms, but instead require stipulations of a "Dirac -function" character.[11] The tacit counterfactual assumptions buried within this species of "initial condition" enlargement are rather subtle, and we won't attempt to detail them here.[12] As every sophisticated text in modern applied mathematics amply demonstrates, the "spaces of adjoined possibilities" pertinent to modeling circumstances of this character need to be very precisely specified and will typically vary from one explanatory format to another. We claim that our opening question of "how can counterfactual conditionals assist real life science?" hasn't been adequately addressed until the reasons why different modeling situations call upon different classes of restricted fictional circumstance have been diagnosed.

4 Motives for Counterfactual Family Enlargement: Boundary Conditions

The ranges of counterfactual variability correlated with the *boundary conditions* of a problem are adopted for quite different reasons than the free variability just considered, although they likewise trace to a desire that a well-set problem reveal the inner workings of a target system in an effective manner. As noted above, boundary conditions can only be assigned to the special spatial locales where we can characterize in advance how the target material will behave there. Two standard exemplars already illustrated are: (1) Dirichlet conditions, where we know that the cord will remain totally immobile at its two endpoints and no energy will leach from the system; (2) Neumann conditions, where we instead know the rule whereby vibrational energy gets lost at these same endpoints. Fruitful S versus E cuts cannot be situated just anywhere; nature itself must suggest the special locations where such divisions can be fruitfully implemented.

But how do we locate these descriptive opportunities? Frequently simply from the advice of experiment and/or common experience: we find that ascertaining the surface condition of a blob of material effectively "screens off" its internal

[11] Mathematicians call these stipulations "distributions" (or something fancier, if modeling requirements require). In such circumstances, neither P(x, t_0) nor V(x, t_0) can be credited with normal numerical values.

[12] See "Semantic Mimicry" in [3]. For novelty's sake, the diagram illustrates a string composed of two sections (gray and black) welded together, causing a finite change in wave speed when the join is transversed. "Side conditions" pertinent to interfacial transport naturally emerge in such contexts.

behaviors from the greater complexities of its surrounding environment (at least up to first order effects). Here "screens off" means (roughly) that with respect to some effects of interest E, conditional on information about what happens on the surface, further information about the interior makes "effectively" no further difference to E. Knowing that our violin bridge and nut will remain (nearly immobile) for the temporal interval under inspection allows us to largely ignore what happens in the supportive wood below. These assurances largely stem from direct manipulationist data: we have found that if we control the endpoints of our string so that they remain fixed at both ends, most other varieties of exterior alteration will make little different to the interior wave behaviors (with the exception of forcing conditions such an ambient hum). When we frame the larger family of behaviors into a well-set problem, we always extend these boundary screening assumptions to all of the counterfactual conditions included in the collection.

When analytic metaphysicians claim that such boundary region counterfactuals "demand a grounding in law," we are not sure what they mean. Insofar as the pinning of a violin string depends upon "laws of nature," they must presumably reflect the tensile and cohesive properties of the wooded bridges and pins that hold the attached wire fast. But in typical boundary condition assignments, practitioners do not probe these supportive underpinnings within the boundary conditions further, in contrast to the interior string (for which "system laws" are explicitly articulated to capture the central physical processes active in these central regions). This is one of the central ways in which physicists and engineers effectively *efface* the interior behaviors of their target systems from the greater complexities of the surrounding environments beyond. Well-chosen "boundary conditions" exploit the descriptive conveniences of what [3] calls "unequal data registration," for the salient behaviors of a violin bridge or nut can be adequately captured in relatively crude terms ("these parts remain fixed") for significant blocks of time, at least in so far as the significant interior behaviors of the violin string is concerned. But the justifications for these descriptively simplified choices usually derive from direct manipulationist testing, rather than relying upon substantive "laws of nature" in any evident manner.[13] We believe that the philosophers who insist upon "groundings" owe us a richer explanation of why their demands appear so greatly at variance with standard physical practice.

Before we move ahead, let us observe several further proclivities that mar the thinking of many who write on counterfactuals. Consider how [6] criticizes Woodward for not further explicating the "truth-conditions" of the manipulationist counterfactuals invoked in [7]:

> While Woodward relies heavily on counterfactuals, he says surprisingly little about their truth conditions... This raises puzzles because standard theories [of counterfactuals] appeal directly to natural laws: "A → C" is true iff A, background facts, and actual laws

[13]To be sure, certain forms of Neumann condition call upon principles such as "Newton's law of cooling," although such provisos rarely satisfy the "law of nature" expectations of the metaphysicians.

> jointly imply C . . . . Woodward's arguments support only the weaker contention that there is some close connection between counterfactuals and causal explanation. For example, it remains open to say that this connection is that the truth conditions of counterfactuals immediately involve laws, and that their causal and explanatory force derives from that fact. [Hiddleston, p. 547]

The concluding sentence apparently presumes that the modal force that drives the explanatory deduction of C from A stems entirely from the inductive generalizations captured within the underlying "laws of nature." In contrast, the salient "background facts" (which includes both initial and boundary conditions) are regarded as "modally inert" in the sense that they allegedly represent simple indicative facts that supply no bearing on the future until they become harnessed to the deductive power of the "laws." Hiddleston's phrase "actual laws" suggests that he believes that the laws exclusively obtain this power to drive counterfactuals inferentially forward because they represent inductive generalizations that have been straightforwardly framed on the basis of real world evidence. Allied opinions are widely shared among present day metaphysicians.

Be this as it may, this picture is scarcely defensible along a variety of fronts. For example, consider Hiddleston's invocation of "standard theories of counterfactuals." He believes that such accounts supply a full *reduction* of counterfactuals to contentions that carry no modal or counterfactual commitments except within the "laws" explicitly invoked in the analysis. But how can this be true? In particular, how can the standard boundary conditions for a violin string be plausibly viewed as "modally inert" in the manner Hiddleston requires? After all, a string's evolving behaviors will be as strongly affected by the future-looking behaviors of its endpoint provisos as by the interior wave equation "law" that only registers the physical processes active *inside* the string (viz., that string curvature directly correlates with accelerative force). But equally crucial events occur over time at the string's endpoints, and these provide the critical factors that force waves traveling along the string to reflect backwards into the interior and to disperse in a manner that allows the system's internal energy to resettle[14] into the standing wave patterns that provide the string with its characteristic tonal features (for this to happen, the majority of the string's vibrational energy must become allocated to its fundamental tone, to its octave, to the fifth above, and so forth). But these "modally active" ingredients are largely codified within the Dirichlet *boundary conditions* that we attach to our string modeling, not to "laws" in any conventional sense. To be sure, the interior wave equation ($\partial y^2/\partial t^2 = c^2 \partial y^2/\partial x^2$) may *look* more like a law-like "causal generalization" than a standard Dirichlet endpoint stipulation, but that deceptive appearance merely reflects the fact that endpoint behaviors can be adequately captured in simpler and cruder syntactic terms than the interior behaviors. By any reasonable standard, Hiddleston should have grouped "boundary conditions" together with his "laws" as coequal participants in providing the "modal force" that

[14]This energetic redistribution is governed by further physical factors that are *not* captured within the wave equation proper.

drives a true counterfactual claim from assumption A to conclusion C.[15] We believe that loose criticisms such as Hiddleston's need to be scrutinized more carefully than is generally the case in the philosophical literature.

In a similar manner, the *constraints* discussed in the next section also represent inferentially active, time-forward generalizations that do not look much like "laws" in any conventional sense.

The misdiagnosis of the inferential force of boundary conditions and constraints just sketched is scarcely idiosyncratic to Hiddleston, but is characteristic of the misapplications of applied mathematical terminologies that have marred philosophical writings on scientific explanation from the days of the logical empiricists onward. Wilson [3] further documents the widespread tendency to lump "initial conditions" and "boundary conditions" into a common, confused category, together with other modeling considerations that are properly neither. We'll later see that the notion of a "law of nature" is itself subject to similar terminological abuses.

One of our background motives in writing this essay is to protest against criticisms of Woodward's work that rest upon mistaken methodological claims akin to Hiddleston's (their varieties are many[16]). As we remarked at the outset,

[15] Indeed, a more detailed study of violin tone requires that the waves passing through the bridge and nut to the instrument's body should be scrutinized more closely, at which point our former Dirichlet endpoint condition will open out into a very complex process involving wave equations rather like the one we applied previously only to the string.

[16] Although we have selected Hiddleston as our chief target due to the pithy manner in which he articulates his claims, the presumptions he articulates—viz., that counterfactuals require modal "backing" stemming entirely from "laws of nature"—have been widely accepted for decades within philosophy. From this vantage point, philosophical accounts that appeal to "undischarged" (= unanalyzed) counterfactuals are dismissed as inadequately "grounded." These popular prejudices follow from (or, at least, are naturally suggested by) "metalinguistic" accounts of counterfactuals in which a counterfactual qualifies as true only if its consequent is derivable from its antecedent in conjunction with other premises, including the applicable "laws of nature." Accounts of this sort trace back to Goodman [1], but broadly similar assumptions have been defended more recently by Maudlin [8] and by Paul and Hall [9], where the "grounding laws" are now assumed to adopt a more specific form—they must represent laws of temporal evolution, and the systems to which they apply should constitute well-posed initial value problems. These apparently represent the background doctrines to which Hiddleston tacitly appeals. The major alternatives to these metalinguistic accounts invoke "similarity relations" among possible worlds in the manner of Lewis [10]. These alternative treatments also assign a preeminent role to "laws of nature," without special regard for boundary conditions or the other ingredients of normal scientific specification. Many authors within these schools further believe that the grounding "laws" themselves can be reduced to Humean claims about "actual" regularities, leading to the conclusion that all counterfactuals can be assigned fully "actualist" truth-conditions. We firmly contend that none of these purported reductions have been adequately established.

It is worth noting that a number of divergent motivations have been offered for these basic "grounding" assumptions. Some writers (e.g. [11, 12]) articulate *epistemic concerns*—they maintain that counterfactuals cannot be reliably assessed for truth or falsity without information about grounding laws. For example, the second article mentioned criticizes Woodward's use of interventionist counterfactuals on the grounds that we cannot determine which interventions are possible and which results would follow if they were to be carried out unless we already know the laws governing the system in question. But this claim is surely false—we can discover which

pinpointing exactly where and how various restricted classes of counterfactual constructions make themselves useful within real life scientific endeavor represents an important and non-trivial task that has been generally neglected by academic methodologists of science, despite the considerable insights that Hadamard and other mathematicians have contributed to these issues. Presumptions that all counterfactuals "obtain truth-values" in exactly identical manners and advance science in equally beneficial ways strike us as patently insensitive to the problems of characterizing the subtleties of real life effective procedure in accurate terms, whether these policies arise in the context of framing a suitable well-set problem (as we discuss here) or within Woodward's work on the experimental underpinnings of common forms of causal attribution.

5 Motives for Counterfactual Family Enlargement: Constraints and System Laws

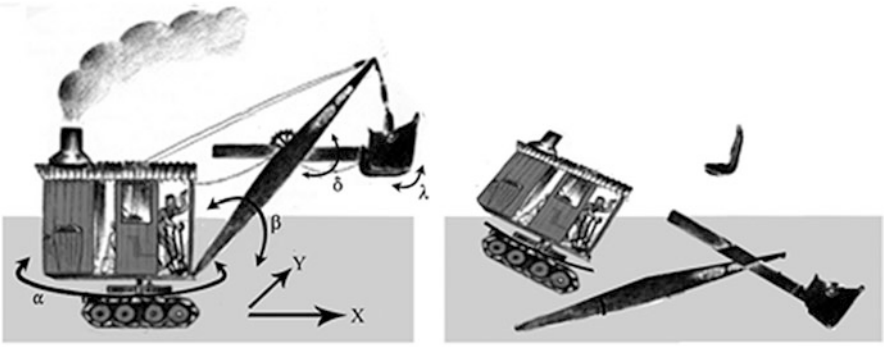
Counterfactually extended manipulationist data of other sorts enter physical practice in a wide variety of further ways, two of which we'll now discuss. Let's first consider the case of the Intact Steam Shovel. At time t_0, let us subject the device to sundry initial conditions, such as a large bump when the mechanism rolls over a rock. Classic Lagrangian methods (exploited by mechanical engineers on a daily basis) decompose the gizmo's possible movements into its evident freedoms of movement, (labeled as x,y,α,β,γ,δ in the diagram) that reflect the basic directions (x,y) and angular rotations (α,β,γ,δ) that our shovel can potentially make without breaking apart (physicists call these new variables "generalized coordinates"[17]).

interventions are possible and what happens when they are performed simply by doing experiments and performing manipulations. As we stress elsewhere in the paper, the great effectiveness of Lagrangian methodology within engineering traces precisely to the fact that we can learn the constraints that restrict a system's movements through direct manipulation without gaining any further information about any pertinent underlying laws.

Other writers (e.g. [8], but also [11]) appeal to broadly *semantic concerns*—they claim that stand-alone counterfactuals without grounding laws are commonly vague, context-dependent and unclear in a manner that makes them unsuitable for use in science. Providing backing laws is required to repair these deficiencies in truth-condition. We agree that *some* counterfactuals exhibit these flaws, but these criticisms rarely apply to the manipulationist counterfactuals under review in this essay. In more recent literature, it is often acknowledged that such epistemic and semantic contentions are unconvincing, and current fashion directly appeals to considerations of an overtly "metaphysical" nature. Often these replacement doctrines are articulated in manners that we find unedifying: viz., "counterfactuals cannot be barely true, but require grounding in what is actual." Insofar as we can determine, these misty claims stem from the same methodological prejudices as motivated the epistemic and semantic concerns of former times. It is striking how "grounding" doctrines continue to thrive even as their philosophical underpinnings shift significantly in the interim.

[17]Contemporary metaphysical opinion frequently presumes that all of the vital traits pertinent to a target system S can be grammatically constructed from the "fundamental qualities" allegedly appearing in the "laws" that govern S (or within their justificatory underpinnings). As [3] points out in considerable detail, this assumption is naively framed. The standing wave tonal spectrum

An engineer can greatly simplify her modeling task by framing her models in terms of these new variables, rather than working with the Cartesian locations of all of the device's pieces. In doing so, she will only consider the positions and velocities that the 6-tuple <x,y,α,β,γ,δ> might possibly assume, rather than worrying about the wider set of Cartesian location possibilities, most of which will situate the shovel in disconnected pieces (the more restricted set of <x,y,α,β,γ,δ> variations are called the device's *mobilities*). Standard computational techniques deftly exploit these significant descriptive simplifications by "walking through" the appropriate "mobility space" (i.e., by continually refining trial solutions until they stabilize on the correct machine behaviors—see [3] for further details). For these techniques to work properly, the pertinent "mobility space" must include all possible <x,y,α,β,γ,δ> values.



acceptable initial condition unacceptable initial condition

In these circumstances, a well-set modeling will restrict its interest in "freely assignable initial conditions" to just the freely assignable x,y,α,β,γ,δ possibilities. But how will our engineer ascertain the relevant set of x,y,α,β,γ,δ degrees of freedom? Typically, by simple induction from manipulative experiment: "Wiggle the sundry parts of the mechanism and you'll discover that you can freely choose the angles x,y,α,β,γ,δ without tearing the damned thing apart."[18] Once she settles on

of a violin string plainly represents one of its most important physical characteristics (natural selection, after all, has fashioned our ears and brain to filter away the extraneous noise that surrounds these vibratory characteristics within everyday life). But these traits do not appear as grammatically constructible vocabulary within the relevant "system law" $\partial y^2/\partial t^2 = c\partial y^2/\partial x^2$; indeed, that formula doesn't pretend to capture all of the physical factors responsible for making standing wave behaviors prominent in our musical lives. As noted above, those traits only become manifest when our interior wave equation is hooked up to boundary conditions that assist the internal energy storage characteristic of tonal vibratory behavior. In terms of present distinctions, the tonal characteristics qualify as a variety of "generalized coordinates."

[18]Note that these manipulative experiments again yield counterfactuals that are not grounded in laws in the sense at issue in this essay: we don't need, to appeal to laws to explain what the counterfactuals mean or how they can be reliably known and there is no reason to think there is a conceptual link of some kind between the counterfactuals and grounding laws.

a suitable range of initial states in this manner, she will want to examine all of them in an even-handed manner, to avoid ersatz externalist projections of a Hadamard-like character.

When a system exhibits locked-together behaviors amongst its parts in the manner of our steam shovel, mathematicians say that its movements have become subject to *constraints*. In the case before us, these limitations on mobility can be captured in entirely geometrical terms: the boom can move relative to the cab only by turning through the angle β and so forth. By employing descriptive variables (such as x,y,α,β,γ,δ) that naturally reflect these geometric limitations, scientists find that they can exploit the easy-to-obtain knowledge of the system's constrained movements to "cut off" a huge amount of unwanted lower scale complexity, for Lagrangian methods allow them to ignore all of the detailed physical processes that keep the steam shovel parts intact.

Expressed in vernacular English, constraints don't look like "laws" in any conventional sense; most folk would be nonplussed to learn that "in a steam shovel, the central boom is hinged to the cab" qualifies as a "law of nature." Nor would most observers agree that our knowledge of the truthful counterfactual "if the cab were turned gently through an angle α, the boom would remain attached" rests upon a "grounding" within more canonical "laws" (if it does, we certainly don't know what they are). Once again, real life practice often commences with an experimentally determined collection of reliable manipulationist data, which is then counterfactually extended to a fuller "mobility space" around which a suitable "well-set problem" is then framed.

Rarely are "side condition" data such as constraints mentioned within the standard counterfactual literature, despite their ubiquity within real life practice. We presume that many writers would be tempted to characterize such provisos as some variety of "initial and/or boundary conditions," although it is hard to see how such a loose assimilation could be justified. And our constraint considerations point out a more serious terminological foible inherent in these same discussions: the great abuses to which the term "law of nature" is commonly subjected. Standard universalist approaches to counterfactuals invariably invoke "laws of nature" as central ingredients in their analyses, yet they rarely delineate what they expect of such "laws," usually preferring to gesture vaguely in the restrictive direction of "Oh, I mean stuff like Newton's laws, Maxwell's equations, the law of gravitation and so forth." On other occasions, they will happily embrace the full set of interior differential equations employed within a modeling as the system's "laws," which, in the case of our central example, consists entirely of the familiar one-dimensional wave equation $\partial y^2/\partial t^2 = c\partial y^2/\partial x^2$.[19] Below we shall label such modeling equations

[19]Many authors cheerfully cite "all ravens are black" and "all dry matches ignite when struck" as candidate "laws," despite the fact that neither of these assertions look like plausible "laws of nature" in any traditionalist sense. Historically, the notion of "law" emerged within the annals of science in a wide variety of highly irregular ways, often carrying along fossilized remnants of archaic conceptions of scientific method. Significant confusions can arise from the common

as the *system laws* pertinent to a modeling. This is a significantly wider notion than the "laws of nature" to which authors like Hiddleston generally appeal.

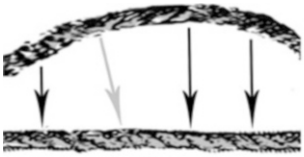
We should immediately observe that standard "law" exemplars such "Newton's laws, Maxwell's equations, etc." rarely provide the modeler with enough material to set up a set of differential equation "system laws" that display adequate "equational closure." What do we mean by that? In our string case, "equational closure" requires that our model's assembled ingredients (side conditions, differential equations and other ingredients) should fit together in a harmonious manner that supply enough formulas that the system's anticipated behaviors can be presumptively[20] projected forward into the future in a unique and stable manner for an appreciable span of time. These demands for equational closure should be viewed as the calculus analog to high school algebra demands upon the solvability of a set of linear equations like $x - 2y + 3z = -6$. How many equations are needed to solve for variables, x, y, z, ...? As a rough rule and barring hidden redundancies, n equations for n variables will be required. Smaller or larger equation sets are apt to prove underspecified or excessively demanding. In the same way, a well-set problem involving differential equations needs to assemble its descriptive ingredients in a manner that can generate stable and well-defined models over useful spans of time. And general principles such as of "Newton's laws, Maxwell's equations" character are rarely adequate to these "adequate system law" purposes.

To underscore this moral, let us briskly survey the derivations one finds of the standard string equation in introductory textbooks. Our basic modeling task is to link the x and y position (x and y) and mass density (Δ) of every point on the string at a given time (t) to the restorative force (f) that will accelerate the string according to its local curvature ($\partial y^2/\partial x^2$). The general laws cited cannot link these variables together in the "equational closure" manner required. Textbook derivations vary considerably in rigor from one source to another but the best ones rest their arguments on three central pillars (1) $\mathbf{F} = m\mathbf{a}$; (2) Hooke's law (in one-dimension along the axis of the string) and (3) the constraint presumption that each section of string will remain directly above its rest position when stretched (as illustrated). Only (1) looks like a "general law of nature" in the expected sense, but its content is far too feeble to provide the equational closure demanded. (3) is clearly a *constraint on movement* in the general manner of the features described above, and its plausibility with respect to real life strings traces largely to direct experimental evidence (counterfactually extended) concerning the approximate motions observed in sufficiently taut strings (viz., wobbles in the x-direction are practically unobservable).

---

practice of presuming that some scientific claim enjoys certain formal features simply because somebody long ago decided to label it as a "law."

[20] We write "presumptively," because simple "n equations for n variable" rules of thumb sometimes fail, requiring more refined studies of solution existence and uniqueness.



Let us finally consider (2): the "Hooke's law" assumption that that a stretched unit of string will develop a restoring force directly proportional to its displacement from its natural rest position (along the axis of the string).[21] Although behavioral assumptions of this linear character have been traditionally labeled as "laws" for nearly five centuries, their provisos do not have the contents that grounding enthusiasts expect under the heading of a "law of nature."[22] Certainly, its scope is not "universal" in any plausible sense, for no realistic materials obey Hooke's principle perfectly, and most materials do not conform to Hooke's law or its analogues even approximately. The carefully wrought products sold as violin strings successfully approximate to Hookean behavior only through devoted attention to fabricational detail (a rather complex and delicate internal structure within the string is required). In modern classificatory parlance within mechanics, "Hooke's law" is classified as a contingent constitutive principle and does *not* qualify as a "general law of mechanics" in the same mode as $\mathbf{F} = m\mathbf{a}$ or the balance of angular momentum.[23] Its applicability to a target material is generally established on the basis of direct testing bench experiment, rather than through any appeal to the elaborate forms of molecular modeling that grounding enthusiasts appear to anticipate in their demands for law-like underpinnings.

We shall not pursue these diagnostic observations further, but believe that they adequately illustrate that standard claims about the alleged "truth-conditional dependencies" of scientifically useful counterfactuals need to be more carefully calibrated than is common today. For example, [12] remarks:

> The [story] I favor ties the truth-conditions of counterfactual assertions to laws of nature. It is then easy to see how the evidence-conditions (that is, actual and hypothetical experiments) are connected with the truth-conditions of a counterfactual: actual and hypothetical experiments are symptoms for the presence of a law.

But what exactly does this mean? In the string case just examined, the differential equation "system law" utilized equally supplies "symptoms of the presence of several previously established families of experimentally supported counterfactuals."

[21] A proper derivation pathway between this axial stress/strain relationship and string curvature is fraught with subtle difficulties that we shall not review here. See [13].

[22] In modern usage, these are often called "constitutive principles."

[23] Worse yet, the forces posited in Hooke's law possess a character prohibited by Newton's third law as normally construed, for they presume a natural rest configuration to which the system strives to return. For a discussion of the general problem of articulating "fundamental force laws" capable of backing up the common procedures of classical physics, see [14].

Philosophers like Psillos assume that the "generalizations" responsible for driving a correct counterfactual A → C from its antecedent A to its consequence C will comprise "laws of nature" in a conventional sense. But this is a misleading claim. To be sure (*modulo* our qualifications with respect to boundary conditions and other side condition factors), the central "system laws" within a standard modeling scheme will supply most of the "A to C" inferential connections required, but the majority of their contents will not derive from certifiable "laws of nature" in any straightforward manner. It is only by muddling "system laws" together with hand -waving appeals to "Newton's laws, Maxwell's equations, etc." that any illusion is created that we thereby escape every form of counterfactual appeal when we look into "evidence-conditions." On any straightforward understanding, a large number of direct manipulationist appeals can be readily located within the "evidence-conditions" of working science.

6 Other Varieties of Well-Set Problems and Our Conclusions

Another of Hadamard's significant achievements was that of distinguishing a number of distinct forms of explanatory architecture and explaining why each naturally demanded different categories of attached side condition. Suppose that we know in advance that a target system will settle into an equilibrium state fairly quickly, in the manner that a soap film will quickly span a metal rim in a stable pattern. Such circumstances generally offer the modeler a golden descriptive opportunity, for she can often calculate the qualities of the interior equilibrium configuration without needing to study how it got there. As noted earlier, the only side condition she now needs is the position of the metal rim and the presumption that the soap will attach itself there. The result is what Hadamard called a "pure boundary problem," as opposed to an "initial-boundary-value problem" of the sort we have already discussed. In the circumstances of our elephant + cord, the relevant "initial-boundary-value problem" tells us how wave disturbances will move through the cord over time, and its relevant "system law" directly mentions time. But behaviors so described will never reach any form of equilibrium at all (frictional influences need to be introduced in the system equations for this to occur). But the operations of friction are subtle and ill-understood, so as long as we know ahead of time that such processes—whatever they are!—will eventually bring our elephant + cord to rest. Accordingly, modelers generally focus upon predicting that final equilibrium state without worrying about any intervening events. The upshot is a "pure boundary value problem" with no mention of initial conditions at all.

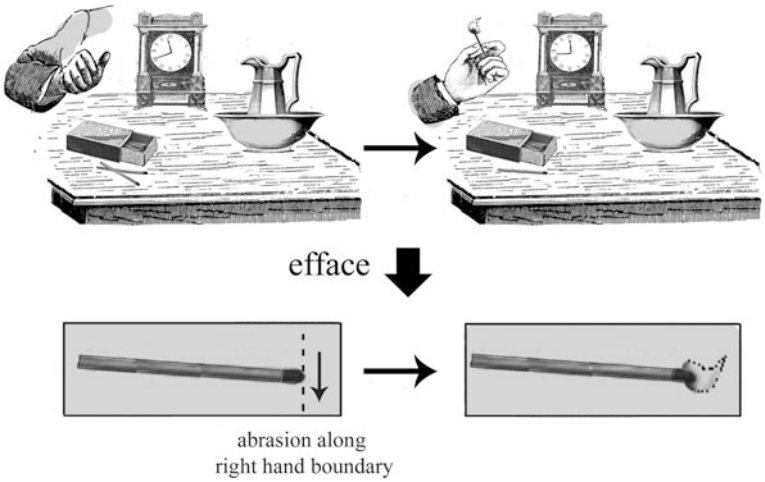
More generally, situations of a greater complexity involving constraints, feedback controls, and much else will invoke a wider variety of explanatory architectures, each with their own characteristic families of naturally attached side conditions. We shall not attempt to illustrate any of these richer architectures here, although applied mathematicians distinguish them precisely. But these considerations allow us to make an important methodological observation. In real life practice

a modeler will begin by assembling the experimentally available data that she believes can be safely extrapolated into specific families of reliable counterfactuals. If S regularly subsides into equilibrium, she will usually format an appropriate well-set model within a "pure boundary problem" format. If experimentally tweaking reveals that the constraint variables x,y,α,β,γ,δ characterize her steam shovel effectively, she will only enforce "free variability" with respect to these quantities. And so forth. In other words, prior knowledge of trustworthy counterfactual extensions serves as a valuable guide to an appropriate choice of modeling format.

Insofar as we can determine (although such issues generally remain tacit within the relevant literatures), standard commentaries with respect to counterfactuals turn these natural methodological dependencies nearly on their ear. Let us consider Goodman's classic counterfactual exemplar: "if this match were struck, it would light." From our point of view, this situation invokes a standard "initial-boundary value problem" format as its natural explanatory architecture.[24] Insofar as we can see, conventional opinion tacitly attempts to bend all counterfactual usage in science to this one stereotyped format, rather than allowing reliable counterfactual data to guide us to a more suitable choice of modeling scheme. Certainly, the proposals of Maudlin [8] can be profitably interpreted as appealing to standard effacement policies in more or less our own manner, for his discussion recognizes the significant role that suitable descriptive "cuts" play in reducing the complexities of nature to a more tractable forms of S versus E evaluation. In contrast, the popular approaches exemplified by the work of Lewis [15] or Stalnaker [16] ignore these redactive policies and attempt to construct full "possible worlds" around Goodman's match without invoking any effacement whatsoever. These embedding-in-a-full-world propensities characteristically embroil their authors in peculiar worries about the "miracles" that must be tolerated to bring these vastly amplified "worlds" into consistent accord. In our view, these conundrums stem from their overly ambitious attempts to supply "truth conditions" for every form of counterfactual statement ever uttered. But recall our opening manifesto: as philosophers of science, we should attend directly to the restricted families of specialized counterfactual claims that palpably assist the development of real world science. Considered from this point of view, various straightforward rationales for effective descriptive effacement become immediately salient, and no need for cutting back maximally detailed possible worlds into smaller possibilities ever arises.[25]

[24] More accurately characterized, the match situation probably invokes an "altered control variable to new equilibrium after an unspecified relaxation time" format commonly invoked within chemical and thermodynamic practice. In this essay, we have tried to sidestep detailed discussion of modeling architectures of this more complex complexion.

[25] For a general discussion of the inadvisability of inflating localized possibilities into possible world behemoths, see [3].



Yet despite its implicit recognition of initial time-slice problem effacement, [8] tacitly attempts to force every natural "pure boundary value problem" counterfactual (such as "if the wire rim were bent to an altered curve C*, the resulting soap film will assume the altered shape S*") into the ill-fitting straitjacket of a standard initial-boundary value problem, contrary to Hadamard's wise recommendations otherwise. But we can confidently ascertain the truths of our soap film counterfactuals without knowing anything at all about the complicated chemical intermediaries that allow the soap to find their new equilibria. Indeed, the great advantage of "pure boundary value" modelings is that they allow the modeler to safely *evade* evolutionary speculations for which they possess little data. Our confidence that physical circumstances S will likely submit to the great conveniences offered by "pure boundary value problem" modeling lies in the brute fact that S quickly returns to equilibrium after being subjected to an appropriate schedule of experimental manipulations. Hadamard classifies his different species of "well-set problem" according to the varieties of side condition data that enter into their formulation, and criticizes earlier authors who tried to jam all modelings together in a one-size-fits-all manner. We believe that the basic utilities of counterfactual appeals within science should be diagnosed in this same variegated fashion.

Indeed, excessively schematic approaches to counterfactuals do not appear to have assisted the project of understanding the effective methodologies of working science very ably, whether they are encountered within the arenas of experimental or quasi- experimental verification (Woodward's original focus) or that of trustworthy mathematical modeling (as surveyed in this essay). In the specific range of examples we have considered, no evidential mysteries whatsoever attach to how our specimen classes of associated counterfactuals "obtain their truth-values": these frequently derive from direct experiment, inductively extended. Moreover, the motivating utilities of these modal extrapolations are equally clear. Puzzles arise only when these firmly founded contentions are rashly lumped together with the speculations about Caesar and Sherlock Holmes. However, we also know, from brute experience,

that most philosophers writing on counterfactuals will find these methodological observations completely boring: "We are only interested in the deep 'metaphysical grounding' that counterfactuals must possess, not in the mere 'epistemologies' of how we come to establish these truths." In retort, we don't find their contrasting interests in "grounding" *boring* exactly, but we also don't fully understand what they seek.

Let us conclude by recalling the query posed in our opening section: where in science do appeals to counterfactual considerations concretely advance the scientific mission and where do they merely introduce unwanted distractions? We claim that the Hadamard-based considerations reviewed in this essay contribute to the positive aspects of this question in very constructive ways. In contrast, sweeping generalist approaches to this same question have not offered parallel enlightenment to date, and their characteristic abuses of Hadamard's classificatory vocabularies have muddled methodological understanding significantly. Philosophers who ignore the complexities of working science shouldn't cast metaphysical stones at the houses of those who do not.

References

1. Goodman N (1955) Fact, fiction and forecast. Harvard University Press, Cambridge
2. Hadamard J (1952) Lectures on Cauchy's problem in linear partial differential equations. Dover, New York
3. Wilson M (2017) Physics avoidance and other essays. Oxford University Press, Oxford
4. Eddington AS (1928) Nature of the physical world. Cambridge University Press, Cambridge
5. Hadamard J (2010) Propagation of waves and the equations of hydrodynamics. Birkhauser, Basil
6. Hiddleston E (2005) Review of Woodward, making things happen. Philos Rev 114:545–547
7. Woodward J (2004) Making things happen. Oxford University Press, Oxford
8. Maudlin T (2007) The metaphysics within physics. Oxford University Press, Oxford
9. Paul LA, Hall N (2013) Causation: a user's guide. Oxford University Press, Oxford
10. Lewis D (1973) Counterfactuals. Harvard University Press, Cambridge
11. Psillos S (2004) A glimpse of the secret connection: harmonizing mechanisms with counterfactuals. Perspect Sci 12(3):288–319
12. Psillos S (2007) Causal explanation and manipulation. In: Persson J, Ylikoski P (eds) Rethinking explanation. Springer, Dordrecht
13. Antman SS (1980) The equations for large vibrations of strings. Am Math Mon 87(5):359–370
14. Wilson M (2008) Determinism: the mystery of the missing physics. Br J Philos Sci 60(1):173–193
15. Lewis D (2001) Counterfactuals. Wiley-Blackwell, Oxford
16. Stalnaker R (1968) A theory of conditionals. In: Rescher N (ed) Studies in logical theory. Blackwell, Oxford