Exact Learning Reviews

# Nature Computational Science

Dear Dr Irwin,  
  
Your manuscript "A Natural Representation of Functions for Exact Learning" has now been seen by 3 referees, whose comments appear below. In the light of their advice, we have decided that we cannot offer to publish your manuscript in Nature Computational Science.  
  
From the reports, you will see that while they find your work of some potential interest, the referees raise concerns about the advance your findings represent over earlier work and the strength of the novel conclusions that can be drawn at this stage. We feel that these criticisms are sufficiently important as to preclude publication of your work in Nature Computational Science.  
  
I am sorry that we cannot be more positive on this occasion, but hope that you find the referees' comments helpful when preparing your paper for resubmission elsewhere.  
  
Best regards,  
  
Jie Pan, Ph.D.  
Associate Editor  
Nature Computational Science  
  
\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  
  
**Reviewers' Comments:**  
  
Reviewer #1 (Remarks to the Author):  
  
Summary:  
The paper proposes to use the Mellin transform as a basic representation for functions to be used in a Machine Learning model for learning exact analytical equations.  
Many known functions have a related Mellin transform and thus the suggestion is to perform fitting in the transformed space and then transform the result back to obtain an interpretable result.  
The proposed method is not evaluated in any way.  
  
Originality and significance:  
The paper is original, although Mellin transforms and all the mathematics is known, its application for function fitting is new to my knowledge. The problem setting of exact learning is very specific and typically far from realistic data. Data is typically corrupted and also not originating from precise analytical functions without distortions. The significance is further questionable as the method might be impractical or useless. The paper lack demonstration of results.  
  
Methodology:  
The mathematics is not simple, so the paper spends time to introduce it and guides the reader through the single variable and multivariable case. However, I was still missing some important discussions and I would maybe suggest changing the focus a bit.  
For me, there is the important point that a large class of functions can be represented by Mellin transforms (Fingerprints) with a set of defined parameters. A list is given in Table 2. As far as I understand, I need to choose one of those function classes and the number of parameters and then try whether it works. I would have liked to see the procedure more explicitly described in the text.  
Why would the fitting of the fingerprint parameters lead to the right solution with gradient descent?  
An analysis of the loss landscape would be important. Either theoretically or if not possible then at least empirically.  
In ML highly overparametrized models have been shown to be "trainable" and largely circumvent the problems of local minima in this way. In the proposed approach it seems that the number of parameters is very small, essentially encoding the prior knowledge of the complexity of the function to be fit. In a naive representation, the fitting of these few parameters is very difficult. It is not clear to me why it should be better in the "fingerprint" space, in particular as the equivalent activation function has an unfavorable shape for optimization.  
  
Claims and Conclusions:  
The conclusions and claims of the paper are only highlighting the potential of the method.  
Stronger claims could be made if the method would be demonstrated and shown where and how it can work.  
Alternatively, results on why the method should work and under which conditions it works.  
What happens if the hypothesis of the function class is wrong, for instance?  
  
  
Clarity and Presentation:  
The presentation is not ideal. The mathematics in the paper is known and the real contribution of the paper should be around using that formulation to present the algorithm/technique for exact learning  
and evaluate it properly. Using machine learning methodology and comparisons, for instance, EQL[1], DSR[3], or other methods, see references, or power series or taylor expansions etc.  
  
References: There is a bulk of literature on symbolic regression, which is the field the paper wants to contribute to. See a classic paper:  
Schmidt, M. and Lipson, H. Distilling free-form natural  
laws from experimental data. Science, 324(5923):81–85,  
2009.  
  
There is more recent interest in the ML community. Here are some examples:  
  
[1] Sahoo, S., Lampert, C. and Martius, G.. (2018). Learning Equations for Extrapolation and Control. Proceedings of Machine Learning Research 80:4442-4450  
  
[2] Udrescu, S and Tegmark, M. AI Feynman: A physics-inspired method for symbolic regression  
Scienve Advances, 6:16 2020  
  
[3] Petersen, B.K. .... Kim, J.K.  
Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients  
ICLR 2021.  
  
Details:  
Sec. 2.1.2 last paragraph. What do you mean by trained individually vs simultaneously? If gradient optimization is performed all parameters are optimized together in either way.  
  
The manuscript contains many small typos.  
  
  
Reviewer #2 (Remarks to the Author):  
  
This paper studies exact learning of analytical functions. The function classes considered are analytical functions characterized by the Mellin transform. The learning procedure is discussed. First, numerical Mellin transform is calculated, and is called the "numerical fingerprint"; Second, the numerical fingerprint is compared with the fingerprint of the function class parameterized by a set of parameters; Third, a least squares problem is solved to find the best fit of the parameters; Finally, the exact function is recovered by the inverse of the Mellin transform based on the parameters found in the previous step. The procedure above is similar to a usual machine learning problem, with the function class as the hypothesis space, and the parameterized fingerprint as the model. The relation between the proposed fingerprint and neural network models is discussed in the paper.  
In my understanding, what the authors did is proposing a new model that better suites the learning of analytical functions listed in the paper. The model can be understood as a neural network model with a new activation function. However, from my perspective this work is incomplete and should not be published at this point. At least the following points should be carefully and thoroughly studied before the paper becomes publishable:  
1. The author should make it clear the significance of considering the function classes listed in the paper. Why it is important to study the learning of analytical functions? How can exact learning of analytical functions help in practical problems, like in solving PDE or discovering physical laws? Examples should be provided and discussed.  
2. The steps discussed in the paper, as illustrated in Figure 1, is not ready for implementation. Many questions remain unsolved. For example, how to determine the best fingerprint class; how to choose s and how many s should be chosen; and how to minimize the loss function.  
3. Shown by figure 2, the log of Gamma function is highly non-convex. Hence, the square loss function between the numerical fingerprint and the parameterized fingerprint will be very complicated and hard to optimize. Is it possible to find a global minimizer at which the loss function is 0? If not, then there's no "exact learning". Either numerical experiments or theoretical proofs should be provided.  
4. It seems the "exact learning" still depends on rounding the learned parameters to integers. Like from 0.0013 to 0, and from 1.0015 to 1 (in appendix 5). This is an artificial step, and whether and how to round the numbers depend on subjective thresholds. Why integers are preferred than decimals? It there any scientific explanation? What if the a and b's are decimals themselves? Then is exact learning still possible? The name "exact learning" does not make sense to me because errors always exist unless you have a lot of prior knowledge of the parameters in advance.  
5. Though the authors may not agree with me, but in the reviewer's opinion numerical justification is crucial and necessary when a new method is proposed. Numerical experiments should be conducted and presented to show the feasibility and the performance of the proposed method. implementation details of the experiments should also be provided.  
  
  
  
  
  
Reviewer #3 (Remarks to the Author):  
  
Summary: In ``A Natural Representation of Functions for Exact Learning' the author discusses questions of function representation and touches on the subject of machine learning from a philosophical direction. The key observation is that modern machine learning models are not interpretable in the sense that it is hard to understand e.g., in neural networks with millions of parameters how a data input is transformed through the function. The goal of the article is to be a `foundational paper that sparks an idea' and introduce the Mellin transform as a key concept into machine learning.  
  
I believe that this article misses the point of machine learning entirely (see below). My recommendation is to reject this paper.  
  
1. I concede that it is undeniable that neural networks are not an intuitive function class, and that interpretability is an important topic to study.  
  
2. The author formulates as a goal that `a user can enter a high precision number -- e.g.\ 14.687633495788080676 -- and the algorithm would return a plausible closed form solution - e.g.\ $\pi^2 + e\sqrt\pi$'. This goal is not mathematically well-posed. For example, the set $Z\_{m,n} = \{m \pi^2 + n e\sqrt\pi : m,n\in\mathbb Z\}$ is dense in $\mathbb R$. In particular, there exist infinitely many `plausible' closed form solutions just of the same form, and infinitely many closed form solutions for $14.687633495788080676 \approx m\pi + n e$ etc.  
  
In every application, it requires great care to decide what is `plausible'. A `one size fits all' solution to interpretability like the author proposes here is doomed to fail from the start.  
  
3. The author considers a subset of the class of analytic functions to be good enough to work with. He is clearly knowledgeable of this class and associated integral transforms. Sadly, this function class is inadequate for almost any problem one may think of in scientific computing, let alone machine learning.  
  
4. It is not clear to me in what fashion power series coefficients are more interpretable than neural network weights. Additionally, polynomials (finite power series) which interpolate data are well-known to become highly oscillatory at the edge of a domain unless the sample points are carefully curated. This may not be the case if the coefficients are chosen by integral transform, but I do not see the author address any of these complications.  
  
5. The author presents his ideas in one dimension only. The major obstacle in machine learning problems is the high dimensionality of the data which leads to geometric sparsity (the `curse of dimensionality'). I do not believe that the author is aware of this phenomenon, or at least does not address it in this work.  
  
6. I am not sure why the Mellin transform was singled out as the `fingerprint' of a function. The author states that `it is not out of the question that the Mellin transform representation of the `fingerprint' also be a natural choice for exact learning' in Section 2.1.2 before becoming more optimistic and claiming that `[w]e have identified the Mellin transform is the method of extracting the fingerprint of a function' in Section 3. The process in between is unclear to me.  
  
7. The Mellin transform is a linear map. It is well known that linear models suffer from the curse of dimensionality in machine learning in situations where modern models like neural networks succeed (see e.g.\ {\em Barron, A.E.: Universal approximation bounds for superpositions of a sigmoidal function} for an illustration). I do not see these methods scaling favourably with dimension.  
  
8. The article is written in a very readable fashion, although not in scientific style. At times, the presentation suggests that the author does not have a deep knowledge of the topics they are discussing. As an example, the author mentions `a form of neural network, or generalised unsupervised learning implementation' while the entire article discusses function fitting, a form of supervised learning.  
  
9. In some fashion, the author seems to recover shallow neural networks in the end.  
  
It is possible, albeit unlikely, that the author's ideas are not entirely without merit in the context of machine learning. To convince a reviewer of that, the author should improve the article in several ways before resubmitting:  
  
a. Argue why the subclass of analytic functions for which they define the Mellin transform is of interest in machine learning.  
b. Motivate their choice of `fingerprint' and demonstrate that it efficiently conveys useful information.  
c. Present their ideas in dimension $d\gg 1$.  
d. Present their ideas in a mathematically convincing fashion. While complete rigor may not be required at every step, it is bad form to leave it to the reader to fill in all the difficult parts. Some of the issues that the author dismisses as `engineering problems' in the beginning are among the essential questions of machine learning and cannot be glossed past.  
e. State their main results clearly in the beginning. The best I could extract was that the Mellin transform can be applied to functions and has been studied previously, more rigorously, by other people. Apparently it has been useful in other fields and the author believes that it could be useful in machine learning to derive a function fitting procedure at discrete data points.

# AI Applied Letters Query

**From:** Olexandr Isayev <olexandr@olexandrisayev.com>  
**Sent:** 07 July 2021 16:24  
**To:** Ben Irwin <ben@optibrium.com>  
**Subject:** Re: Suitable for submission to Applied AI Letters

Dear Ben:

thanks for sending a paper inquiry our way. I briefly glanced through the paper and the math and see why you are having a hard time. It's probably very hard to find qualifying referees with appropriate math backgrounds. I also see the interesting general idea of your work. You are more than welcome to submit it to AAIL, but I worry that I will have the same problems, it's a very theoretical paper. To ensure the best fit with our scope, would it be possible to include a simplistic practical example like learning from a small dataset of your choice, or a function, etc. This will help your readers to understand the utility of your approach.

On a side note (from a scientist, not an editor) have you tried to send it to usual CS venues like NeurIPS, JMLR, JMVA etc?

Oles

# Nature Machine Intelligence

Dear Dr Irwin  
  
Thank you for submitting "A Natural Representation of Functions for Exact Learning" to Nature Machine Intelligence. Regretfully, we cannot offer to publish it.  
  
We receive many more papers than we can publish, which means we must decline a substantial proportion of manuscripts without sending them to referees, so that they may be sent elsewhere without delay. Decisions of this kind are made by the editorial staff when it appears that papers, even when technically correct, are unlikely to succeed in the competition for limited space.  
  
In this case, we have no doubt that your findings on your framework for exact learning will be of interest to experts in theoretical and applied machine learning. However, I regret that we are unable to conclude that the paper provides the sort of substantial practical or conceptual advance that would be of immediate interest to a broad readership of researchers in artificial intelligence or robotics. For us, there would need to be more of a connection made between the concepts, as you mention in the article, and perhaps more of a demonstration of your ideas.  
  
We thus feel that the present manuscript would be better suited to a journal other than Nature Machine Intelligence. Below, I have included links to other journals in the Springer Nature portfolio, in case you would like to transfer your manuscript. One possible fit is a new journal, Nature Computational Science. Please note that I have not consulted with my editorial colleagues at other journals, who would make their own independent editorial decisions.  
  
I am sorry that we cannot respond more positively, and I hope you will understand that our decision in no way reflects any doubts about the quality of the work reported.  
  
Best regards,  
  
Trenton Jerde, PhD  
Senior Editor  
Nature Machine Intelligence

# Review from a Mathematician Friend (also called Benedict)

**From:** Benedict Williams <benedict.williams@gmail.com>  
**Sent:** 16 August 2020 01:20  
**To:** Ben Irwin <ben@optibrium.com>  
**Subject:** Re: Paper

Dear Ben,

I am very much looking forward to seeing photographs of you and your daughter. I hope Mingfei is doing well!

I apologize for the excessive time it has taken me to respond to your paper. I do have some comments about it, which I hope might be useful to you. Bear in mind that while I criticise mathematics writing for a living, it's always pure mathematics, so I may not be the critic you really want for this paper.

In the first place, the main idea as I understand it is excellent: use the Mellin transform on numerical data, try to fit something like a hypergeometric series to the Mellin transform, then transform back to get an approximation to the original data that lies in a set of 'good' functions.

This is beautifully summarized by the diagram on p4.

To my pure-mathematicians eye, there are some points of ambiguity or vagueness. These you may disregard: I don't know if applied mathematicians would be as sensitive to them as I was.

A. Much of the paper is written for real numbers, but the complex numbers show up from time to time (in contour integrals e.g.). Can you do the whole thing for C (selectively avoiding poles of the functions being used)? Since you seem to be working with analytic functions throughout and use analytic continuation here and there, I would imagine that this is not only possible but even desirable.

B. You refer to 'closed form' and 'exact solutions' without really defining the terms. One might say 'closed form' is any representation of a function without using an integral sign or an infinite summation, but the question arises what other 'standard' functions one may use. It might be helpful to be precise here. The same precision would tell me what an 'exact solution' is.

C. The concept of 'fingerprint' is important, but not defined. A Mellin transform would seem to be a kind of fingerprint, but what else might serve as a fingerprint? Is a fingerprint an embedding of U -> V where U, V are spaces of functions? Is the identity a fingerprint?

D. The paper is too discursive. I think the section on the 'bigger picture' could go, and some of the historical remarks, and the more interesting metaphors. The truth is, the people who might make use of this ingenious method are going to want to get to the method quickly. That's not to say the discursions aren't interesting, or would not be valuable to know, but they belong in notes, or a lecture or a textbook, not in this paper. That's my opinion, at least.

E. The paper proper is missing a killer example: an instance of a problem that was difficult to solve without the use of this method, but is easy to solve with this method. If you think the examples in the supplement meet this criterion, you should move them into the main paper.

F. In fact, I think the material of the supplement should all be in the main paper, at the expense of the more conversational matters in the paper itself.

G. I do not understand the sections on neural networks, and I consequently wonder if they might belong in a follow-up paper to this one. You could have a first paper that talks about the Mellin->fitting->reverse-Mellin method, and second talking about the use of neural networks. They don't seem to be necessary for the main idea of this paper.

H. It would be helpful to be told about limitations to the idea that you have encountered in using it. I am not an expert in this field, but I think the main thing I would worry about is numerical stability of the Mellin transform (and its inverse). You do touch on stability problems briefly when you discuss the Gamma function on the whole line.

I attach a copy of the paper after some proofreading. I apologize for the amount of red ink… it's really a good paper, and the writing is especially good in the more technical parts.

Ben

# Cell Press Patterns

A Natural Representation of Functions that Facilitates `Exact Learning'  
PATTERNS-D-20-00065  
  
Jun 23, 2020  
  
Dear Dr. Irwin,  
  
I hope this email finds you well. I am enclosing the comments that the reviewers made on your paper. Unfortunately, the consensus recommendation is against publication in Patterns. The reviewers refer to the potential interest of the topic and the approach, but suggest that this would be more suitable for a specialised AI and machine learning journal. Given their comments, we feel that it would therefore be premature to proceed further with the manuscript.

I also have to apologise for the length of time it took to get the reviews in - finding reviewers capable of, and interested in reviewing, was a struggle.

I am sorry that the outcome for this manuscript was not more positive. However, I hope that the reviewers' comments are clear and constructive and that you will continue to consider Patterns for future submissions.  
  
Best wishes,  
  
Sarah Callaghan  
Editor-in-Chief, Patterns  
  
**Reviewers’ Comments**:  
  
Reviewer #1: The paper presents a general framework to interpret generalized hypergeometric functions as exact models to represent unknown numeric distributions as mathematical functions. The paper is easy to read, well structured and scientific sound. As stated, the paper is a foundational paper that sparks ideas and lay concepts. It is fundamentally a methodological paper.  
  
Using generalized hypergeometric functions is, in my opinion, the future of applied mathematics. Problems which could be solved by using elementary and low-order generalized functions are now scarce. Therefore, the framework proposed shall be useful for all sorts of problems.  
  
The parallels drawn between NN and DL are quite interesting and shall be further explored in future works.  
  
One of the main issues which, apparently, was not adequately treated in the paper is the uniqueness of the solutions. The Mellin-transform direct-inverse method provides unique mathematical solutions. But since the paper deals with the physical meaning of the solutions, there may be more than one model which provides similar solutions. A quite sensitive case shows up on multivariate statistics.  
  
In short, we are dealing with moment-matching techniques. This may be valid for univariate distributions, but not in general for multivariate ones. The author may check [1] to account for that. The unique representativeness of multivariate distributions may need additional assumptions on the dependence structure of the random variables involved (copulas etc) to really build a 1 to 1 correspondence. This must be mentioned in the paper to avoid numerical matching despite physical inconsistency of the models.  
  
Besides that, why did the author used the RMT instead of the residue theorem itself? It seems the results would be more general if the residue theorem was considered instead, as the RMT is a consequence of the  
residue theorem when taking the poles of the function Gamma(s) only.  
  
Overall, the paper can be accepted for publication after the issues raised have been properly addressed.  
  
[1] <https://doi.org/10.1016/j.jmva.2011.06.001>  
  
  
Reviewer #2: This paper proposes a machine learning framework, called exact learning, which aims at learning an exact mathematical expression representing an input-output functional relationship given as a dataset.  
The key technical contribution is to use the Mellin transformation to represent any function, where parameters can be learned from data.  
  
The idea of using the Mellin transformation for learning real-valued functions appears to be novel and interesting.  
Also, analyzing the connection between the proposed Mellin transformation based approach and deep neural networks is interesting.  
However, I have the following concerns:  
  
- First and most importantly, the motivation of "exact learning" is not clear. The authors discuss that the general form of the expression, called the theoretical fingerprint, given in Equations (6), (14) can be used to represent a functional relationship in a dataset. Then, if its expression is exact, meaning that the training loss is zero, which immediately means that the model is overfitting to the dataset, resulting in poor performance for prediction or extrapolation. This is a natural consequence of the problem setting of machine learning, and how to avoid this situation and obtain a model that generalizes well is the central question of ML. Currently, it is not clear how to treat this problem in the proposed approach. More careful discussion of the motivation of the proposed method is required.  
  
- Related to the above problem, I would recommend to prove the soundness and completeness of the proposed approach, in which the soundness means that any expression is well defined, and the completeness means that any function can be represented (or, at least approximated at any precision) by the proposed approach. For example, are discontinuous functions learnable?  
  
- There is no technical details of the learning process. In particular, there are three steps introduced in P.12, however, the step 2 is vaguely discussed and it is not clear how to achieve it. For example, if one uses the gradient descent method, one needs to check at least (1) is the optimization problem convex? (2) is it possible to analytically compute the gradient in the closed form? Since such properties depend on the model, the step 2 should be carefully designed according to the function used in step (1), while such discussion is missing.  
  
- Since there are a number of machine learning approaches that try to learn functional relationship from simple linear regression to k nearest neighbor regression, comparison with such method is needed to empirically evaluate the significance of the proposed method.  
  
- Can the proposed fingerprint represent data compactly? In other words, if it needs longer and longer expression to exactly encode more and more complex functional relationship in data, it means that the information of data is just translated into the fingerprint and no "learning" occurs. This point also should be carefully discussed.  
  
- Relationship between the RKHS is briefly discussed. Since RKHS is one of central concepts in ML and it is heavily used in various models such as kernel methods, more detailed analysis might be interesting.  
  
  
  
  
Reviewer #3:   
In this article, the author tries to solve a fundamental problem, fitting unknown function or distribution of artificial intelligence models. But the paper it too mathematical and is out of my knowledge to provide constructive comments and make reliable or faithful decision. I do apologise.  
  
One attractiveness is, the paper includes results that are reproducible using code published by the author.