INCENTIVIZING EMPIRICAL SCIENCE IN MACHINE LEARNING: PROBLEMS AND PROPOSALS

Preetum Nakkiran & Mikhail Belkin
Halıcıoğlu Data Science Institute
University of California, San Diego
{preetum,mbelkin}@ucsd.edu

ABSTRACT

We introduce a proposal to help address a structural problem in ML publishing: the lack of community support and perceived lack of legitimacy for experimental scientific work that neither proves a mathematical theorem, nor improves a practical application. Such experimental work is the bedrock of many fields of science, yet is not well appreciated by top ML publication venues (e.g. NeurIPS, ICML, ICLR). The problem is twofold: reviewers are often unaware of the value of such work, and thus authors are disincentivized from producing and submitting such work. The result is a suffocation of a scientific methodology that has a long history of success in the natural sciences, and has recently been fruitful in machine learning.

To address this, we propose introducing a Best Paper Award specifically for foundational experimental work in machine learning. The award targets scientific work that is missed by existing communities: we exclude primarily theoretical work and application-motivated work, both of which are well supported by existing venues (e.g. COLT, CVPR). We propose that ML venues include a subject-area for "scientific aspects of machine learning", and consider papers in this subject for the award. More ambitiously, it can be implemented as an endowed yearly award considering all papers in the prior year. We expect that establishing an award will help legitimize this research area, establish standards for such scientific work, and encourage authors to conduct this work with the support of the community.

In this proposal, we first discuss the structural problems in ML publication which we hope to address. We then present a call-for-papers for the "science of ML" subject area, describing the type of work we want to encourage. We argue that it is not only a scientifically legitimate type of work, but perhaps even a *necessary* type of work. Finally, we discuss broad guidelines for how such papers may be evaluated by reviewers.

1 INTRODUCTION

The field of Machine Learning— as well as computer science more generally— is an umbrella field harboring many different subfields, each with different goals. Among subfields of ML, an important distinction is whether the subfield is primarily *technological* (i.e. with the goal of building and improving learning systems), or *scientific* (i.e. with the goal of understanding the nature of learning). There is well-understood value in interactions between these two types of work— however, their motivations, methodology, and evaluation procedures are fundamentally different[1]. Nevertheless, many top ML conferences (e.g. NeurIPS, ICML, ICLR) do not cleanly distinguish between scientific and technological tracks: they have neither designated subject areas, nor specific reviewer guidelines for purely scientific work. The result is that the standards for technological work— which are the dominant output of the ML community— often become de-facto standards for scientific work. For example, reviewers are known to often judge scientific work in terms of its performance improve-

[1] See Eden (2007), Wegner (1976) and Goldreich & Wigderson (1996) for discussion of differences between scientific and technological goals in CS.

ment in practice.[2] Moreover, top venues do not explicitly encourage purely scientific work in their call-for-papers, or in their stated subject-areas. This creates at least two scales of problems: in the short term, authors are discouraged from producing purely scientific work, since it is unclear if their work will be respected at top venues. In the long term, this prevents a strong scientific community from flourishing within the field of ML, since mainstream ML venues do not provide a welcoming home for scientific activity.

There is one notable exception, where scientific work is properly supported within ML: the subfield of fully rigorous mathematical theory. This particular type of science is well-respected at top conferences, and is judged on its own terms (largely independent of technological considerations). Conferences have explicit "Theory" subject areas[3], and purely theoretical papers have won best-paper awards in top venues frequently in the past. However, the fully rigorous mathematical theory is one of many modes of scientific inquiry— and should not be mistaken for the *only* valid mode of inquiry. This mistake is unfortunately both frequently made and not at all new. Hoare in 1989, for example, dismissed experimental methods in CS as wholly unscientific:

> *I find digital computers of the present day to be very complicated and rather poorly defined. As a result, it is usually impractical to reason logically about their behavior. Sometimes, the only way of finding out what they will do is by experiment. Such experiments are certainly not mathematics.* ***Unfortunately, they are not even science****, because it is impossible to generalize from their results or to publish them for the benefit of other scientists.* (Hoare, 1989)

These sentiments, which are still prevalent in the ML community, results in a lack of support for other types of scientific work which are valid, important, and even necessary for future progress. We aim to help address this by focusing on the following structural problem in the ML publication system:

> **Problem 1:** *In mainstream ML venues, there is a perceived lack of legitimacy and real lack of community for good experimental science— which neither proves a theorem, nor improves an application. This effectively suppresses a mode of scientific inquiry which has historically been critical to scientific progress, and which has shown promise in both ML and in CS more generally.*

Below, we first give evidence that Problem 1 exists (Section 2). Then we argue that Problem 1 is important— that experimental methods are both valid and essential to scientific progress in ML (Section 3). These arguments are not new: we will survey arguments for experimental computer science as early as Turing, and place some of them in a modern ML context. Finally, we will outline our proposal to address Problem 1: a dedicated subject-area for "science of ML", and an accompanying best-paper award for this subject area (Section 4).

2 The Failure of Scientific Incentives

Here we list evidence that Problem 1 exists– that experimental scientific work is not well supported in mainstream ML venues. This is "folklore" in the community, but we briefly list concrete arguments below.

1. There was no designated subject area for scientific work that does not involve theorems in the most recent NeurIPS, ICML, or ICLR. Thus, such works get scattered among different reviewer pools, with unclear evaluation standards.
2. Scientific aspects of ML (towards understanding existing systems, not building new ones) did not appear in the call-for-papers in the most recent NeurIPS, ICML, or ICLR.
3. There are no reviewer guidelines at mainstream generic ML conferences for evaluating purely scientific work. In contrast, it is well-known that reviewers often ask for application improve-

[2]The ACL reviewing guidelines explicitly warns against this failure mode: "SOTA results are neither necessary nor sufficient for a scientific contribution" (Rogers & Augenstein, 2021; Rogers, 2020).

[3]These subject areas usually refer to mathematically rigorous theory, although this is not the only kind of scientific theory.

ments, and "improving SOTA" is often essentially necessary for acceptance (Sculley et al., 2018).

4. It is well-known that reviewers ask for "theoretical justification" for purely experimental papers, even when the experiments alone constitute a valid scientific contribution[4].
5. There are many workshops organized around scientific investigation of machine learning[5] which reveals the lack of support for such work in existing conference venues.

3 THE CASE FOR EXPERIMENTAL SCIENCE IN ML

The importance of experiments in scientific inquiry is not new: experimental science has a long history in both the natural sciences, and in computer science. Interestingly, the AI research community in the 1950-70s was one of the first to push for empirical scientific methods in CS, beyond the purely mathematical methods that had dominated previously[6]. This was famously presented by Allen Newell and Herbert Simon in their 1975 Turing Award lecture "Computer Science as Empirical Inquiry: Symbols and Search":

> *"Computer science is the study of the phenomena surrounding computers [...] Each new program that is built is an experiment. It poses a question to nature, and its behavior offers clues to an answer."* (Newell & Simon, 1976)

To Newell and Simon, computer science is the study of computers, in the same way that botany is the study of plants, or optics the study of light[7]. Thus, they argue that phenomena in computing can be treated as aspects of nature, and studied using the same methodology as the natural sciences—incorporating both theory and experiment as appropriate. Moreover, the core tenet of Newell & Simon (1976) and Simon (1995) is just as relevant today: in AI and ML, empirical methods are not only appropriate, but *essential* to scientific progress.

To see why, observe that there are at least two obstacles faced by fully rigorous mathematical theory, which are bypassed by experimental methods. First, the curse of complexity: computational systems can exhibit dynamics that are beyond the reach of current formal analytical tools.

> *"In computer science in general, and in AI in particular, we are usually operating in areas of greater complexity than those in which theorems can be proved. [...] Often the most efficient way to predict and understand the behavior of a novel complex system is to construct the system and observe it."* (Simon, 1995).

This intractable complexity is appreciated in many areas of science, including physics[8]. It is particularly evident in machine learning today, where we are far from having rigorous mathematical theories which characterize the behavior of real deep learning systems. However, we can still make progress with experimental methods, where we hope to empirically characterize behaviors rather than prove them.

The second obstacle is the "definitional barrier" to theory in ML (Nakkiran, 2021, Section 1.3.2). We cannot even *formally state* many relevant behaviors of ML methods, because we cannot precisely define the objects involved. For example, we would like to describe the behavior of a learning algorithm when trained and tested on data distributions from the real world (e.g. natural images or language). However, we do not have a precise enough understanding of these "natural distributions" to state formal theorems involving them. But this should not prevent us from studying ML: we can perform experiments on a variety of distributions we heuristically consider "natural", in the belief that they share a common structure we are yet to formally identify. As noted by Newell & Simon

[4] For example, see Tom Goldstein's recently NSF Town Hall on "how the ML community became anti-science": `https://twitter.com/tomgoldsteincs/status/1484609273162309634`.

[5] For example, "Identifying and Understanding Deep Learning Phenomena" at ICML 2019 `https://deep-phenomena.org/`.

[6] See Eden (2007) and Polak (2016) for more historical context and discussion.

[7] This phrasing based on Knuth (1974).

[8] The theoretical physicist and Nobel laureate Philip W. Anderson warned against the reductionist viewpoint: "The ability to reduce everything to simple fundamental laws does not imply the ability to start from those laws and reconstruct the universe" (Anderson, 1972).

(1976), many other areas of science lack formal definitions, but make significant progress through qualitative and informal theories— from cell biology to plate tectonics.

If computer science was once primarily, as Knuth said, "the study of algorithms" (Knuth, 1974), then machine learning is the study of learning algorithms and the data they learn from. We can use analytical tools to describe the algorithms, but primarily empirical tools to probe their data.

4 Our Proposal

We propose that ML conferences introduce a subject area specifically for "Experimental Science of Machine Learning," and explicitly include this in their call-for-papers. Moreover, we propose a best-paper-award for submissions in this area. To facilitate this, we present a candidate call-for-papers for this subject area, and discuss goals and evaluation of experimental science.

4.1 Call for Papers: Experimental Science of Machine Learning

We invite submissions which conduct experimental investigation into the nature of learning and learning systems. We welcome scientific work which involves either purely experimental characterization, or a synthesis of experiment and mathematical theory. We do not attempt to enumerate the full scope of what may be considered "scientific work"— which is expansive and ever-growing. But as guiding principles: (1) the primary motivation is to understand, rather than to improve, and (2) the results identify structure in Nature— they teach us something about what was previously unknown or uncertain. For ML in particular, "Nature" means computational learning systems— those used in the past, present, future, and even those never to be used but instructive to conceive. And "structure" can take many different forms, from precise quantitative conjectures to general qualitative principles. We expect that, as an experimental call, most submissions will contain at least one experiment that teaches something new to a reasonable fraction of the community[9]. We elaborate on potential types of papers, and give guidance to authors in the following sections.

To limit scope, and develop a focused community, we exclude the following types of work which are already served by existing venues:

1. Primarily theoretical work on learning, of the type that would be welcomed at COLT, ALT, or the Theory subject-area in general ML conferences (NeurIPS, ICML, ICLR).
2. Primarily technological (or application-motivated) work in learning, of the type that would be welcomed at domain-specific venues (e.g. CVPR), or existing areas of general ML conferences.

4.2 Types of Papers

We describe several prototypical types of scientific papers in this subject area. This is far from a comprehensive list[10]

Surprising Experiments. This type of paper performs a new and "surprising" experiment. A well-designed experiment alone is worthy to be published, with or without a candidate explanation.

We now attempt to define "surprising" in the ML context. This is subtle: Kuhn (1962) defines surprise as when an experiment is not accounted for by (or is inconsistent with) the dominant scientific paradigm. This does not apply to many areas of machine learning, such as deep learning, because *there is no dominant scientific paradigm.* That is, there is not yet a common conceptual framework and methodology used by researchers for analyzing deep learning systems: different communities apply different paradigms (from statistics, computational learning theory, physics, cognitive science, etc.), and derive results which can be inconsistent with each other (and even with themselves). The field is in a Kuhnian crisis of sorts, exploring many possible frameworks in search of one that will be

[9]A new experiment is not required, however. For example, we welcome work which unifies existing experimental behaviors under a general empirical conjecture.

[10]Attempting to strictly define types of "good scientific work" encounters similar issues as attempting to define "good mathematics"— issues discussed by Terence Tao in Tao (2007). We thus only present frequent themes and prototypical examples.

appropriate for modern ML. This makes defining "surprise" especially subtle— surprising to whom? We propose the following informal sociological definition of surprise: an experiment is surprising if a reasonable fraction[11] of experts in the ML community incorrectly predict the experiment's result. That is, surprise is implied by community disagreement.

Finally, this is not the only reasonable definition of surprise: an experiment can also be deemed surprising if it disagrees with the predictions or intuitions of a certain specific research paradigm— for example, as experiments on overparameterized models disagreed with the intuitions from the bias-variance tradeoff in statistics.

Empirical Conjectures. This type of paper unifies a set of experiments via a general conjecture— describing (as precisely as appropriate) the observed behavior, and the range of settings for which it holds. It is the analogue of conjecturing Kepler's laws[12] after the experimental observations of Tycho Brahe. In machine learning, this type of paper can lift a behavior observed in *one particular experimental setting* (a particular architecture and dataset, for example), into a more general claim attempting to characterize learning systems which exhibit this behavior. Experimental conjectures need not be fully precise, general, or even completely correct to be scientifically insightful— refining conjectures along these axes is the normal process of science.

Refining Existing Phenomena. This type of paper builds on a prior experimental observation or conjecture, and extends it in at least one way. This can take many forms, including

1. Generalizing scope: Demonstrating that the phenomenon holds in more general settings than was previously known. Either by conducting further experiments, proposing a more general conjecture, or both.

2. Isolating causes: Simplifying the experimental setting, to better understand which factors are important for a certain behavior. This includes ablating factors of a complex system with surprising behaviors (for example, CLIP Radford et al. (2021)) to determine which factors are necessary and sufficient for the surprise (e.g. data diversity, sample size, architecture, loss function, etc). It also includes studying simpler or vastly different learning systems (e.g. random forests vs. neural networks) to understand which learning systems do or do not exhibit a certain behavior.

3. Increasing precision: Addressing limitations in existing conjectures, by making them more precise in either their predictions or their assumptions.

4. Refuting or refining conjectures: Presenting experimental counterexamples to prior conjectures, thus helping refine them to be more correct. For example, the Michelson–Morley experiment refuting the "luminiferous aether" theory of light (Michelson & Morley, 1887).

Formalizing Intuition. This type of paper takes a "folklore" intuition and attempts to formalize it as precisely as possible, and test it empirically. This may involve proposing new definitions to capture notions which were hitherto vaguely defined. In machine learning, for example, one may start with the folklore that "deep neural networks learn good representations when trained on real data", and attempt to formalize this by proposing concrete definitions of each object: representations, real data, and deep neural networks. This type of paper will be "unsurprising" to most, since it is based on existing folklore. However, the process of formalizing and testing intuition is itself scientifically valuable: it allows us to better understand the scope and limitations of our knowledge, and to to better identify flaws in our intuition (which are often revealed in the process of formalization).

New Measurements. This type of paper presents a new measurement tool (analogous to the microscope, or the telescope), and argues that this tool can enable new kinds of scientific discovery. A prototypical example of this kind of paper is the line of work on "feature visualization" (Olah et al., 2017), which introduced a new tool for inspecting networks, and revealed interesting qualitative structures in trained networks. Our description of this as a scientific tool follows the presentation of Olah et al. (2020).

[11] We leave "reasonable" up to interpretation, but we intend something closer to 20% than 99%.

[12] Note that Kepler's laws were indeed "just" experimental conjectures at the time, since they preceded Newtonian mechanics.

4.3 REVIEWER GUIDELINES

Reviewing scientific work in a rapidly evolving field is challenging: there are no "short cuts" for reviewers, no "benchmarks" measuring scientific quality, not even universally accepted central questions for the field. Each paper must be evaluated on an individual basis. Nevertheless, we describe here some general principles which can guide reviewers in the "science of ML" subject area. Only a subset of these guidelines will apply for any particular paper, depending on the type of paper (whether it introduces a conjecture, presents an experiment, etc). We focus on aspects of reviewing which are specific to this subject area, omitting those aspects common to peer reviewing in general (e.g. clarity of presentation, discussion of related work, etc).

Evaluating Surprise. Reviewers should be generous in evaluating whether an experiment that is claimed to be "surprising" is in fact surprising. Our proposed definition of surprise involves determining if a "reasonable fraction" of experts in the community would disagree with the experimental result. However, the reviewer is an individual, and cannot know the opinions of all experts in the community. The reviewer should thus use their best judgement, and *err on the side of generosity*, in evaluating such claims. Reviewers *should not* reject a paper simply because they are personally not surprised. There is little harm in publishing an unsurprising experiment, but serious harm in suppressing a surprising one.

We can also confirm surprise via the following hypothetical: If the overwhelming majority of experts in the community could be convinced of both the true experimental conclusion, *and its negation*, then the paper should be accepted, since it reduces inconsistency in beliefs.

Evaluate Only Stated Claims. Reviewers should evaluate only the paper's stated claims, and in context of the paper's stated motivations. For example, a paper that attempts to formalize a folklore intuition should not be rejected because it is "unsurprising," when the paper never claimed surprise. Similarly, a paper that presents a surprising experiment should not be criticized because it fails to theoretically explain the experiment.

Justification and its Limitations. Claims should be justified reasonably, at the level of generality that they are made. In experimental work, unlike theoretical work, it is often impossible or infeasible to justify a claim fully rigorously.[13] Rather, the best we can do is try to falsify our conjectures in reasonable ways, and acknowledge the limitations of both our conjectures and our experimental verification. Reviewers should check that authors have made reasonable effort to validate claims, and have described potential limitations of their validation.

A reviewer can ask for more experiments, but should specify exactly which claim is unsupported by the current experiments, and which specific new experiments would address this gap.

Precision and its Limitations. For papers which propose a quantitative conjecture, authors should attempt to state the conjecture as precisely and formally as possible, including quantifiers over all relevant quantities (e.g. sample size, distribution, etc). The gold standard of an empirical conjecture is a "theorem without a proof": a statement that is sufficiently formal to admit a proof. This is often not possible in machine learning, for good reason: we do not yet have formal definitions of many objects involved (such as the "natural distributions" of real data, and the "natural architectures" used in deep learning). These definitional obstacles to formalism in deep learning are described in Nakkiran (2021, Section 1.3.2). However, it is still possible to make scientific progress without formal definitions— just as biologists can study "life" without precisely defining life. We propose that authors attempt to state conjectures as precisely as possible, and acknowledge reasons for their imprecision: is it due to known definitional obstacles, new definitional obstacles, or an incomplete experimental understanding? Partial or informal conjectures are still scientifically valuable, and can lead to greater precision in future work. However, it is important to establish where and why claims are imprecise, to better understand the scope of our knowledge.

Theoretical Claims. Although this is primarily an experimental call-for-papers, we welcome incorporating theoretical results as appropriate: theory and experiment are complementary tools in scientific work. As with all theoretical work, reviewers should evaluate theory based on the insight

[13] This is because empirical conjectures will often involve quantifiers which can never be fully tested in finite time. For example, we are rarely interested in a claim about one specific experiment, but rather in claims that involving certain universal quantifiers (e.g. "for all sample sizes" or "for all sufficiently-large network widths"). Such quantifiers cannot be verified by experiments alone.

it brings (either in the result or the approach), and not based on the technical difficulty of proofs. This is consistent with principles in the Theoretical CS community:

> *"Although non-trivial mathematics plays a significant role in our research, we are in the business of understanding the nature of computation, not enumerating difficult-to-prove theorems about it."* (Parberry, 1989).

Axes: Generality, Precision, and Justification. For most if not all types of scientific work, increasing its *generality*, *precision*, and *experimental justification* will increase the strength of the paper. An interesting behavior observed in a single experimental setting could be strengthened if it were observed in a wide variety of settings. A conjecture about a particular learning method (e.g. SGD on a given architecture) could be strengthened if it held for an entire family of methods. A qualitative phenomena observed when training on one dataset (say, emergence of semantic feature visualizations) could be strengthened if it occurred on many diverse datasets and sample sizes. A claim about "most" training distributions could be straightened by specifying exactly which distributions. And all claims supported by experimental validation are strengthened by more extensive experiments, pushing the boundaries of the experimental claims. These overarching axes— of generality, precision, and justification— can be encouraged by reviewers to help improve the strength of papers.

5 Conclusion

We highlighted a structural problem in ML publishing: the lack of support for experimental scientific work that neither proves a mathematical theorem, nor improves a practical application. We discussed the historical importance of experimental science, and presented a concrete proposal to encourage it: a Best Paper Award for experimental scientific work at top ML conferences. To help implement this proposal, we presented a candidate call-for-papers and reviewer guidelines for the award.

References

P. W. Anderson. More is different. *Science*, 177(4047):393–396, 1972. doi: 10.1126/science.177.4047.393. URL https://www.science.org/doi/abs/10.1126/science.177.4047.393.

Amnon H Eden. Three paradigms of computer science. *Minds and machines*, 17(2):135–167, 2007.

Oded Goldreich and Avi Wigderson. Theory of computing: a scientific perspective. *ACM Computing Surveys (CSUR)*, 28(4es):218–es, 1996.

Charles Antony Richard Hoare. *The Mathematics of Programming*. Prentice-Hall, Inc., USA, 1989. ISBN 0132840278. URL https://dl.acm.org/doi/10.5555/63445.C1104376.

Donald E Knuth. Computer science and its relation to mathematics. *The American Mathematical Monthly*, 81(4):323–343, 1974.

Thomas Kuhn. *The structure of scientific revolutions*. University of Chicago Press, 1962.

Albert A Michelson and Edward W Morley. On the relative motion of the earth and of the luminiferous ether. *Sidereal Messenger, vol. 6, pp. 306-310*, 6:306–310, 1887.

Preetum Nakkiran. *Towards an Empirical Theory of Deep Learning*. Doctoral dissertation, Harvard University Graduate School of Arts and Sciences, 2021. URL https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37370110.

Allen Newell and Herbert A. Simon. Computer science as empirical inquiry: Symbols and search. *Commun. ACM*, 19(3):113–126, mar 1976. ISSN 0001-0782. doi: 10.1145/360018.360022. URL https://doi.org/10.1145/360018.360022.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

Ian Parberry. A guide for new referees in theoretical computer science. *ACM SIGACT News*, 20(4): 92–99, 1989.

Paweł Polak. Computing as empirical science–evolution of a concept. *Studies in Logic, Grammar and Rhetoric*, 48(1):49–69, 2016.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Anna Rogers. Peer review in nlp: reject-if-not-sota, Apr 2020. URL https://hackingsemantics.xyz/2020/reviewing-models/.

Anna Rogers and Isabelle Augenstein. How to review for acl rolling review, Feb 2021. URL https://aclrollingreview.org/reviewertutorial#6-check-for-lazy-thinking.

David Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. Winner's curse? on pace, progress, and empirical rigor. 2018.

Herbert A Simon. Artificial intelligence: an empirical science. *Artificial Intelligence*, 77(1):95–127, 1995.

Terence Tao. What is good mathematics? *Bulletin of the American Mathematical Society*, 44(4): 623–634, 2007.

Peter Wegner. Research paradigms in computer science. In *Proceedings of the 2nd international Conference on Software Engineering*, pp. 322–330. Citeseer, 1976.