Exploring the Deep Learning Toolbox for Mathematics

Liam Hodgkinson

Mathematics provides a fascinating use case for deep learning.

An entirely new paradigm beyond images / text / tabular data.

This is the new frontier, and it is gaining a lot of momentum in ML

Continuous

PDE Theory
Differential Geometry

The neural network is a regular function approximator; provides a unified performant strategy across multiple spatial dimensions.

Discrete

Combinatorics
Algebra

Learning patterns, underlying structure, circuits with models

Demands are very different!

Discrete Problems

Heuristic

PatternBoost

Finding counterexamples or guiding intuition; ***transformers*** perform extremely well (closer to classical ML territory).

Liu, Bingbin, et al. "Transformers learn shortcuts to automata." *arXiv preprint arXiv:2210.10749* (2022).

Theorem Discovery

Learning Group Operations

(Near-)exact solutions; interpretability is critical; depends strongly on model bias.

Chughtai, Bilal, Lawrence Chan, and Neel Nanda. "A toy model of universality: Reverse engineering how networks learn group operations." *International Conference on Machine Learning*. PMLR, 2023.

The Transformer

Theorem (Liu et al., 2023):
Transformers can simulate all solvable semiautomata $\mathcal{A} = (Q, \Sigma, \delta)$ with depth $\mathcal{O}(|Q|^2 \log |Q|)$, embedding dimension and attention width $2^{\mathcal{O}(|Q| \log |Q|)}$ and MLP width $|Q|^{\mathcal{O}(2^{|Q|})} + 2^{\mathcal{O}(|Q| \log |Q|)} n$ where n is the sequence length.

This global efficiency is unparalleled among machine learning models.

Transformers are even more effective at finding shortcuts.

The Transformer

Theorem (Liu et al., 2023):
Transformers can simulate all solvable semiautomata $\mathcal{A} = (Q, \Sigma, \delta)$ with depth $\mathcal{O}(|Q|^2 \log |Q|)$, embedding dimension and attention width $2^{\mathcal{O}(|Q| \log |Q|)}$ and MLP width $|Q|^{\mathcal{O}(2^{|Q|})} + 2^{\mathcal{O}(|Q| \log |Q|)}$ where T is the sequence length.

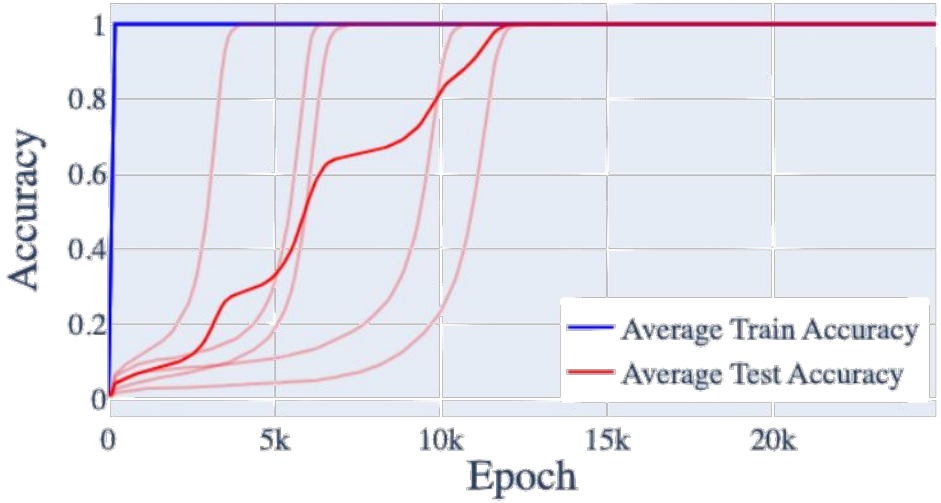
This global efficiency is unparalleled among machine learning models.

Transformers are even more effective at finding shortcuts.

Unfortunately, the reality is not so simple...

Grokking

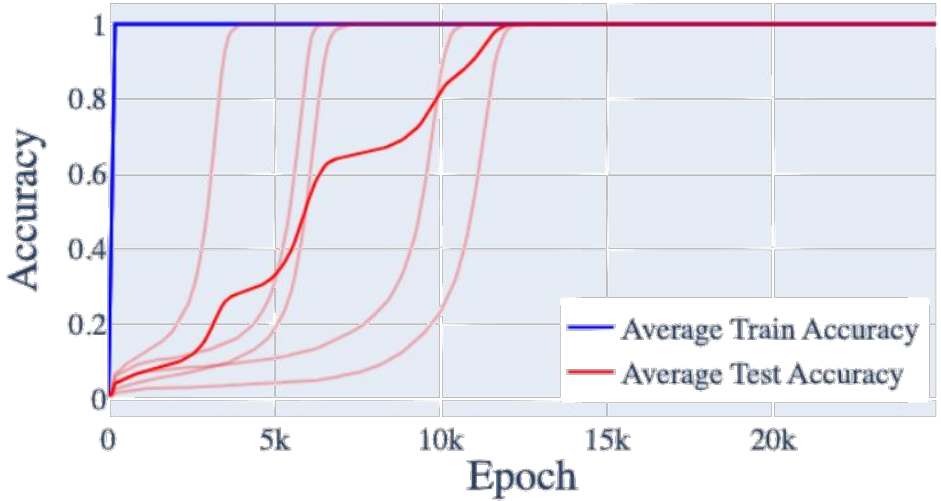
Nanda, Neel, et al. "Progress measures for grokking via mechanistic interpretability." *arXiv preprint arXiv:2301.05217* (2023).



Remark: If test accuracy steadily decreases once training accuracy is at 100%, *grokking will not occur with the current setup.*

Grokking

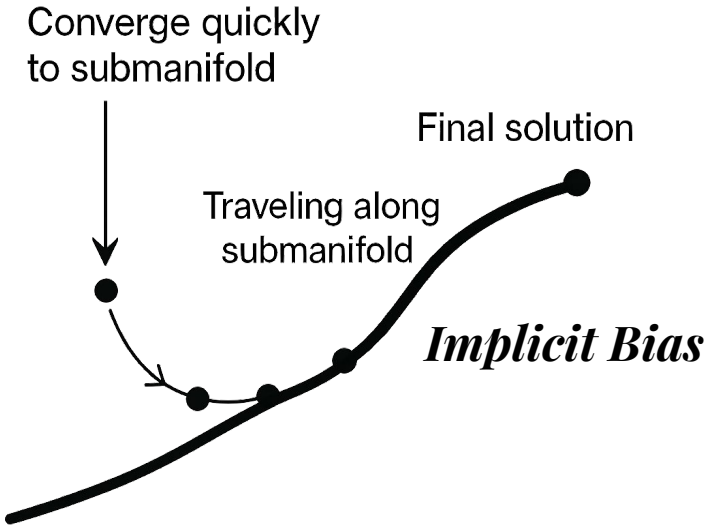
Nanda, Neel, et al. "Progress measures for grokking via mechanistic interpretability." *arXiv preprint arXiv:2301.05217* (2023).



Remark: If test accuracy steadily decreases once training accuracy is at 100%, *grokking will not occur with the current setup.*

"Transition from a lazy training regime to a rich training regime"

Lyu, Kaifeng, et al. "Dichotomy of early and late phase implicit biases can provably induce grokking." *arXiv preprint arXiv:2311.18817* (2023).



The Implicit Bias of Depth

Giampouras, Paris, et al. "A novel variational form of the Schatten-p quasi-norm." *Advances in Neural Information Processing Systems* 33 (2020): 21453-21463.

Weight decay introduces L^2 regularisation into the loss function (Frobenius norm of weight matrices); this induces **low-rank structure** in models involving compositions of matrices.

$$\underbrace{\|W\|_{\mathcal{S}_{2/n}}^{2/n} = \sum_i \sigma_i(W)^{2/n}}_{\textbf{Schatten norm}} = \min_{W=W_1\cdots W_n} \frac{1}{n}\left(\|W_1\|_F^2 + \cdots + \|W_n\|_F^2\right)$$

increasing depth + weight decay $\Longrightarrow$ **reduced rank in multi-layer networks**

The Pattern Avoidance Problem

- **Multi-layer perceptron (MLP):** Not high accuracy, but exhibits **low rank** structure with more layers — ripe for interpretation.

- **Transformers:** Better performance, exhibit symmetries inherent in the problem, seem to be (roughly) following the human theorem. *Smarter tokenisation* may give perfect accuracy.

- **Gated recurrent unit (GRU):** Perfect (100%) validation accuracy within few epochs; fewer parameters, but no low-rank structure.

Many other architectures exist ***built in to PyTorch*** and may be more effective (better alignment with the "algorithm" we look to discover).

Finding Polynomial Coefficients

Can we get a neural network to learn how to perform a change-of-basis between two (multivariate) polynomial bases?

This is a linear map for a fixed degree, but we would like it to be able to extrapolate to larger degrees.

Kostka numbers (monomials ↔ Schur) can be computed by recursion: *can a network learn this?*

Finding Polynomial Coefficients

Can we get a neural network to learn how to perform a change-of-basis between two (multivariate) polynomial bases?

This is a linear map for a fixed degree, but we would like it to be able to extrapolate to larger degrees.

Kostka numbers (monomials ↔ Schur) can be computed by recursion: *can a network learn this?*

We've had a lot of failures!

Regimes of Learning

Yang, Yaoqing, et al. "Taxonomizing local versus global structure in neural network loss landscapes." *Advances in Neural Information Processing Systems* 34 (2021): 18722-18733.

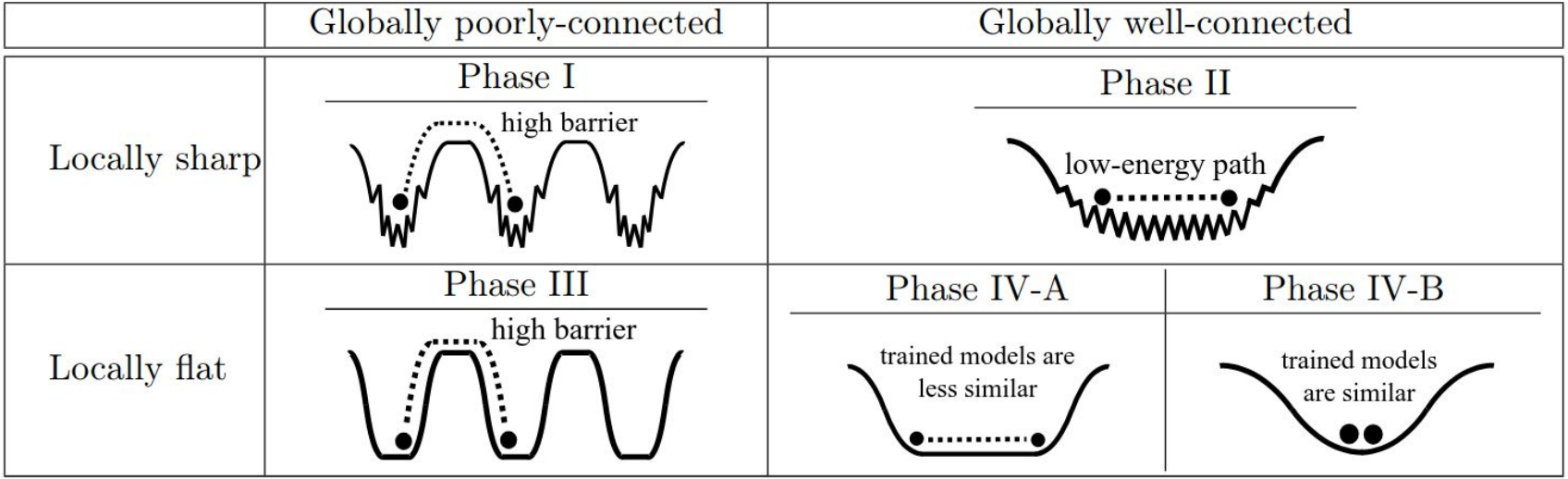
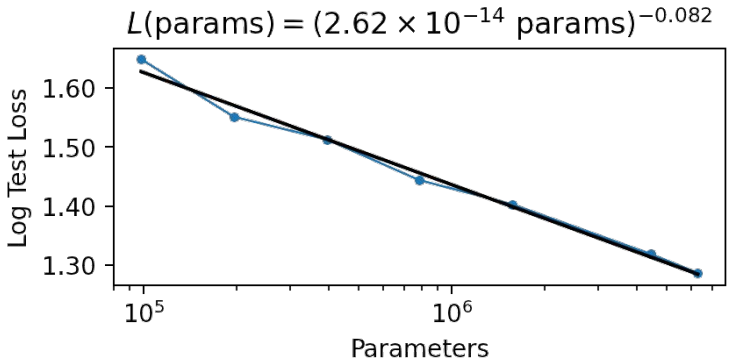


Figure 1: **(Caricature of different types of loss landscapes).** Globally well-connected versus globally poorly-connected loss landscapes; and locally sharp versus locally flat loss landscapes. Globally well-connected loss landscapes can be interpreted in terms of a global "rugged convexity"; and globally well-connected and locally flat loss landscapes can be further divided into two sub-cases, based on the similarity of trained models.

Practical Considerations

Neural Scaling Law $L \sim cd^{-\nu}$



...if the implicit bias is adapted to the problem.

- **Increasing model size** should increase "global connectivity"; if increasing model size makes things worse, *you should change something.*
- Higher learning rate / smaller batch size combinations (temperature) ensure locally stable solutions, but **good solutions should be easy to replicate** (minimal fiddling with hyperparameters).
- You may need more data for this to happen (especially for transformers!).

Continuous Problems

Discovery of Unstable Singularities

Yongji Wang[1, 2], Mehdi Bennani[3], James Martens[3], Sébastien Racanière[3], Sam Blackwell[3], Alex Matthews[3], Stanislav Nikolov[3], Gonzalo Cao-Labora[1, 4], Daniel S. Park[5], Martin Arjovsky[3, †], Daniel Worrall[3, †], Chongli Qin[3, †], Ferran Alet[3, †], Borislav Kozlovskii[3, †], Nenad Tomašev[3, †], Alex Davies[3], Pushmeet Kohli[3], Tristan Buckmaster[1,*], Bogdan Georgiev[3,*], Javier Gómez-Serrano[6,*], Ray Jiang[3,*] and Ching-Yao Lai[2,*]

[1]New York University, Department of Mathematics, New York, NY 10012, USA
[2]Stanford University, Department of Geophysics, Stanford, CA 94305, USA
[3]Google DeepMind, London, N1C 4DJ, UK
[4]École Polytechnique Fédérale de Lausanne, Institute of Mathematics, Lausanne, VA 1015, Switzerland
[5]Google DeepMind, New York, NY 10011, USA
[6]Brown University, Department of Mathematics, Providence, RI 02912, USA
[†]These authors have comparable contributions
[*]Corresponding author: buckmaster@cims.nyu.edu
[*]Corresponding author: bogeorgiev@google.com
[*]Corresponding author: javier_gomez_serrano@brown.edu
[*]Corresponding author: rayjiang@google.com
[*]Corresponding author: cyaolai@stanford.edu
[*]All corresponding authors contributed equally and are listed in alphabetical order

.AP] 17 Sep 2025

Discovery of Unstable Singularities

**Yongji Wang[1, 2], Mehdi Bennani[3], James Martens[3], Sébastien
Alex Matthews[3], Stanislav Nikolov[3], Gonzalo Cao-Labora[1, 4
Arjovsky[3, †], Daniel Worrall[3, †], Chongli Qin[3, †], Ferran Alet[3, †],
Tomašev[3, †], Alex Davies[3], Pushmeet Kohli[3], Tristan Buckma
Javier Gómez-Serrano[6,*], Ray Jiang[3,*] and Ching-Yao Lai[2,*]**

[1]New York University, Department of Mathematics, New York, NY 10012, US
[2]Stanford University, Department of Geophysics, Stanford, CA 94305, USA
[3]Google DeepMind, London, N1C 4DJ, UK
[4]École Polytechnique Fédérale de Lausanne, Institute of Mathematics, Lausa
[5]Google DeepMind, New York, NY 10011, USA
[6]Brown University, Department of Mathematics, Providence, RI 02912, USA
[†]These authors have comparable contributions
[*]Corresponding author: buckmaster@cims.nyu.edu
[*]Corresponding author: bogeorgiev@google.com
[*]Corresponding author: javier_gomez_serrano@brown.edu
[*]Corresponding author: rayjiang@google.com
[*]Corresponding author: cyaolai@stanford.edu
[*]All corresponding authors contributed equally and are listed in alphabetical

.AP] 17 Sep 2025

arXiv:2201.06780v3 [math.AP] 7 May 2023

Asymptotic self-similar blow-up profile for three-dimensional axisymmetric Euler equations using neural networks

Y. Wang,[1] C.-Y. Lai,[1] J. Gómez-Serrano,[2,3,4] and T. Buckmaster[5,6,*]

[1]*Department of Geosciences, Princeton University Princeton, NJ 08540, USA*
[2]*Department of Mathematics, Brown University, Kassar House, 151 Thayer St., Providence, RI 02912, USA*
[3]*Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Gran Via de les Corts Catalanes, 585, 08007, Barcelona, Spain*
[4]*Centre de Recerca Matemàtica, Edifici C, Campus Bellaterra, 08193 Bellaterra, Spain*
[5]*School of Mathematics, Institute for Advanced Study Princeton, NJ*
[6]*Department of Mathematics, Princeton University, Princeton, NJ 08544, USA*

(Dated: May 9, 2023)

Whether there exist finite time blow-up solutions for the 2-D Boussinesq and the 3-D Euler equations are of fundamental importance to the field of fluid mechanics. We develop a new numerical framework, employing physics-informed neural networks (PINNs), that discover, *for the first time*, a smooth self-similar blow-up profile for both equations. The solution itself could form the basis of a future computer-assisted proof of blow-up for both equations. In addition, we demonstrate PINNs could be successfully applied to find *unstable* self-similar solutions to fluid equations by constructing the first example of an unstable self-similar solution to the Córdoba-Córdoba-Fontelos equation. We show that our numerical framework is both robust and adaptable to various other equations.

PACS numbers: 47.10.-g, 07.05.Mh, 47.11.-j, 47.54.Bd

A celebrated open question in fluids is whether or not from smooth initial data the 3-D Euler equations may develop finite-time singularities (the inviscid analogue of the Navier-Stokes Millennium-prize problem). For non-smooth $C^{1,\alpha}$ initial data with $0 < \alpha \ll 1$, finite-time self-similar blow-up was proven in the groundbreaking work of Elgindi [1, 2]. The question of finite-time blow-up from *smooth* initial data remains *unresolved*.

In the presence of a cylindrical boundary, Luo and Hou [3] (cf. [4]) performed compelling numerical simulations proposing a scenario – sharing similarities with Pumir and Siggia [5] (cf. [6, 7]) – for finite time blow-up of the axi-symmetric 3-D Euler equations. They simulated the time-dependent problem and observed a dramatic growth in the maximum of vorticity (by a factor of $3 \cdot 10^8$), strongly suggesting formation of a singularity. The work is also suggestive of asymptotic self-similarity.

To confirm the existence of the finite-time singularity in the Luo-Hou scenario, and find its self-similar structure, we need to solve the self-similar equations associated with the axisymmetric 3-D Euler equations in the local coordinates near the singularity, which poses an cylindrical boundary is intrinsically linked to the same problem for the 2-D Bousinessq equations (cf. [4, 8–10]), another fundamental question in fluid mechanics, mentioned in Yudovich's *'Eleven great problems of mathematical hydrodynamics'* [11]. The mechanism for blow-up for the two equations is believed to be identical. The 2-D Boussinesq equations take the form

$$\begin{aligned} \partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} + \nabla p &= (0, \theta), \\ \partial_t \theta + \mathbf{u} \cdot \nabla \theta &= 0, \ \operatorname{div} \mathbf{u} = 0, \end{aligned} \tag{1}$$

where the 2-D vector $\mathbf{u}(\mathbf{x}, t)$ is the velocity and the scalar $\theta(\mathbf{x}, t)$ is the temperature. We consider the spatial variable $\mathbf{x} = (x_1, x_2)$ to be taken on the half plane $x_2 \geq 0$ and impose a non-penetration boundary condition at $x_2 = 0$ (x_1-axis), namely $u_2(x_1, 0) = 0$.

To search for singularity formation for the Boussinesq equations (1), we look for *backwards* [12] self-similar solutions of the form $\mathbf{u} = (1-t)^{\lambda} \mathbf{U}(\mathbf{y})$ and $\theta = (1-t)^{-1+\lambda} \Theta(\mathbf{y})$, where we define the self-similar coordinates as $\mathbf{y} = (y_1, y_2) = \frac{(x_1, x_2)}{(1-t)^{1+\lambda}}$ with $\lambda > -1$ yet to be determined. Under such a self-similar ansatz, the equations (1) become

.AP] 17 Sep 2025

Discovery o

Yongji Wang[1, 2], **Mel**
Alex Matthews[3], **Sta**
Arjovsky[3, †], **Daniel**
Tomašev[3, †], **Alex Da**
Javier Gómez-Serra

[1]New York University, De
[2]Stanford University, Dep
[3]Google DeepMind, Lond
[4]École Polytechnique Fé
[5]Google DeepMind, New
[6]Brown University, Depar
[†]These authors have con
[*]Corresponding author: b
[*]Corresponding author: b
[*]Corresponding author: j
[*]Corresponding author: r
[*]Corresponding author: c
[*]All corresponding autho

:2412.01184v1 [math.DG] 2 Dec 2024

Computationally-assisted proof of a novel $\mathsf{O}(3) \times \mathsf{O}(10)$-invariant Einstein metric on S^{12}

Timothy Buttsworth[1] and Liam Hodgkinson[2]

[1] *School of Mathematics and Statistics, The University of New South Wales, Kensington, Sydney, NSW 2052, Australia,* E-mail address: t.buttsworth@unsw.edu.au
[2] *School of Mathematics and Statistics , The University of Melbourne , Parkville, Melbourne, VIC 3052, Australia,* E-mail address: lhodgkinson@unimelb.edu.au

Abstract

We prove existence of a non-round Einstein metric g on S^{12} that is invariant under the usual cohomogeneity one action of $\mathsf{O}(3) \times \mathsf{O}(10)$ on $S^{12} \subset \mathbb{R}^{13} = \mathbb{R}^3 \oplus \mathbb{R}^{10}$. The proof involves using several rigorous numerical analysis techniques to produce a Riemannian metric $\hat{g}$ which approximately satisfies the Einstein condition to known high precision, and then demonstrating that $\hat{g}$ can be perturbed into a true Einstein metric g.

1 Introduction

Let M be a smooth manifold. An *Einstein metric* on M is a smooth Riemannian metric g for which there exists a constant λ so that

$$\mathrm{Ric}(g) = \lambda g \text{ on } M, \tag{1.1}$$

where $\mathrm{Ric}(g)$ is the Ricci curvature of g. Finding and classifying complete Einstein metrics on a given smooth manifold M (i.e., those Einstein metrics g for which the Riemannian metric space (M, g) is complete) is one of the most important topics in Riemannian geometry, mainly because the Einstein condition offers a coherent notion of what it means for a manifold's geometry to be 'ideal'. See [5] for a more thorough discussion of this idea, including why Einstein metrics are also 'ideal' from the perspective of physics.

If the dimension of M is at most 3, then any Einstein metric has constant sectional curvature, so the Killing-Hopf theorem implies that any complete Einstein manifold (M, g) must be a quotient of the round sphere, flat Euclidean space, or hyperbolic space. Starting in dimension 4, this classification result fails, and we are forced to use other methods in the study of the Einstein condition. In local coordinates, (1.1) is a system of quasi-linear, second order, non-elliptic partial differential equations, so one might hope to use PDE methods to solve it. While the local theory of solutions to this problem is now well-understood (see, for example, Chapter 5 of [5]), the global theory is much more complicated.

ree-dimensional axisymmetric Euler
al networks

[2, 3, 4] and T. Buckmaster[5, 6, *]
sity Princeton, NJ 08540, USA
e, 151 Thayer St., Providence, RI 02912, USA
a, Universitat de Barcelona,
08007, Barcelona, Spain
s Bellaterra, 08193 Bellaterra, Spain
anced Study Princeton, NJ
sity, Princeton, NJ 08544, USA
023)

2-D Boussinesq and the 3-D Euler equa-
mechanics. We develop a new numerical
(PINNs), that discover, *for the first time,*
The solution itself could form the basis of
tions. In addition, we demonstrate PINNs
olutions to fluid equations by constructing
Córdoba-Córdoba-Fontelos equation. We
daptable to various other equations.

drical boundary is intrinsically linked to the same
lem for the 2-D Bousinessq equations (cf. [4, 8–10]),
her fundamental question in fluid mechanics, men-
d in Yudovich's *'Eleven great problems of mathe-*
cal hydrodynamics' [11]. The mechanism for blow-
r the two equations is believed to be identical. The
Boussinesq equations take the form

$$\begin{aligned} \partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} + \nabla p &= (0, \theta), \\ \partial_t \theta + \mathbf{u} \cdot \nabla \theta &= 0, \ \operatorname{div} \mathbf{u} = 0, \end{aligned} \tag{1}$$

e the 2-D vector $\mathbf{u}(\mathbf{x}, t)$ is the velocity and the scalar
) is the temperature. We consider the spatial vari-
$\mathbf{x} = (x_1, x_2)$ to be taken on the half plane $x_2 \geq 0$ and
se a non-penetration boundary condition at $x_2 = 0$
xis), namely $u_2(x_1, 0) = 0$.
search for singularity formation for the Boussi-
equations (1), we look for *backwards* [12] self-
ar solutions of the form $\mathbf{u} = (1-t)^{\lambda}\mathbf{U}(\mathbf{y})$ and
$(1-t)^{-1+\lambda}\Theta(\mathbf{y})$, where we define the self-similar
linates as $\mathbf{y} = (y_1, y_2) = \frac{(x_1, x_2)}{(1-t)^{1+\lambda}}$ with $\lambda > -1$ yet
determined. Under such a self-similar ansatz, the
ions (1) become

Main Strategy

Proving existence of particular solution to $\mathcal{F}[\varphi] = 0$.

I. Obtain an approximate solution $\hat{\varphi}$ to *very high accuracy*.

II. Using interval arithmetic and rigorous / symbolic integration procedures, compute an *a posteriori error bound* $\|\mathcal{F}[\hat{\varphi}]\|$.

III. By rigorous computation (low prec.), find a *lower bound* for $\|(\mathcal{D}\mathcal{F}[\hat{\varphi}])^{-1}\|_{\mathrm{op}}$.

IV. Bound the second-order differential in the theory (in a nbhd of $\hat{\varphi}$).

V. If the bounds are "good enough", apply *fixed point theory*.

Deep Learning Finds Approximate Solutions

Warning: It will find *a* solution, often the most “regular” one.

Physics-Informed Neural Networks (PINNs)

Find a solution to a differential equation by solving the problem:

$$\inf_{\varphi} \|\mathcal{F}[\varphi]\|^2_{L^2(\Omega)} \approx \inf_{\varphi} \frac{1}{N} \sum_{i=1}^{N} \|\mathcal{F}[\varphi](x_i)\|^2,$$

where $x_1, \ldots, x_N$ are uniformly distributed on Ω. Can also include $\|\mathcal{D}\mathcal{F}[\varphi]\|^2_{L^2(\Omega)}$

This often fails

unless you restrict to a clever class of solutions (e.g. self-similar) for φ

Blow Up in Navier-Stokes: Wang, Yongji, et al. "Discovery of Unstable Singularities." *arXiv preprint arXiv:2509.14185* (2025).

Variational Problems (Mean Curvature Flow)

Let $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ denote an arbitrary diffeomorphism; $S_r,\ r \in \mathcal{A}$ is a sweepout of smooth manifolds embedded in $\mathbb{R}^n$.

Theorem: Any critical point $\varphi(S_r)$ to the following **min-max problem** is a mean curvature flow shrinker:

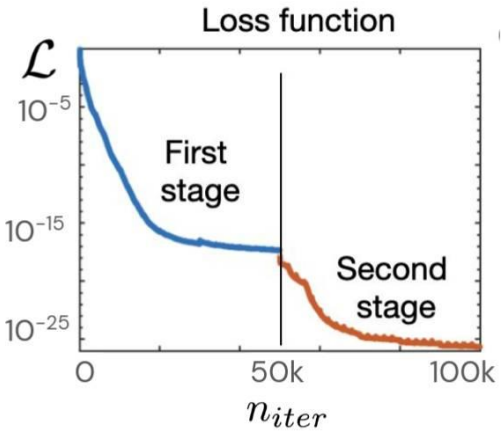
$$\inf_{\varphi} \sup_{r \in \mathcal{A}} \int_{\varphi(S_r)} e^{-\frac{1}{2n}\|y\|^2} d\sigma(y)$$

$$= \inf_{\varphi} \sup_{r \in \mathcal{A}} \int_{S_r} e^{-\frac{1}{2n}\|\varphi(z)\|^2 + \log|\det D\varphi(z)|} \|D\varphi(z)^{-\top}\nu(z)\| d\sigma(z)$$

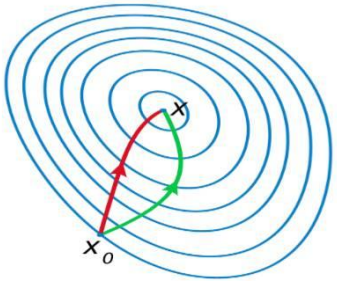
where $\nu(z)$ is a normal vector to S_r at z. **Better adapted to the problem.**

Why Use Neural Networks vs. Meshes?

Wang, Yongji, et al. "Discovery of Unstable Singularities." *arXiv preprint arXiv:2509.14185* (2025).



Gauss-Newton optimization



Exponential Convergence of the Deep Neural Network Approximation for Analytic Functions

Weinan E[1,2,3] and Qingcan Wang[4]

[1]Department of Mathematics and PACM, Princeton University
[2]Center for Big Data Research, Peking University
[3]Beijing Institute of Big Data Research
[4]PACM, Princeton University

Dedicated to Professor Daqian Li on the occasion of his 80th birthday

Abstract

We prove that for analytic functions in low dimension, the convergence rate of the deep neural network approximation is exponential.

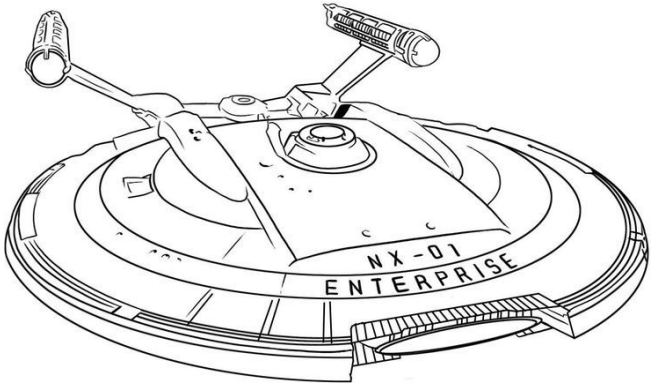
What Types of Models?

Emphasis in continuous problems is on the ***objective***, so the model choice is rarely critical.

- **Arbitrary Map:** Multi-layer perceptron (vanilla neural network)
- **Homeomorphism** / **Diffeomorphism**: Normalising / autoregressive flows; neural ODEs
- **Symmetries** / **Invariances:** Invariant and equivariant learning

Ma, George, et al. "A canonicalization perspective on invariant and equivariant learning." *Advances in Neural Information Processing Systems* 37 (2024): 60936-60979.

So What Now?



- Deep learning has a lot of tools for *all* mathematicians; it is still unclear how best to use them, but existing insights seem to help.

- **Mathematics offers a different paradigm** from other classical ML applications. Discrete applications might provide clarity toward the advantages of certain architectures and problem designs.

- More certainty, but still early days for continuous applications. Further engineering into high-precision neural network training is required.