Man, Machine, and Mathematics

Akshunna S. Dogra

Group Leader, Mathephysics
NSF IAIFI Fellow, Dept. of Physics, MIT
President's Ph. D. Scholar, Imperial College London

January 22, 2026

Mathematical Modeling

Problems

$$Lu = h$$

$$\frac{\partial u}{\partial t} = \Delta u + N(u) + h$$

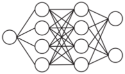
 or 

Architectures

$$\mathcal{N}(\mathbf{w}) = \begin{bmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{m1} & \cdots & w_{mn} \end{bmatrix}$$

$$\mathcal{N}(\mathbf{w}, \cdot) = \sum_{-M}^{M} w_m e^{i2\pi \frac{m}{P}(\cdot)}$$

Neural Network



Optimization

Customized $\mathcal{L}$ vs Default

Convex vs Non-convex $\mathcal{L}$

Strict vs Loose Constraints

$\mathcal{L}^{-1}[0]$ vs $\mathcal{L}[t] \xrightarrow[t \to \infty]{} 0$

Stochastic vs Deterministic

User Needs

???

Many-fold Learning: A generalized approach to modeling

- **Problem Setup**: Represent the problem as a map between separable spaces $\mathbf{F} : G \to H$.
- **Modelling Setup**: Parameterize the target space G using a modelling method (or architecture) $\mathcal{A} : \mathbb{R}^M \to G$.
- **Mathematical Analysis**: Prove tractability of gradient based optimization, usually by showing $\mathbf{F}$ is "well-behaved" and then carrying it over to $\mathbf{F} \circ \mathcal{A}$.
- **Optimization**: Optimize using an appropriate gradient flow variant.
- **Error Correction**: Boost initial performance by perturbing/expanding $\mathcal{A}$.

Gradient Flows: Are they too incredibly useful? Why? How?

- **Pros**: Gradient flows work for many problems with a small trick-set
- **Cons**: Problems $\rightarrow$ nonlinear, Architectures $\rightarrow$ nonlinear, Optimization $\rightarrow$ nonlinear $+$ stochastic, Formal analysis $\rightarrow$ intractable**??**
- **Motivation**: Build a generic theory covering many kinds of problems and architectures. Ideally, one that can be empirically useful
- **Inspirations**: Neural Tangent Kernels, Spectral Theory, PDEs, QM, QFT
- **Applications**: PDE solvers, Shape/Visual recognition, Classification, etc

A Universal Convergence Theorem

- **Problem**: $\mathbf{F} : G \to H$, where G, H are separable Hilbert spaces
- **Architecture**: $\mathcal{A} \in \mathscr{C}^2(\mathbb{R}^M, G)$, produces models in G, M is countable
- **Solution**: Φ, exists and the problem satisfies some conditions near it
- **Assumption** 1: Problem **F** is "well-behaved"

  Assumption 2: Architecture $\mathcal{A}$ is "well-initialised"
- **Claim**: A Gradient flow strategy will get us to Φ

Setting the table: A problem is well-behaved if

- it can be cast as a map between separable Hilbert spaces $\mathbf{F} : G \longrightarrow H$,
- $\exists \Phi \in G$ with a neighbourhood $\mathcal{B}_\Phi$, and a coercive **nominal loss** $\mathcal{L} \in \mathscr{C}^2(G, \mathbb{R})$, s.t. Φ is a global minimum for both $\langle \mathbf{F}[g] | \mathbf{F}[g] \rangle_H$ and $\mathcal{L}[g]$,
- $\mathcal{L}$ satisfies the following Lojasiewicz inequality (LI) [Ref. D]:

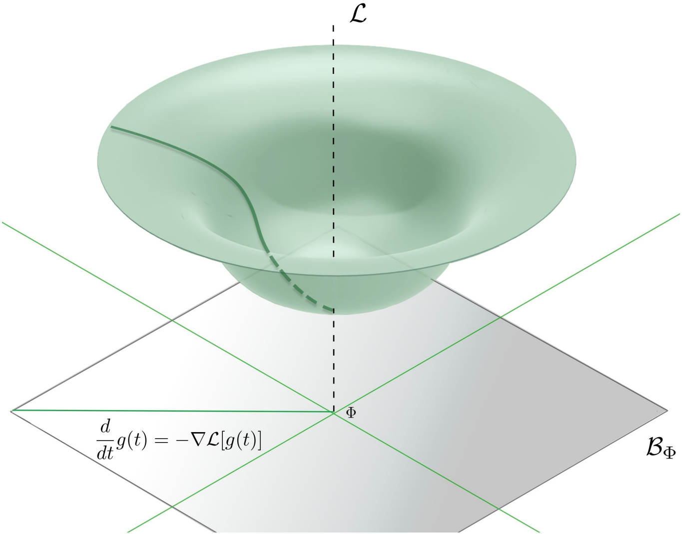
$$|\mathcal{L}[g] - \mathcal{L}[\Phi]|^\alpha \leq C \|\nabla \mathcal{L}[g]\|, \qquad g \in \mathcal{B}_\Phi,\ \alpha \in [1/2, 1),\ C > 0 \quad (1.1)$$

Theorem

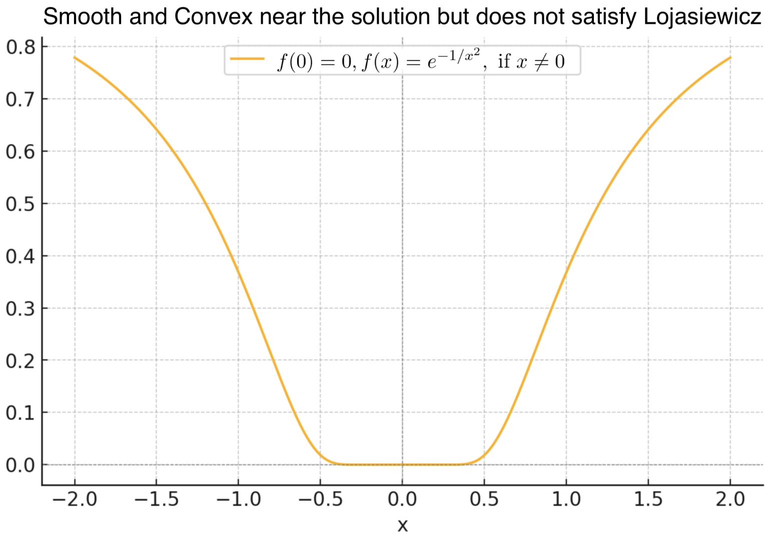
Let **F** *be well-behaved with the associated nominal loss being* $\mathcal{L}$. *Then*

$$\frac{d}{dt} g(t) = -\nabla \mathcal{L}[g(t)],\ g(0) \in \mathcal{B}_\Phi \implies \|g(t) - \Phi\| = \begin{cases} O(e^{-Ct}), & \text{if } \alpha = \frac{1}{2} \\ O(t^{\frac{-\alpha}{2\alpha - 1}}), & \text{if } \alpha > \frac{1}{2} \end{cases}$$

A well-behaved problem is solved by gradient flows (in principle)



CAUTION: LI is not Convexity and vice versa



Examples of well-behaved problems

- **Regression problems** [7], **Polynomial-fitting problems,** etc

- **Scientific ODEs/PDEs** [2, 5], such as the nonlinear Poisson Eqn (nPBE):

$$\mathbf{F}[g] = -\Delta g + \sinh(g) + h, \qquad G = W^{2,2}(\mathbb{R}^{n_{in}}), H = L^2(\mathbb{R}^{n_{in}})$$

- **Shape/Visual recognition** solvers [4] using Wasserstein distances b/w a Euclidean distribution(s) Φ and manifold(s) of model distributions $\mathcal{A}$:

$$\mathbf{F}[g] = \left(\int_{-\infty}^{x} \Phi(q)dq\right)^{-1} - \left(\int_{-\infty}^{x} g(q)dq\right)^{-1}, \quad G = \mathscr{P}_2(\mathbb{R}^d), H = \mathbb{R}$$

- **Classification problems** [3]. **Slide too small for F!!**, but $G = L^2(\Omega) \otimes \mathcal{H}$, where $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, $L^2(\Omega)$ is the space of square integrable Bochner measurable functions, and $\mathcal{H}$ is some apt Hilbert space.

Well-behaved problems seem easy enough. Why do we need Architectures?

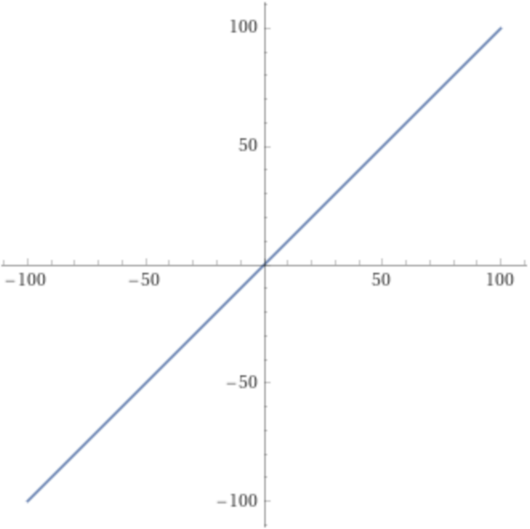
- Nominal loss dynamics are usually in infinite dimensional spaces

- Computers can only handle finite dimensional dynamics.
  We need to parametrize G through some M parameter architecture $\mathcal{A}$

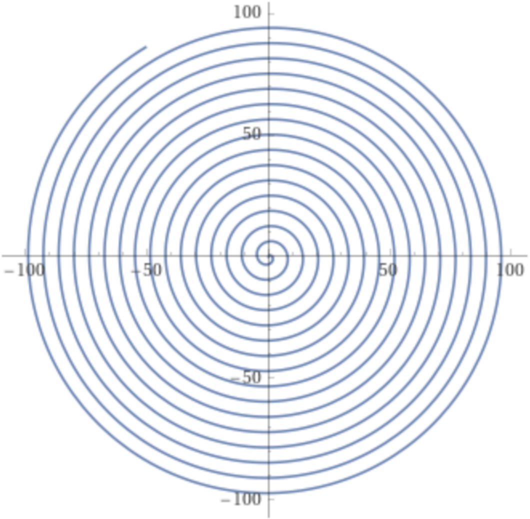
- Linear $\mathcal{A}$ are great if $\textbf{dist}(\Phi, \{\mathcal{A}(\mathbf{w}) : w \in \mathbb{R}^M\})$ is small. Such guarantees are impossible if $\dim(G) > M$, let alone $\dim(G) = \infty$

- Nonlinear $\mathcal{A}$ can use a finite number of parameters to cover G more densely than linear methods. But too many local minimum

1 Parameter, Linear Architecture



1 Parameter, Nonlinear Architecture



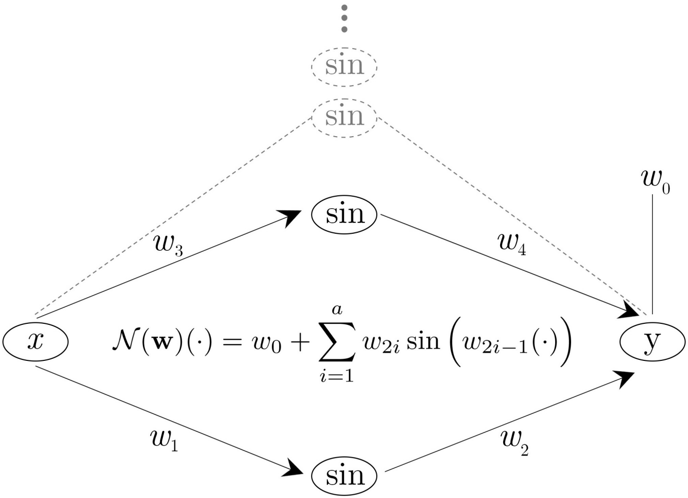
Setting the table: Architectures

- **Architectures**: $\mathcal{A} \in W^{2,2}(\mathbb{R}^M, G)$ are M-parameter maps that produce models $\mathcal{A}(\mathbf{w}) \in G$. Derivatives and adjoints are denoted by $\mathcal{A}_{\mathbf{w}}$ and $\mathcal{A}_{\mathbf{w}}^{\dagger}$.
- **Model set**: $G_M := \{\mathcal{A}(\mathbf{w})\} \subset G$: Set of models produced by $\mathcal{A}$.
- $\vartheta := \mathcal{A}_{\mathbf{w}}^{\dagger}\mathcal{A}_{\mathbf{w}}, \quad \Theta := \mathcal{A}_{\mathbf{w}}\mathcal{A}_{\mathbf{w}}^{\dagger}, \quad \mu(\mathbf{w}) = \inf(\mathrm{Spec}(\Theta) - \{0\})$
- Θ and ϑ share their non-zero spectrum.
- G_M is usually an M dimensional immersed submanifold in G near almost all $\mathcal{A}(\mathbf{w})$. Alternatively, ϑ is invertible almost everywhere on $\mathbb{R}^M$.
- $G_\infty = \bigcup_{M\in\mathbb{N}} G_M$ is dense in G (Universal Approximation Theorems [A]).

Architecture Examples

- Linear methods like Fourier Series, Chebyshev polynomials, etc
- Input/Output maps between Euclidean spaces modeled by Neural Nets
- Binary Classifiers
- Audio Signal Processing
- Image and Visual Recognition techniques

Architecture Examples: A “deep” Fourier Method



$$\mathcal{N}(\mathbf{w})(\cdot) = w_0 + \sum_{i=1}^{a} w_{2i} \sin\Big(w_{2i-1}(\cdot)\Big)$$

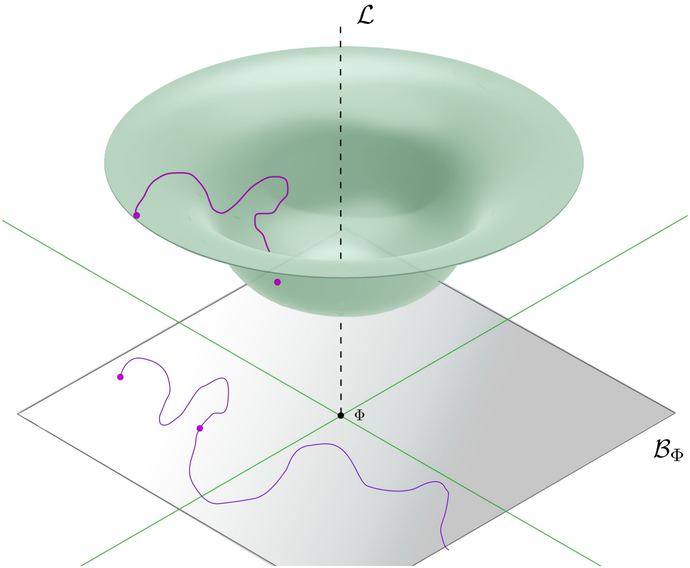
Parametric Optimization

- **Parametric Loss and Gradient Flow**: $\mathscr{L}[\mathbf{w}] \coloneqq \mathcal{L}[\mathcal{A}(\mathbf{w})]$, $\dot{\mathbf{w}} = -\nabla\mathscr{L}[\mathbf{w}]$
- **Well-initialization**: $\mathcal{A}$ is well-initialized with parameters $\mathbf{w}(0)$, if under a parametric gradient flow, there exists $t \in \mathbb{R}^+$ s.t. $\mathcal{A}(\mathbf{w}(t)) \in \mathcal{B}_\Phi$.
- Parametric gradient flows are given by the following equation:

$$\dot{\mathbf{w}}(t) = -\nabla_{\mathbf{w}}\mathscr{L}[\mathcal{A}(t)] = -\mathcal{A}_{\mathbf{w}}^{\dagger}\nabla\mathcal{L}[\mathcal{A}(t)] \tag{1.2}$$

- These translate into the following flows on the model-set G_M

$$\dot{\mathcal{A}}(t) = \mathcal{A}_{\mathbf{w}}\dot{\mathbf{w}}(t) = -\underbrace{\mathcal{A}_{\mathbf{w}}\mathcal{A}_{\mathbf{w}}^{\dagger}}_{\text{NTK}^{++}}\nabla\mathcal{L}[\mathcal{A}(t)] = -\Theta(t)\nabla\mathcal{L}[\mathcal{A}(t)] \tag{1.3}$$


$\mathcal{L}$
Φ
$\mathcal{B}_\Phi$

When is $\mathscr{L}$ well-behaved too?

Theorem

Assume **F** *is well-behaved with* $\mathcal{L}$ *as the associated nominal loss and* $\mathcal{A}$ *is well-initialised. Then* $\mathscr{L}[\mathbf{w}]$ *is a well-behaved map in the neighbourhood of all its critical points* $\mathbf{w}^*$ *s.t.* $\mathcal{A}(\mathbf{w}^*) \in \mathcal{B}_\Phi$*, if*

1. $\exists \alpha^* \in (0, 1/2], C^* > 0$, *s.t.* $|\mathscr{L}[\mathbf{w}] - \mathscr{L}[\mathbf{w}^*]|^{\alpha^*} < C^* \|\nabla \mathscr{L}[\mathbf{w}]\|$ *for all* $\mathbf{w} \in \mathcal{B}_{\mathbf{w}^*}$.
2. $\mathscr{L}$ *is analytic and* $M \in \mathbb{N}$.
3. $\mathcal{A}$ *is analytic and* $M \in \mathbb{N}$.
4. $\vartheta(\mathbf{w}^*)$ *is a Fredholm operator.*
5. G_M *is a "weakly" singular manifold.*

A simple example: Convex $\mathcal{L}$ and effectively linear $\mathcal{A}$

- In the large width regimes, we have: $a = \infty, \mathbf{F}[\mathcal{A}(\mathbf{w})] = \mathcal{A}(\mathbf{w}) - \Phi$
- If $\mathcal{L} = \langle \mathbf{F}[g] | \mathbf{F}[g] \rangle$, then $\nabla \mathcal{L}[\mathcal{A}(\mathbf{w})] = \mathcal{A}(\mathbf{w}) - \Phi$

$$\implies \dot{\mathcal{A}}(t) = -\Theta(t)[\mathcal{A}(t) - \Phi]$$

- As $a \to \infty$, the operator Θ tends to a static object, giving us

$$\mathcal{A}(t) = \Phi + e^{-\Theta(0)t}[\mathcal{A}(0) - \Phi]$$

- Even for finite M, we have

$$\mathcal{A}(t) = \Phi + e^{-\int_t^\infty \mu(s)ds}[\mathcal{A}(0) - \Phi]$$

- Classical NTK results obtained/generalized (with rigor) !!!!

How does mathematical analysis come into play for specific problems?

- Practically speaking, the strategy is to either directly establish $\mathscr{L}$ is well-behaved or that **F** is well-behaved and carry it over: $\mathcal{A}$ is an immersion almost everywhere for almost any conventional choice.
- **Example 1**: For the nPBE, we prove $D\mathbf{F}$ is invertible and pair the problem with an analytic $\mathcal{A}$ (see [2])
- **Example 2**: For infinite parameter regimes $\mu(t) = \mu(0)$ [B]
- **Example 3**: $\mathscr{L}[\mathbf{w}]$ is analytic for classification [3] and shape recognition [4]
- **Example 4**: In general, we can estimate $\mu(t)$ for real NNs (at huge costs [1]), **in situ** during optimization without needing new computations

When does the gradient flow for a fixed $\mathcal{A}$ stop?

- $\dot{\mathbf{w}}(t) = -\mathcal{A}_{\mathbf{w}}^{\dagger} \nabla \mathcal{L}[\mathcal{A}(t)] \implies \dot{\mathcal{A}}(t) = -\Theta(t) \nabla \mathcal{L}[\mathcal{A}(t)]$

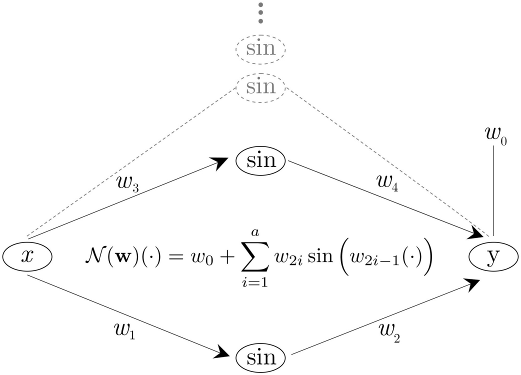
- $\mathbf{w}^*$ is a critical point iff one or more of the following conditions hold:

$$\text{(i) } \nabla \mathcal{L}[\mathcal{A}(\mathbf{w}^*)] = 0, \qquad \text{(ii) } \mu(\mathbf{w}^*) = 0, \qquad \text{(iii) } \nabla \mathcal{L}[\mathcal{A}(\mathbf{w}^*)] \in \ker(\Theta)$$

- LI ensures (i) holds only at $\mathcal{A}(\mathbf{w}) = \Phi$. If $\mathcal{A}$ is an immersion, as is the case in most applications, (ii) holds with 0 probability.

- We combat (iii) by either using a stochastic method with an apt annealing schedule or using iterative architecture expansions (IAE)

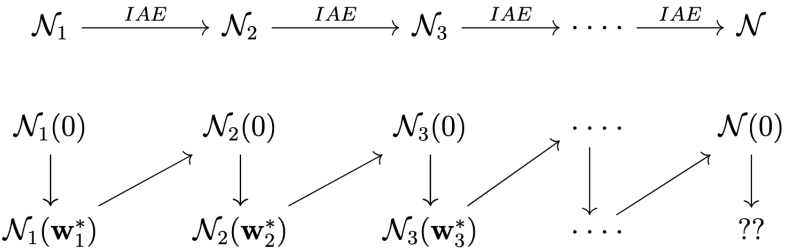
Iterative Architecture Expansions (IAE)



Lemma

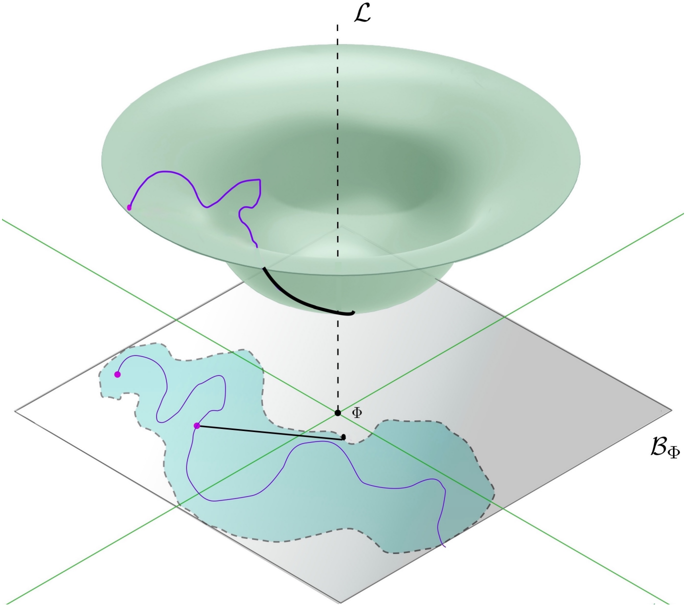
There exists an architecture $\mathcal{A} \in \mathscr{C}^2(\mathbb{R}^\infty, G)$ *s.t. given any* M *and any* $\mathbf{w} \in \mathbb{R}^M$, $\mathcal{A}(\mathbf{w}) = \mathcal{A}_M(\mathbf{w})$.

Error Correction via Iterative Architecture Expansions



Theorem

Assume G_∞ is a dense subspace of G, $\mathcal{A}_1$ is well-initialized, and each $\mathcal{A}_i$ is an immersion almost everywhere in $\mathbb{R}^{M_i}$. Then, $\mathcal{A}_i(\mathbf{w}_i^*) \xrightarrow[i\to\infty]{} \Phi$


$\mathcal{L}$
Φ
$\mathcal{B}_{\Phi}$

Applications for Neural Network Differential Equation solvers

Relative Errors across different F, $\mathcal{N}$, and optimization methods

System (Baseline Codebase)	Baseline (10^{-4})	IAE (10^{-4})	Total flops (10^3) (Baseline, EC)
1D+1D Burgers (RAR-PINN)	36.3	0.908	(7.3, 14.6)
2D+1D Henon Heiles (HNN)	12.9	0.0933	(10.5, 21.0)
2D+1D Heat (XPINN)	26.7	22.7	(0.3, 0.6)
1D+1D NL Oscillator (HNN)	4.76	0.00488	(10.3, 20.6)
2D nPBE (PINN)	436	0.37	(20.3, 40.6)
4D nPBE (PINN)	87.3	0.491	(20.5, 41.0)
2D Poisson (SPINN)	4.35	2.34	(4.9, 9.8)

Applications in shape and visual recognition, classification and anomaly detection, boosted optimization dynamics, and model reduction

- FINDER [3]: An anomaly detection tool + a general theory for binary classification that efficiently builds stochastic features that allow faster and more accurate identification in noisy datasets
- SHAPER [4]: a tool for defining and computing shape observables within collider physics datasets that generalizes several related methods
- Pruning [6]: Iterative Magnitude Pruning, a common sparsification tool in ML, was shown to be a renormalization process, with insight into how and where it does and does not work and how it could be more efficient.
- Koopman training [7]: A technique that identifies, estimates, and makes use of the Koopman operators associated with ML optimization dynamics to evolve parameters at lower computation costs

External References

A Kurt Hornik *et. al*, *Multilayer Feedforward Networks are Universal Approximators*, **Neural Networks** (Vol. 2, pp. 359–366), 1989

B Arthur Jacot *et. al*, *Neural Tangent Kernel: Convergence and Generalization in Neural Networks*, **NeurIPS** 31 (pp. 8571–8580), 2018

C M. Reed and B Simon, *Functional Analysis (Methods of Modern Mathematical Physics, Volume 1)*, **Academic Press**, 1980

D S. Lojasiewicz, *Une propriete topologique des sous-ensembles analytiques reels*, **Colloques internationaux du C.N.R.S 117. Les ´Equations aux D´eriv´ees Partielles**, 1963

Personal References

1. Many-fold Learning, AS Dogra, *under review*, 2025.
2. SoLvER: Solution Learning via Equation Residuals allows unsupervised error analysis and correction, AS Dogra *et. al*, *under review*, 2025.
3. FINDER: Feature Inference on Noisy Datasets using Eigenspace Residuals, T Murphy, AS Dogra *et. al*, *under review*, 2025.
4. SHAPER: Can You Hear the Shape of a Jet?, D Ba, AS Dogra, R Gambhir *et. al*, ***Journal of High Energy Physics*** 2023 (195), 2023.
5. Hamiltonian neural networks for solving equations of motion, M Mattheakis *et. al*, ***Physical Review E*** 105, 065305, 2022.
6. Universality of Winning Tickets: A Renormalization Group Perspective WT Redman, T Chen, Z Wang, AS Dogra, **ICML** 2022
7. Optimizing neural networks via Koopman operator theory, AS Dogra, W Redman, ***NeurIPS*** 33 (pp. 2087-2097), 2020.

People I am really grateful for

- Ms. Konstancja Maria Myszkowska (Austin Architects),
- Prof. Demba Ba (Harvard University, IAIFI),
- Prof. Julio Castrillon (Boston University),
- Dr. Sanchit Chaturvedi (NYU/Courant Institute)
- Mr. Rikab Gambhir (MIT, IAIFI),
- Prof. Mark Kon (Boston University),
- Mr. Jeffrey B. Lai (**Student**, UT Austin/Oden Institute),
- Prof. Jeroen Lamb (**Advisor**, Imperial College London/MRS CDT),
- Mr. Trajan Murphy (Boston University),
- Mr. Martin Peev (Imperial College London/MRS CDT),
- Dr. William T. Redman (APL/Johns Hopkins University),
- Prof. Jesse Thaler (MIT, IAIFI),
- Dr. Kevin Webster (**Advisor**, Imperial College London), and
- Mr. Dominic Wynter (UT Austin)