AI in the Sciences and Engineering 2024: Lecture 8

Siddhartha Mishra

Seminar for Applied Mathematics (SAM), D-MATH (and),
ETH AI Center,
ETH Zürich, Switzerland.

What you learnt so far

- Introduction to Deep Learning.
- Physics-Informed Neural Networks (PINNs) for solving PDEs.
  - Algorithms
  - Successes
  - Limitations
- Goal for the Today's Lecture:
  - Theoretical insights into why PINNs work and why they don't

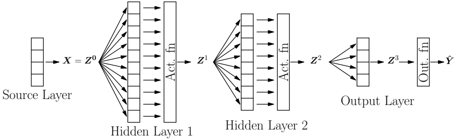
PDE forward problem

- Let X, Y be Function spaces with $Y = L^p(\mathbb{D}; \mathbb{R}^m)$.
- $\mathbb{D} = D$ or $\mathbb{D} = D \times (0, T)$, with $D \subset \mathbb{R}^d$
- Generic Abstract PDE:

$$\mathcal{D}(\mathrm{u}) = \mathrm{f},$$

- $\mathcal{D} : X \mapsto Y$ is the Differential operator, with input $\mathrm{f} \in Y$
- Boundary (Initial) conditions are implicit.
- Example: Heat Equation
  - $\mathcal{D} := \partial_t - \Delta$

Deep Neural networks



- $\mathrm{u}^*(y) = \sigma_o \odot C_K \odot \sigma \odot C_{K-1} \ldots\ldots\ldots \odot \sigma \odot C_2 \odot \sigma \odot C_1(y)$.
- At the k-th Hidden layer: $y^{k+1} := \sigma(C_k y^k) = \sigma(W^k y^k + B^k)$
- Parameters: $\theta = \{W_k, B_k\} \in \Theta$.
- Scalar Activation function σ
- Sigmoid, Tanh

PINNs for the PDE

- For Parameters $\theta \in \Theta$, $\mathrm{u}_\theta : \mathbb{D} \mapsto \mathbb{R}^m$ is a DNN, with $\mathrm{u}_\theta \in X^*$
- Aim: Find $\theta \in \Theta$ such that $\mathrm{u}_\theta \approx \mathrm{u}$ (in suitable sense).
- Compute PDE Residual by Automatic Differentiation:

$$\mathcal{R} := \mathcal{R}_\theta(y) = \mathcal{D}\left(\mathrm{u}_\theta(y)\right) - \mathrm{f}(y), \;\; y \in \mathbb{D} \quad \mathcal{R}_\theta \in Y^*, \quad \forall \theta \in \Theta$$

- PINNs are minimizers of $\|\mathcal{R}_\theta\|_Y^p \sim \int\limits_{\mathbb{D}} |\mathcal{R}_\theta(y)|^p \, dy$
- Replace Integral by Quadrature !
- Let $\mathcal{S} = \{y_i\}_{1 \leq i \leq N}$ be quadrature points in $\mathbb{D}$, with weights w_i
- PINN for approximating PDE is defined as $\mathrm{u}^* = \mathrm{u}_{\theta^*}$ such that

$$\theta^* = \arg\min_{\theta \in \Theta} \sum_{i=1}^{N} w_i |\mathcal{R}_\theta(y_i)|^p$$

- Minimize Very high-d Non-Convex loss with ADAM,L-BFGS

Do PINNs work ?

- Multi-D Heat Equation
- PINN with Depth 4, Width 20, Interior training points 2^{16}, Boundary points 2^{15}

Dimension	Training Error	Generalization error
1	2.8×10^{-5}	0.0035%
5	0.0002	0.016%
10	0.0003	0.03%
20	0.006	0.79%
50	0.006	1.5%
100	0.004	2.6%

- No Curse of dimensionality !!

When and Why do PINNs work for a PDE $\mathcal{D}(\mathrm{u}) = \mathrm{f}$?

- PDE solution u, DNN u_θ with parameters $\theta \in \Theta$
- AIM is to ensure small Total Error:

$$\mathcal{E}(\theta) := \|\mathrm{u} - \mathrm{u}_\theta\|_p$$

- PINNs may not have access to samples from Exact Solution u
- On the other hand, PINNs minimize PDE Residual:

$$\mathcal{E}_G(\theta) = \|\mathcal{R}_\theta\|_p = \|\mathcal{D}\left(\mathrm{u}_\theta\right) - \mathrm{f}\|_p$$

- In practice, we only have access to Training Error:

$$\mathcal{E}_T(\theta) = \left(\sum_{i=1}^{N} w_i |\mathcal{R}_\theta(y_i)|^p\right)^{\frac{1}{p}}$$

Key Theoretical Questions

- Is the PDE Residual small in the class of Neural Networks that approximate the exact solution u ? i.e.

$$\text{Does } \exists \hat{\theta}, \tilde{\theta} \in \Theta, \quad \mathcal{E}_G(\hat{\theta}), \mathcal{E}_T(\tilde{\theta}) < \epsilon, \text{ and } \mathcal{E} \sim \mathcal{O}(\epsilon) \text{ ?.}$$

- Does small PINN Residual $\Rightarrow$ small Total Error ? i.e.,
- Can we derive a bound of the form:

$$\mathcal{E}(\theta) \leq C\mathcal{E}_G(\theta), \quad \forall \theta \in \Theta$$

- Does small Training Loss $\Rightarrow$ small PINN Residual ? i.e.,
- Can we derive a bound of the form ?

$$\mathcal{E}_G(\theta) \leq \overline{C}\left(\mathcal{E}_T(\theta), N\right) \sim o\left(N^{-1}\right) \quad \forall \theta \in \Theta$$

On the smallness of PDE Residuals

- For sufficiently smooth u solving $\mathcal{D}(u) = f$ observe that

$$\mathcal{E}_G(\theta) = \|\mathcal{D}(u_\theta) - f\|_p = \|\mathcal{D}(u_\theta) - \mathcal{D}(u)\|_p \leq C(u, u_\theta)\|u - u_\theta\|_{W^{s,p}}$$

- Here s depends on the number of derivatives in the Differential Operator $\mathcal{D}$.
- Universal Approximation Theorems for DNNs:

$$\exists \hat{\theta} \in \Theta, \quad \|u - u_{\hat{\theta}}\|_{L^p} < \epsilon$$

- Extensions of (DeRyck,Lanthaler,SM, 2021): $\|u - u_{\hat{\theta}}\|_{W^{s,p}} < \epsilon$
- smoothness of u $\Rightarrow$ small PINN Residuals: $\mathcal{E}_G(\theta) \leq \epsilon$
- Smooth Activations + Sufficient Quadrature points:

$$\min_\theta \mathcal{E}_T(\theta) \leq \epsilon + o\left(N^{-1}\right)$$

On bounds on total error in terms of Residuals

- Sufficient Conditions of SM, Molinaro, 2021:
- Coercivity of the PDE $\mathcal{D}\mathrm{u} = f$: for any $\mathrm{u}, \bar{\mathrm{u}} \in X$:

$$\|\mathrm{u} - \bar{\mathrm{u}}\|_p \leq C_{pde}(\bar{\mathrm{u}}, \mathrm{u})\|\mathcal{D}(\bar{\mathrm{u}}) - \mathcal{D}(\mathrm{u})\|_p$$

- Coercivity $\Rightarrow$ Bounds in terms of Residuals as,

$$\begin{aligned}
\mathcal{E}(\theta) &= \|\mathrm{u}_\theta - \mathrm{u}\|_p, \\
&\leq C_{pde}(\mathrm{u}, \mathrm{u}_\theta)\|\mathcal{D}(\mathrm{u}_\theta) - \mathcal{D}(\mathrm{u})\|_p \quad (\text{Coercivity}), \\
&\leq C_{pde}(\mathrm{u}, \mathrm{u}_\theta)\|\mathcal{D}(\mathrm{u}_\theta) - f\|_p \quad \text{as } \mathcal{D}(\mathrm{u}) = f, \\
&\leq C_{pde}(\mathrm{u}, \mathrm{u}_\theta)\mathcal{E}_G(\theta) \quad (\text{Definition of } \mathcal{E}_G)
\end{aligned}$$

- Training Error $\mathcal{E}_T$ is Quadrature Approximation of $\mathcal{E}_G$:

$$\mathcal{E}_G \leq \mathcal{E}_T + C_{quad}(u_{\theta^*})^{\frac{1}{p}} N^{-\frac{\alpha}{p}} \quad \text{quadrature error},$$

A Strategy for PINN Error Bounds SM, Molinaro, 2021

- Use smoothness of exact solution to u of PDE $\mathcal{D}(\mathrm{u}) = f$
- And DNN approximation results in high-order Sobolev spaces to show that:

$$\exists \theta \in \Theta : \quad \mathcal{E}_G(\theta), \mathcal{E}_T(\theta) \leq \underline{C}(\mathrm{u}, \mathrm{u}_\theta) \|\mathrm{u} - \mathrm{u}_\theta\|_{W^{s,p}}.$$

- Use Coercivity of a given PDE to show that

$$\|\mathrm{u} - \mathrm{u}_\theta\|_p \leq \overline{C}(\mathrm{u}, \mathrm{u}_\theta) \mathcal{E}_G(\theta), \quad \forall \theta \in \Theta.$$

- Use Quadrature bounds to show that,

$$\mathcal{E}_G \leq \mathcal{E}_T + C_{quad}(u_{\theta^*})^{\frac{1}{p}} N^{-\frac{\alpha}{p}}$$

- Prove explicit growth bounds on the constants $\underline{C}, \overline{C}, C_{quad}$ in terms of Neural Network architecture and number of collocation points.

Kolmogorov PDEs

- Linear Parabolic PDEs of form:

$$\partial_t u = \sum_{i=1}^{d} \mu_i(x)\partial_{x_i} u + \frac{1}{2}\sum_{i,j,k=1}^{d} \sigma_{ik}(x)\sigma_{kj}(x)\partial_{x_i x_j} u,$$

$$u|_{\partial D\times(0,T)} = \Psi(x,t), \quad u(x,0) = \varphi(x)$$

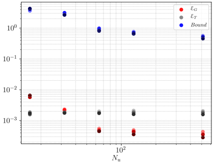
- μ, σ are Affine
- Examples:
  - Heat Equation: $\mu = 0$, $\sigma = ID$
  - Black-Scholes Equation for Option Pricing:
  - Interest rate μ, Stock Volatilities β and correlations ρ

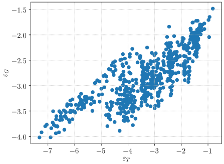
$$u_t = \sum_{i,j=1}^{d} \beta_i \beta_j \rho_{ij} x_i x_j u_{x_i x_j} + \sum_{j=1}^{d} \mu x_j u_{x_j}$$

- Note that $d >> 1$ (Very high-dimensional)

- $\exists$ Tanh PINN $\hat{u}$ of size $\mathcal{O}(\epsilon^{-\alpha(d)})$: $\mathcal{E}_{G,T}(\hat{\theta}) \sim \epsilon$,
- Uses Dynkin's formula to overcome curse of dimensionality.
- Stability of PDE: $\|u - u_\theta\|_2 \leq C \left(\|\mathcal{R}_{int,\theta}\| + \|\mathcal{R}_{sb,\theta}\|^{\frac{1}{2}} \right)$
- Use Hoeffding's inequality + Lipschitz bounds on u_θ:

$$\mathcal{E}_G^2(\theta) \sim \mathcal{O}\left(\mathcal{E}_T^2(\theta) + \frac{C\left(M, \log(\|W\|)\right)\log(\sqrt{N})}{\sqrt{N}} \right)$$





Numerical Results: (SM, Molinaro, Tanios, 2021)

- Heat Equation:

Dimension	Training Error	Total error
20	0.006	0.79%
50	0.006	1.5%
100	0.004	2.6%

- Black-Scholes type PDE with Uncorrelated Noise:

Dimension	Training Error	Total error
20	0.0016	1.0%
50	0.0031	1.5%
100	0.0031	1.8%

- Heston option-pricing PDE

Dimension	Training Error	Total error
20	0.0064	1.0%
50	0.0037	1.3%
100	0.0032	1.4%

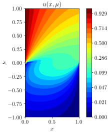
Radiative Transfer Equations

- $2d + 1$-dim Integro-Differential PDE for Intensity

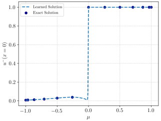
$$\frac{1}{c} u_t + \omega \cdot \nabla u + (k(x,\nu) + \sigma(x,\nu))\, u - \frac{\sigma(x,\nu)}{s_d} \int\limits_{R_+} \int\limits_{S} \Phi(\omega, \omega', \nu, \nu') u \, d\omega' d\nu' = f(x,t,n,\nu).$$

- High-dimensional, non-local, mixed-type, multiphysics
- PINNs applied and bound derived in SM, Molinaro 2020.

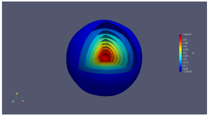
Numerical Results



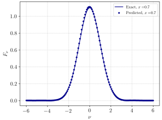
2-D, Intensity



2-D, Boundary



6-D, Inc. Radiation



6-D, Radial flux

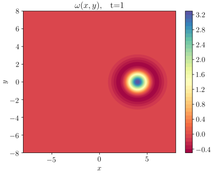
Dimension	Network Size	Error	Training Time
2	24×8	0.3%	57 min
6	20×8	2.1%	66 min

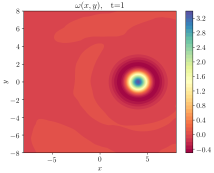
Navier-Stokes Eqns: $\mathrm{u}_t + (\mathrm{u}\cdot\nabla)\mathrm{u} + \nabla p = \nu\Delta u,\ \mathrm{div}\ \mathrm{u} = 0$

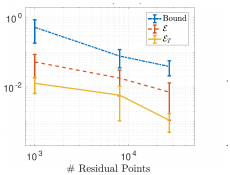
- Theory in DeRyck, Jagtap, SM, 2022.
- Smooth $u \in H^k$: PINN with size$(\hat{u}) \sim \mathcal{O}\left(M^{d+1}\right)$:
  $\mathcal{E}_G(\hat{\theta}) \leq \mathcal{O}\left(M^{1-k}\log(M)\right)$
- Use PDE theory to prove for $C = C\left(\|\mathrm{curl}\ u\|_{L^\infty}\right)$

$$\|u - u_\theta\|_2 \leq C\left(\|\mathcal{R}_{int,\theta}\| + \|\mathcal{R}_{tb,\theta}\| + \|\mathcal{R}_{sb,\theta}\|^{\frac{1}{2}} + \|\mathcal{R}_{div,\theta}\|^{\frac{1}{2}}\right)$$

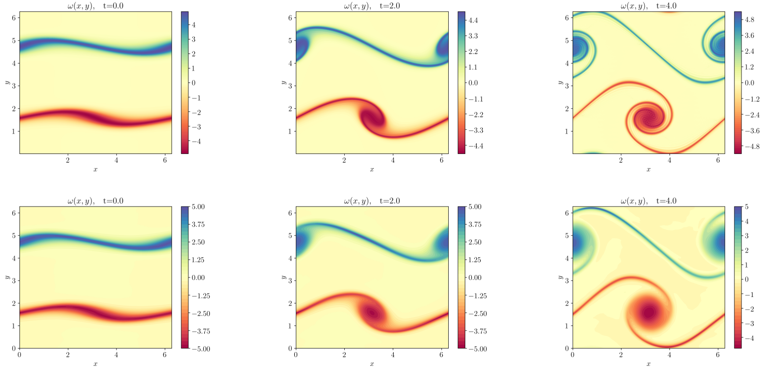
- Use Quadrature bounds: $\mathcal{E}_G^2(\theta) \sim \mathcal{O}\left(\mathcal{E}_T^2(\theta) + N^{-\alpha}\right)$





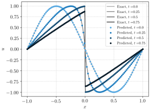


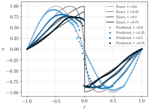
Results for 2-D Double Shear Layer

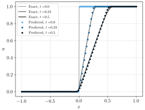


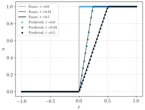
Viscous Burgers': $u_t + \mathrm{div}\ f(u) = \nu \Delta u$

- Error $\mathcal{E} \leq Ce^{CT}\left(\mathcal{E}_T + C_q N^{-\alpha}\right)$, $C = C\left(\|\nabla u^\nu\|_{L^\infty}\right)$
- $\|\nabla u^\nu\|_{L^\infty} \sim \frac{1}{\sqrt{\nu}}$ $\Rightarrow$ Error can blow up near shocks !!


$\nu = 10^{-3}$, Sh


$\nu = 0$, Sh


$\nu = 10^{-3}$, RF


$\nu = 0$, RF

ν	$\mathcal{E}$ (Shock)	$\mathcal{E}$ (Rarefaction)
10^{-3}	1.0%	2.2%
10^{-4}	11.2%	1.6%
0	23.1%	1.2%

- Alternatives: wPINNs of De Ryck, Molinaro, SM, 2023.

Summary (so far)

- For generic PDE: $\mathcal{D}(\mathrm{u}) = f$
- Rigorous Error estimate for PINNs:

$$\|\mathrm{u} - \mathrm{u}_\theta\| \sim C_{\rm pde}\left(\mathrm{u}, \mathrm{u}_\theta\right)\left[\mathcal{E}_T(\theta) + C_{\rm quad}(\mathrm{u}_\theta) N^{-\alpha}\right]$$

- Training Error is a blackbox
- We have that $\min_\theta \mathcal{E}_T(\theta) \leq \epsilon$
- But can we train to reach close to the global minimum ?

Theoretical Framework of De Ryck et al 2023

- Gradient Descent with Physics-Informed Loss:

$$\theta_{k+1} = \theta_k - \eta \nabla_\theta L, \quad L = \frac{1}{2} \int_D |\mathcal{D}(\mathrm{u}(x,\theta) - f(x)|^2 dx.$$

- Taylor Expansion:

$$\mathrm{u}(x,\theta_k) = \mathrm{u}(x,\theta_0) + \nabla_\theta \mathrm{u}(x,\theta_0)(\theta_k - \theta_0) + \langle H_k \theta_k - \theta_0, \theta_k - \theta_0 \rangle$$

- Rewritten GD: $\theta_{k+1} = (I - \eta \mathcal{A})\theta_k + \eta(\mathcal{A}\theta_0 + \mathcal{C}) + \eta \epsilon_k$
- Gram Matrix: $\mathcal{A}_{i,,j} = \langle \mathcal{D}\varphi_i, \mathcal{D}\varphi_j \rangle_{L^2}$, $\varphi_i = \partial_{\theta_i} \mathrm{u}(x,\theta_0)$
- Bias vector: $\mathcal{C}_i = \langle \mathcal{D}\mathrm{u}(\theta_0) - f, \mathcal{D}\varphi_i \rangle$

Dynamics of simplified GD

- if $\epsilon_k \sim \mathcal{O}(\epsilon)$, then GD can be approximated by simpGD:

$$\theta_{k+1} = (I - \eta\mathcal{A})\theta_k + \eta(\mathcal{A}\theta_0 + \mathcal{C})$$

- Small error terms correspond to the NTK regime for $\mathrm{u}_\theta, \mathcal{D}\mathrm{u}_\theta$:

$$TKf_\theta(x,y) = \nabla_\theta f_\theta(x)^\top \nabla_\theta f(y).$$

- For simpGD, easy to show that

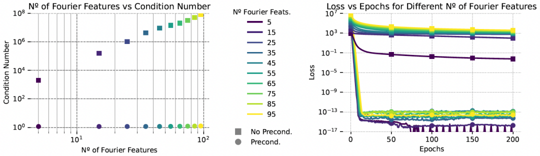
$$\|\theta_k - \theta^*\|_2 \leq \left(1 - \frac{c}{\kappa(\mathcal{A})}\right)^k \|\theta_0 - \theta^*\|_2, \quad N(\delta) \sim \mathcal{O}(\kappa(\mathcal{A})\log(1/\delta))$$

- Key role played by Condition Number: $\kappa(\mathcal{A}) = \frac{\lambda_{\max}(\mathcal{A})}{\lambda_{\min}(\mathcal{A})}$

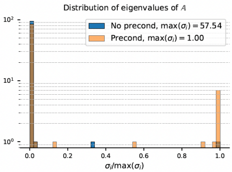
More on Simp GD

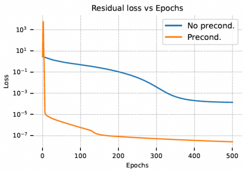
- Introduce $\mathcal{A} = \mathcal{D}^*\mathcal{D}$, the Hermitian-Square of $\mathcal{D}$.
- Under suitable assumptions, $\kappa(\mathcal{A}) = \kappa\left(\mathcal{A} \odot TT^*\right)$,
- $T : v \mapsto \sum_k v_k \varphi_k$ connects the vector and function spaces.
- Ex: if $\mathcal{D} = -\Delta$, then $\mathcal{A} = \Delta^2$
- in general $\kappa(\mathcal{A})$ can be very high.
- Key difference in Supervised Learning and Physics-Informed learning
- Need to precondition $\mathcal{D}^*\mathcal{D}$.
- Most techniques to accelerate PINNs training can be viewed as Preconditioning

1-D Possion: $-u'' = -k^2 \sin(kx)$

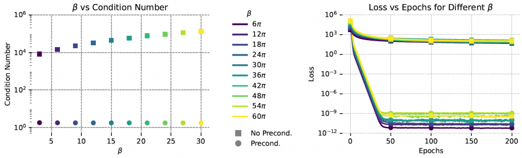


1-D Possion: $-u'' = -k^2 \sin(kx)$

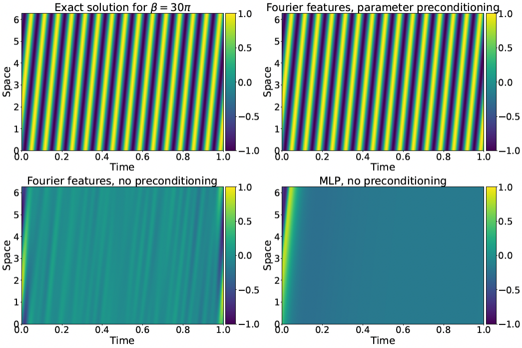




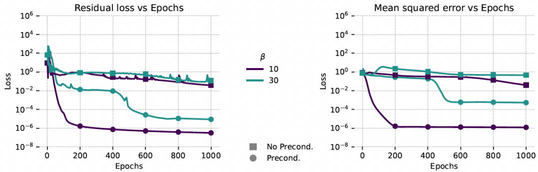
1-D Advection: $u_t + \beta u_x = 0$



1-D Advection: $u_t + \beta u_x = 0$



1-D Advection: $u_t + \beta u_x = 0$



1-D Advection: $u_t + \beta u_x = 0$