Error analysis for physics informed neural networks

Tim De Ryck

Seminar for Applied Mathematics, ETH Zürich

Bernouillis Tafelrunde - Universität Basel
4 October 2021

Outline

- Neural networks & physics informed neural networks (PINNs)
- Neural network approximation in Sobolev norms
- Error analysis case studies
  - PINNs for the Navier-Stokes equation
  - PINNs for linear Kolmogorov PDEs
- Physics informed operator learning

Joint work with

- S. Lanthaler (ETH Zürich),
- S. Mishra (ETH Zürich),
- A.D. Jagtap (Brown University).

Neural networks

A **feedforward (artificial) neural network** of depth L is a map of the form

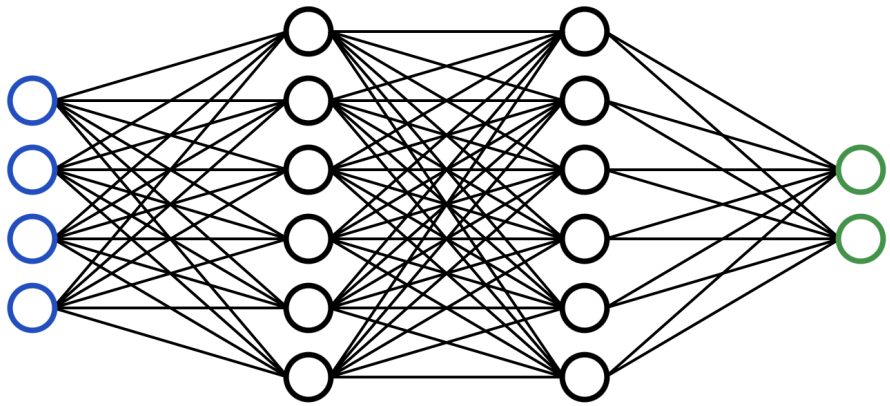
$$\Phi : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L} : x \mapsto (\mathcal{A}^L \circ \rho^{L-1} \circ \cdots \circ \rho^1 \circ \mathcal{A}^1)(x)$$

where

- $\mathcal{A}^\ell : \mathbb{R}^{n_{\ell-1}} \to \mathbb{R}^{n_\ell}$ are affine linear maps,

$$\mathcal{A}^\ell(x) = W^\ell x + b^\ell \text{ with weights } W^\ell \text{ and bias } b^\ell,$$

- ρ^ℓ are activation functions.



Neural networks

A **feedforward (artificial) neural network** of depth L is a map of the form

$$\Phi : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L} : x \mapsto (\mathcal{A}^L \circ \rho^{L-1} \circ \cdots \circ \rho^1 \circ \mathcal{A}^1)(x)$$

where

- $\mathcal{A}^\ell : \mathbb{R}^{n_{\ell-1}} \to \mathbb{R}^{n_\ell}$ are affine linear maps,
- ρ^ℓ are activation functions.

A feedforward NN is called a **tanh NN** when

- $\tanh : \mathbb{R} \to \mathbb{R} : x \mapsto \frac{e^x - e^{-x}}{e^x + e^{-x}}$,
- $\rho^\ell(x) = (\tanh(x_1), \ldots, \tanh(x_{n_\ell}))$ for $\ell = 1, 2, \ldots, L-1$,

Network has depth L, width $\max_\ell n_\ell$ and $\sum_\ell n_\ell$ neurons

Neural network approximation theory

Central question

Given function class $\mathcal{F}$ and $f \in \mathcal{F}$, what size should a neural network $\widehat{f}$ have such that $\left\| f - \widehat{f} \right\|_{\mathcal{F}} < \epsilon$?

Typical result (e.g. [Yarotsky, 2017]):

Theorem

For every $\epsilon > 0$ and every $f \in W^{n,\infty}([0,1]^d)$, there exists a tanh neural network $\widehat{f}$ with $O(\epsilon^{-d/n})$ neurons such that $\left\| f - \widehat{f} \right\|_{\infty} < \epsilon$.

Neural network approximation theory

NN structure allows for constructive proofs

Sum or composition of NN is a NN

For $\text{ReLU}(x) = \max\{x, 0\} = (x)_+$:

- $x = (x)_+ - (-x)_+$,
- $\max\{a, b\} = a + (b - a)_+$.

For smooth σ with $\sigma'(0) \neq 0$ and $|x| \leq B$:

$$\forall h > 0 : x = \frac{\sigma(xh) - \sigma(-xh)}{2h\sigma'(0)} + O(B^3 h^2)$$

Setting

PDE is characterized by:

- domain D,
- differential operator $\mathcal{D} : X \to Y$,
- boundary operator $\mathcal{B} : X \to Z$.

We assume that there exists $u \in X$ with

- $\forall x \in D : (\mathcal{D}u)(x) = 0$,
- $\forall y \in \partial D : (\mathcal{B}u)(y) = 0$.

Goal: approximate u with neural network $\widehat{u}$ s.t. $\|u - \widehat{u}\|_X$ is small

Finding the neural network $\widehat{u}$

Goal: find NN $\widehat{u}$ such that $\|u - \widehat{u}\|_X$ is small

Supervised learning consists of

- select (grid) points y_i and use PDE solver to approximate $u(y_i)$
- training set $\mathcal{S} = \{(y_1, u(y_1)), \dots (y_N, u(y_N))\} \subset D \times u(D)$,
- approach: minimize $\frac{1}{N} \sum_{n=1}^{N} \|u(y_n) - \widehat{u}(y_n)\|^2$,
- problem: generating training data can be expensive,

Finding the neural network $\widehat{u}$

Goal: find NN $\widehat{u}$ such that $\|u - \widehat{u}\|_X$ is small

Problem: generating training data can be expensive

Recall that $\mathcal{D}u = \mathcal{B}u = 0 \Rightarrow$ idea: minimize residuals $\|\mathcal{D}\widehat{u}\|_Y + \lambda\|\mathcal{B}\widehat{u}\|_Z$

Physics informed (unsupervised) learning consists of

- e.g. [Lagaris et al., 2000; Raissi et al., 2019],
- select (grid) points $x_i \in D$ and $y_i \in \partial D$
- training sets $\mathcal{S}_i = \{x_1, \dots x_N\} \subset D$ and $\mathcal{S}_b = \{y_1, \dots y_M\} \subset \partial D$ are free $\rightarrow$ no data generation necessary,
- approach: minimize $\frac{1}{N}\sum_{n=1}^{N} \|\mathcal{D}\widehat{u}(x_n)\| + \frac{\lambda}{M}\sum_{m=1}^{M} \|\mathcal{B}\widehat{u}(y_m)\|$.

Physics informed learning

Goal: find NN $\widehat{u}$ such that $\|u - \widehat{u}\|_X$ is small

$\mathcal{D}u = \mathcal{B}u = 0 \Rightarrow$ idea: minimize PINN residual $\|\mathcal{D}\widehat{u}\|_Y + \lambda\|\mathcal{B}\widehat{u}\|_Z$

Physics informed (unsupervised) learning consists of

- training sets $\mathcal{S}_i = \{x_1, \dots x_N\} \subset D$ and $\mathcal{S}_b = \{y_1, \dots y_M\} \subset D$,
- approach: minimize $\frac{1}{N}\sum_{n=1}^{N} \|\mathcal{D}\widehat{u}(x_n)\| + \frac{\lambda}{M}\sum_{m=1}^{M} \|\mathcal{B}\widehat{u}(y_m)\|$.

Questions

1. **Existence**: Is there $\widehat{u}$ such that $\|\mathcal{D}\widehat{u}\|_Y + \lambda\|\mathcal{B}\widehat{u}\|_Z$ is small? If yes, what is the size of $\widehat{u}$?

Physics informed learning

Goal: find NN $\widehat{u}$ such that $\|u - \widehat{u}\|_X$ is small

$\mathcal{D}u = \mathcal{B}u = 0 \Rightarrow$ idea: minimize PINN residual $\|\mathcal{D}\widehat{u}\|_Y + \lambda\|\mathcal{B}\widehat{u}\|_Z$

Physics informed (unsupervised) learning consists of

- training sets $\mathcal{S}_i = \{x_1, \dots x_N\} \subset D$ and $\mathcal{S}_b = \{y_1, \dots y_M\} \subset D$,
- approach: minimize $\frac{1}{N}\sum_{n=1}^{N} \|\mathcal{D}\widehat{u}(x_n)\| + \frac{\lambda}{M}\sum_{m=1}^{M} \|\mathcal{B}\widehat{u}(y_m)\|$.

Questions

1. **Existence**: Is there $\widehat{u}$ such that $\|\mathcal{D}\widehat{u}\|_Y + \lambda\|\mathcal{B}\widehat{u}\|_Z$ is small? If yes, what is the size of $\widehat{u}$?
2. **Stability**: If $\|\mathcal{D}\widehat{u}\|_Y + \lambda\|\mathcal{B}\widehat{u}\|_Z$ is small, will $\|u - \widehat{u}\|_X$ be small as well?

Physics informed learning

Goal: find NN $\widehat{u}$ such that $\|u - \widehat{u}\|_X$ is small

$\mathcal{D}u = \mathcal{B}u = 0 \Rightarrow$ idea: minimize PINN residual $\|\mathcal{D}\widehat{u}\|_Y + \lambda\|\mathcal{B}\widehat{u}\|_Z$

Physics informed (unsupervised) learning consists of

- training sets $\mathcal{S}_i = \{x_1, \dots x_N\} \subset D$ and $\mathcal{S}_b = \{y_1, \dots y_M\} \subset D$,
- approach: minimize $\frac{1}{N}\sum_{n=1}^{N} \|\mathcal{D}\widehat{u}(x_n)\| + \frac{\lambda}{M}\sum_{m=1}^{M} \|\mathcal{B}\widehat{u}(y_m)\|$.

Questions

1. **Existence**: Is there $\widehat{u}$ such that $\|\mathcal{D}\widehat{u}\|_Y + \lambda\|\mathcal{B}\widehat{u}\|_Z$ is small? If yes, what is the size of $\widehat{u}$?
2. **Stability**: If $\|\mathcal{D}\widehat{u}\|_Y + \lambda\|\mathcal{B}\widehat{u}\|_Z$ is small, will $\|u - \widehat{u}\|_X$ be small as well?
3. **Generalization**: Does small training error imply small generalization error?

Neural network approximation in Sobolev norms

Approach 1: approximates well in Sobolev norm $\Rightarrow$ small PINN residual

Theorem [DR, Lanthaler, and Mishra, 2021b]

For every $N \in \mathbb{N}$ and every $f \in W^{s,\infty}([0,1]^d)$, there exists a tanh neural network $\widehat{f}$ with 2 hidden layers of width N^d such that for every $0 \leq k < s$ it holds that,

$$\left\| f - \widehat{f} \right\|_{W^{k,\infty}} \leq C(\ln(cN))^k N^{-s+k},$$

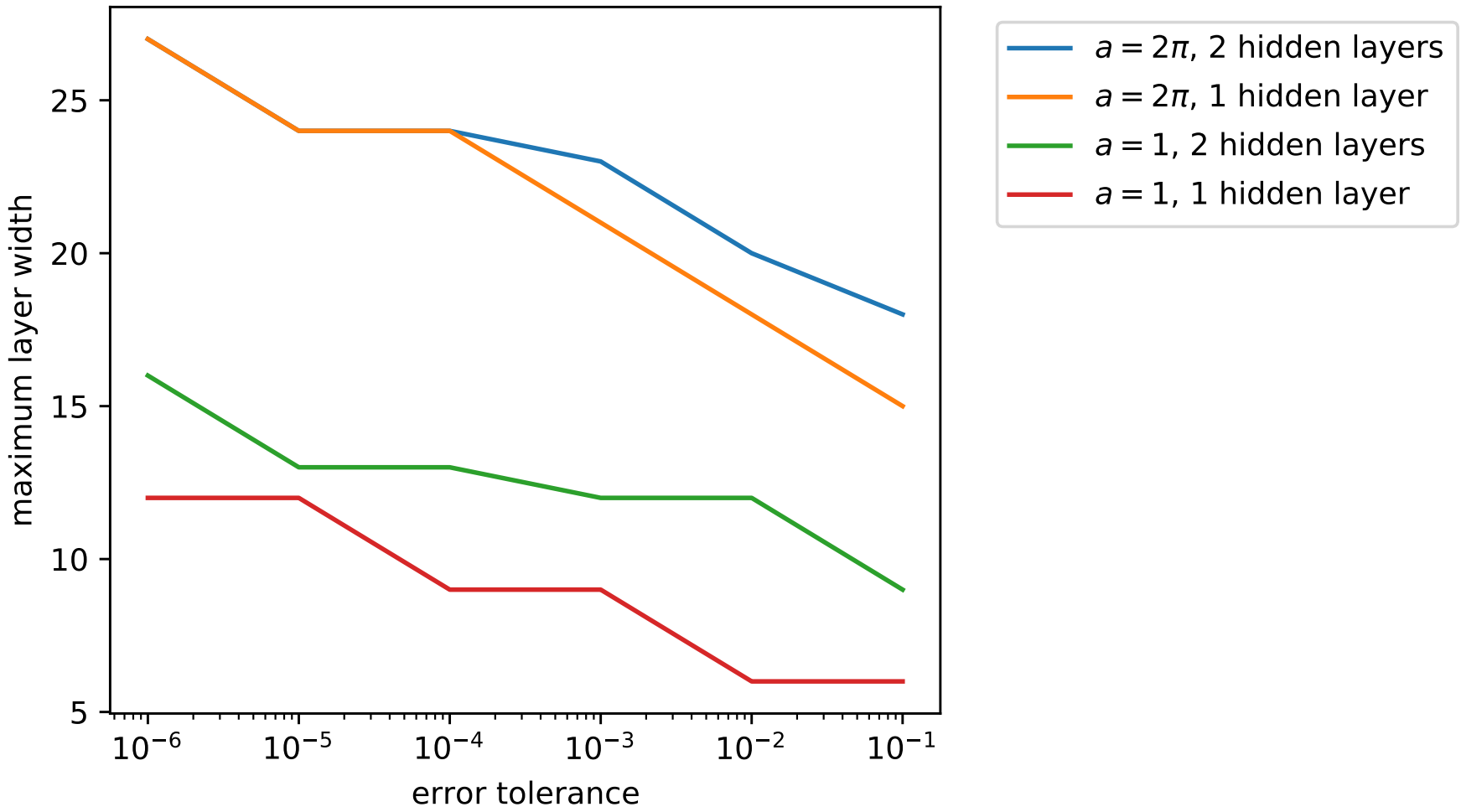
where $c, C > 0$ are independent of N and explicitly known.

Summary of our work:

- No unknown parameters
- Faster convergence for analytic functions
- More general norm
- Fixed depth

Neural network approximation in Sobolev norms

Layer width of network to approximate
$f_a : [0,1] \to [-1,1] : x \mapsto \sin(ax), \quad a > 0$?



PINNs for Navier-Stokes equations

Bounds in $W^{k,\infty}$-norm $\Rightarrow$ existence of NNs with small PINN error

Case study: Navier-Stokes equations in d space dimensions (periodic BC)

$$\begin{cases} u_t + u \cdot \nabla u + \nabla p - \nu \Delta u = 0 & \text{in } D \times [0,T], \\ \operatorname{div}(u) = 0 & \text{in } D \times [0,T], \\ u(t=0) - u_0 = 0 & \text{in } D. \end{cases} \quad \Rightarrow \quad \text{residuals} \begin{cases} \mathcal{R}_{\text{PDE}} \\ \mathcal{R}_{\text{div}} \\ \mathcal{R}_{\text{t}} \end{cases}$$

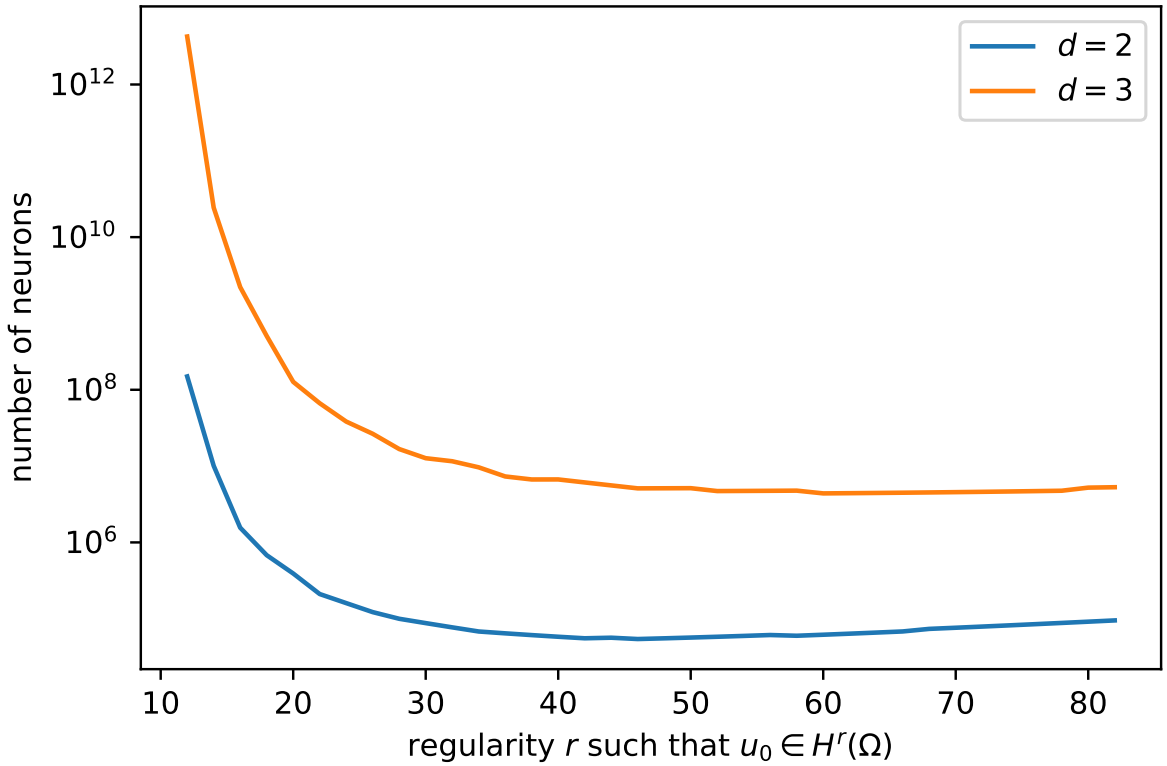
Theorem [DR, Jagtap, and Mishra, 2021a]

Let $u_0 \in H^r$ with $r > \frac{d}{2} + 2k$ and $\operatorname{div}(u_0) = 0$. For every $N \in \mathbb{N}$ there exists a tanh neural network $\widehat{u}$ with 2 hidden layers of width N^{d+1} s.t.

$$\|\mathcal{R}_{\text{PDE}}[\widehat{u}]\|_{L^2} + \|\mathcal{R}_{\text{div}}[\widehat{u}]\|_{L^2} + \|\mathcal{R}_{\text{t}}[\widehat{u}]\|_{L^2} \leq C(\ln(cN))^2 N^{-k+2}.$$

PINNs for Navier-Stokes equations

Needed NN size such that PINN error is $< 1\%$ for $u_0 \in H^r(\mathbb{T}^d)$:



$\Rightarrow$ reasonable NN sizes

PINNs for Navier-Stokes equations

So far: existence of NN with small PINN residuals ❶ ✓

Does small PINN loss imply small L^2-error?

Theorem

$$\|u - \widehat{u}\|_{L^2}^2 \leq C(\|\mathcal{R}_{\text{div}}\|_{L^2} + \|\mathcal{R}_{\text{PDE}}\|_{L^2}^2 + \|\mathcal{R}_{\text{s}}\|_{L^2} + \|\mathcal{R}_{\text{t}}\|_{L^2}^2)$$

Yes! Stability ❷ ✓

Does small training error $\mathcal{E}_T(\mathcal{S})$ imply small generalization error?

Theorem

$$\|u - \widehat{u}\|_{L^2}^2 \leq C(\mathcal{E}_T(\mathcal{S}) + N_t^{-\frac{2}{d}} + N_{\text{int}}^{-\frac{1}{d+1}} + N_s^{-\frac{1}{d}})$$

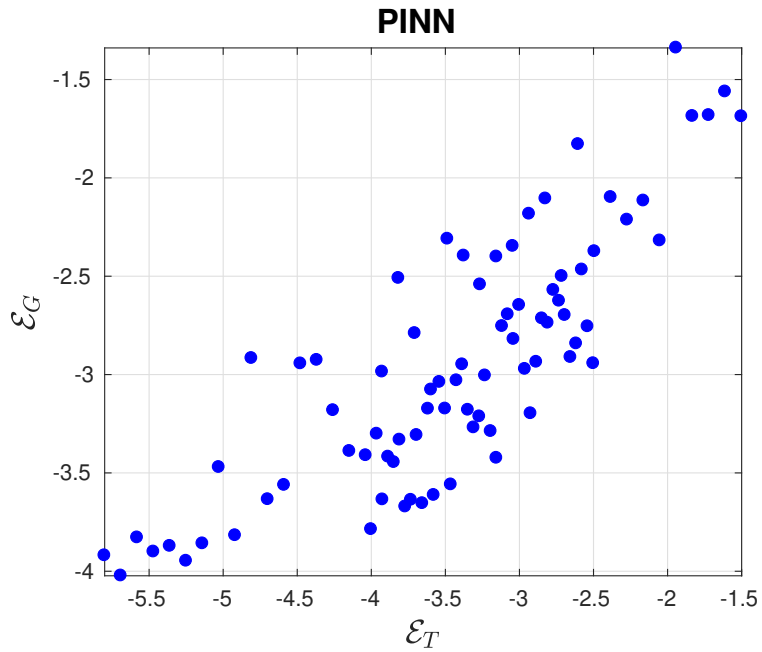
Yes! Generalization ❸ ✓

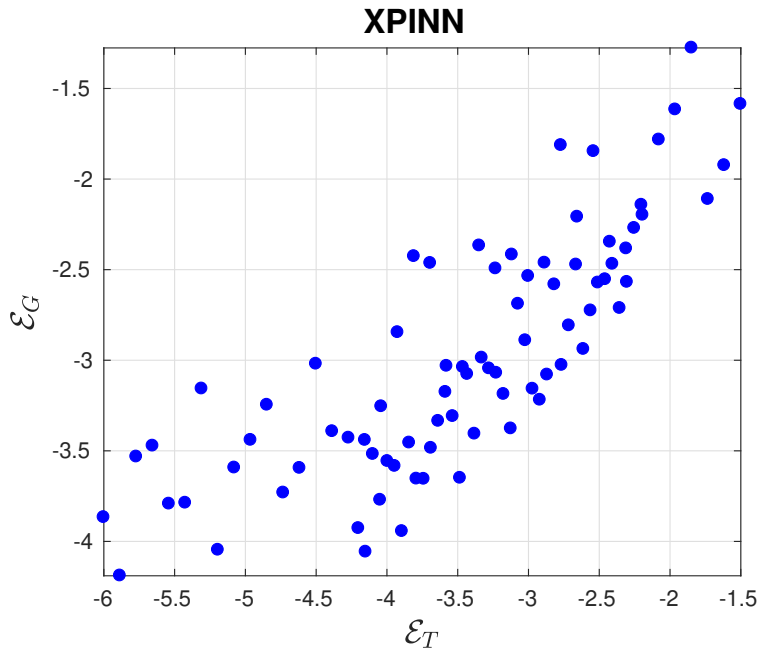
PINNs for Navier-Stokes

Does small training error $\mathcal{E}_T(\mathcal{S})$ imply small generalization error $\mathcal{E}_G(\mathcal{S})$?

Theorem

$$\mathcal{E}_G^2 = \|u - \widehat{u}\|_{L^2}^2 \leq C(\mathcal{E}_T(\mathcal{S}) + N_t^{-\frac{2}{d}} + N_{\text{int}}^{-\frac{1}{d+1}} + N_s^{-\frac{1}{d}})$$





PINNs for Kolmogorov PDEs

So far: Navier-Stokes equations are low-dimensional

Previous estimates suffer from curse of dimensionality (CoD):

- e.g. accuracy $\times 10 \Rightarrow$ size $\times 10^{d/s}$.

Many PDEs are high-dimensional:

- e.g. option pricing: Black-Scholes, Heston ...,
- experiments show that PINNs can overcome CoD [Mishra, Molinaro, Tanios, 2021].

Can we rigourously prove that PINNs overcome the CoD?

PINNs for Kolmogorov PDEs

Focus on linear Kolmogorov PDEs with solution u,

$$\begin{cases} u_t(t,x) = \mathcal{L}[u](t,x) = \frac{1}{2}\text{Trace}(\sigma(x)\sigma(x)^T H_x[u](t,x)) + \mu(x)^T \cdot \nabla_x[u](t,x), \\ u(0,x) = \varphi(x), \\ u(t,y) = \psi(t,y), \qquad \forall x \in D, y \in \partial D, t \in [0,T], \end{cases}$$

where $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ and $\mu : \mathbb{R}^d \to \mathbb{R}^d$ are affine functions.

PINNs for Kolmogorov PDEs

Focus on linear Kolmogorov PDEs with solution u,

$$\begin{cases} u_t(t,x) = \mathcal{L}[u](t,x) = \frac{1}{2}\text{Trace}(\sigma(x)\sigma(x)^T H_x[u](t,x)) + \mu(x)^T \cdot \nabla_x[u](t,x), \\ u(0,x) = \varphi(x), \\ u(t,y) = \psi(t,y), \qquad \forall x \in D, y \in \partial D, t \in [0,T], \end{cases}$$

where $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ and $\mu : \mathbb{R}^d \to \mathbb{R}^d$ are affine functions.

Examples

- Heat equation: $\mu = 0$ and $\sigma = \sqrt{\kappa} I_d$
- Black-Scholes equation: μ and σ linear

Goal: ❶ existence, ❷ stability, ❸ generalization

PINNs for Kolmogorov PDEs

Connection between Kolmogorov PDE and Itô diffusion SDE,

$$dX_t^x = \mu(X_t^x)dt + \sigma(X_t^x)dB_t, \quad X_0^x = x, \quad x \in D, t \in [0, T],$$

with generator $(\mathcal{F}\varphi)(X_t^x) = \sum_{i=1}^d \mu_i(X_t^x)(\partial_i\varphi)(X_t^x) + \frac{1}{2}\sum_{i,j,k=1}^d \sigma_{ik}(X_t^x)\sigma_{kj}(X_t^x)(\partial_{ij}^2\varphi)(X_t^x)$.

Dynkin's formula

If $\varphi \in C^2(\mathbb{R}^d)$ with bounded first partial derivatives, then it holds that

$$(\partial_t u)(x,t) = \mathcal{L}[u](x,t)$$

where u is defined as

$$u(x,t) = \varphi(x) + \mathbb{E}\left[\int_0^t (\mathcal{F}\varphi)(X_\tau^x)d\tau\right] \qquad \text{for } x \in D, t \in [0,T].$$

PINNs for Kolmogorov PDEs

Approach 2:

Theorem [DR and Mishra, 2021a]

If one can approximate φ without the CoD, then there exist constants $\alpha, \beta > 0$ such that for every $\varepsilon > 0$ and $d \in \mathbb{N}$, there exist a constant $\rho_d > 0$ and a tanh neural network $\Psi_{\varepsilon,d}$ with at most $O((d\rho_d)^\alpha \varepsilon^{-\beta})$ neurons such that

$$\|\partial_t \Psi_{\varepsilon,d} - \mathcal{L}[\Psi_{\varepsilon,d}]\|_{L^2(D_d \times [0,T])} + \|\Psi_{\varepsilon,d} - u_d\|_{L^2(\partial(D_d \times [0,T]))} \leq \varepsilon.$$

PINNs for Kolmogorov PDEs

Approach 2:

Theorem [DR and Mishra, 2021a]

If one can approximate φ without the CoD, then there exist constants $\alpha, \beta > 0$ such that for every $\varepsilon > 0$ and $d \in \mathbb{N}$, there exist a constant $\rho_d > 0$ and a tanh neural network $\Psi_{\varepsilon,d}$ with at most $O((d\rho_d)^\alpha \varepsilon^{-\beta})$ neurons such that

$$\|\partial_t \Psi_{\varepsilon,d} - \mathcal{L}[\Psi_{\varepsilon,d}]\|_{L^2(D_d \times [0,T])} + \|\Psi_{\varepsilon,d} - u_d\|_{L^2(\partial(D_d \times [0,T]))} \leq \varepsilon.$$

Convergence rate is dimension-independent

CoD is fully overcome if $\rho_d = O(d^\gamma)$,

- e.g. when μ and σ are constant,
- ρ_d depends on time regularity of Itô diffusion corresponding to PDE.

PINNs for Kolmogorov PDEs

Sketch of the proof: use $u(x,t) = \varphi(x) + \mathbb{E}\left[\int_0^t (\mathcal{F}\varphi)\left(X_\tau^x\right) d\tau\right]$ and

- replace $\varphi \to \widehat{\varphi}$ and $\mathcal{F}\varphi \to \widehat{\mathcal{F}\varphi}$,
- replace $\mathbb{E}$ by average,
- replace $\int_0^t$ by trapezoidal rule,
- μ and σ are affine $\Rightarrow X_t^x = \sum_{i=1}^d \left(X_t^{e_i} - X_t^0\right) x_i + X_t^0$
  $\Rightarrow \mathcal{L}\left[(\mathcal{F}\varphi)\left(X_\tau^x\right)\right]$ is computable.

Conclusion: existence ❶ ✓

PINNs for Kolmogorov PDEs

So far: existence of NN with small PINN residuals ❶ ✓

Does small PINN residual imply small L^2-error?

Theorem [DR and Mishra, 2021a]

$$\|u - \widehat{u}\|_{L^2}^2 \leq C(\|\mathcal{R}_{\text{PDE}}\|_{L^2}^2 + \|\mathcal{R}_{\text{s}}\|_{L^2} + \|\mathcal{R}_{\text{t}}\|_{L^2}^2)$$

Yes! Stability ❷ ✓

PINNs for Kolmogorov PDEs

So far: existence of NN with small PINN residuals ❶ ✓

Does small PINN residual imply small L^2-error?

Theorem [DR and Mishra, 2021a]

$$\|u - \widehat{u}\|_{L^2}^2 \leq C(\|\mathcal{R}_{\text{PDE}}\|_{L^2}^2 + \|\mathcal{R}_{\text{s}}\|_{L^2} + \|\mathcal{R}_{\text{t}}\|_{L^2}^2)$$

Yes! Stability ❷ ✓

Does small training error $\mathcal{E}_T(\mathcal{S})$ imply small generalization error?

Theorem [DR and Mishra, 2021a]

With high probability: $\|u - \widehat{u}\|_{L^2}^2 \leq C(\mathcal{E}_T(\mathcal{S}) + N_t^{-\frac{1}{2}} + N_{\text{int}}^{-\frac{1}{2}} + N_s^{-\frac{1}{4}})$.

Yes! Generalization ❸ ✓

One step further ...

Deep operator learning: map between infinite-dimensional spaces

- PDEs: $u(0,x) \mapsto u(T,x)$
- Antiderivative operator: $f \mapsto \int f$
- DeepONet, Fourier Neural Operator (FNO), ...
- Error estimates e.g. [Lanthaler, Mishra, Karniadakis, 2021], [Kovachki, Lanthaler, Mishra, 2021]

Physics informed deep operator learning (March 2021)

- Training set: sample points and functions

Theorem [DR and Mishra, 2021b]

Error estimate for DeepONets/FNO + regularity assumptions
$\Rightarrow$ Error estimate for physics informed DeepONet/FNO

Summary

Physics informed neural networks

- Surrogate models for PDE solutions
- No training data needed

Theoretical investigation

- Existence
- Stability
- Generalization

Results valid for

- Low-dimensional PDEs
- High-dimensional Kolmogorov PDEs

More information

DR and Siddhartha Mishra. Error analysis for PINNs approximating Kolmogorov PDEs. *SAM report 2021-17*, 2021a.

DR and Siddhartha Mishra. Error estimates for physics informed operator learning. *In preparation*, 2021b.

DR, Ameya D. Jagtap, and Siddhartha Mishra. Error analysis for PINNs approximating the Navier-Stokes equations. *In preparation*, 2021a.

DR, Samuel Lanthaler, and Siddhartha Mishra. On the approximation of functions by tanh neural networks. *Neural Networks*, 2021b.

Course

'Deep Learning in Scientific Computing' at ETHZ (spring semester)
Audience: mathematics and CSE master students