A statistical tour of physics-informed machine learning

Focus on physics-informed neural networks (PINNs)

* for simplicity.

Setting:

$\Omega \subset \mathbb{R}^{d_{in}}$ ($d_{in} = d^*$) bounded Lipschitz domain

ex: bounded convex domains

ex: manifolds with $\mathcal{C}^1$ boundaries.

$C^K(\Omega, \mathbb{R}^{d_{out}})$ ($d_{out} = 1^*$) space of functions from Ω to $\mathbb{R}^{d_{out}}$ K-time continuously differentiable.

$$C^\infty(\Omega, \mathbb{R}^{d_{out}}) = \bigcap_{K \geq 0} C^K(\Omega, \mathbb{R}^{d_{out}})$$

Hölder norm $\quad \|f\|_{\mathcal{C}^K(\Omega)} = \max_{|\alpha| \leq K} \|\partial^\alpha f\|_{\infty, \Omega}$

Hybrid modelling

- Goal: to estimate an unknown regression function f^* such that $f^*: \Omega \to \mathbb{R}^{d_{out}}$ ($d_{out} = 1$) and

$$Y = f^*(X) + \varepsilon$$

random noise $\quad \mathbb{E}[\varepsilon | X] = 0$.

In traditional supervised learning, to do so we use training samples (X_i, Y_i) $i = 1..n$ independent copies of (X, Y) distributed as $\circledast$

- Additional information : physical modeling / prior knowledge

$$\mathscr{D}(f^*, .) \simeq 0 \quad \text{on } \Omega .$$

known differential operator

- Additional information : Initial / Boundary conditions on $E \subseteq \partial\Omega$

partial knowledge allowed.

ex: when $E = \partial\Omega$, Dirichlet boundary condition.

- To estimate f^*, we have 3 different data sets

① The "traditional" training samples (unknown distribution)

$$(X_1, Y_1) \dots (X_n, Y_n)$$

② Collocation points : a collection of iid points (uniformly distributed on Ω)

$$X_1^{(r)}, \dots X_{n_r}^{(r)}$$

Those points can be seen as a naive discretization of Ω.

③ Condition points : a collection of iid points $\sim \mu_E$ on E

$$X_1^{(e)}, \dots, X_{n_e}^{(e)}$$

(known distribution μ_E on E)

- Physics-informed empirical risk

can be an absolute value for simplicity

$$R_{n,n_r,n_e}(f) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\|f(X_i)-Y_i\|_2^2}_{\text{Data-fidelity term}} + \underbrace{\frac{\lambda_r}{n_r}\sum_{i=1}^{n_r}\|\mathcal{D}(f, X_i^{(r)})\|_2^2}_{\text{PDE term}}$$

$$+ \underbrace{\frac{\lambda_e}{n_e}\sum_{i=1}^{n_e}\|f(X_i^{(e)})-h(X_i^{(e)})\|_2^2}_{\text{Initial / boundary conditions}}$$

I may drop the initial / boundary conditions for the sake of presentation.

- When this physics-informed risk is minimized over a class of neural networks, we obtain the so-called physics-informed neural networks (PINNs)

to do so, consider the class of fully-connected feed forward NN with H hidden layers of width $(L_1, L_2, \dots, L_H) = (D, D, \dots D)$

an input layer of width $L_0 = d_{in} = d$ $(\Omega \subset \mathbb{R}^{d_{in} = d})$

an output layer of width $L_{H+1} = d_{out} = 1$ $(Y \in \mathbb{R}^{d_{out} = 1})$

activation function tanh

Overall,

$$f_\theta = \mathcal{A}_{out} \circ (\underbrace{\tanh}_{\text{applied component-wise}} \circ \mathcal{A}_H) \circ \ldots \circ (\tanh \circ \mathcal{A}_1)$$

with $\mathcal{A}_k : \mathbb{R}^{L_{k-1}} \to \mathbb{R}^{L_k}$ 　　weight matrix $W_k \in \mathbb{R}^{L_{k-1} \times L_k}$

$x \mapsto W_k x + b_k$ 　　bias $b_k \in \mathbb{R}^{L_k}$

The parameters of the neural network – to be learned – are encoded by

$$\theta = (W_1, b_1, \ldots, W_{H+1}, b_{H+1}) \in \Theta_{H,D}$$

$$\Theta_{H,D} = \mathbb{R}^{\sum_{\ell=1}^{H} (L_\ell+1) \times L_{\ell+1}}$$

Notation: $NN_H(D) \equiv$ neural networks of H hidden layers of width D. 　　H hidden layers of width D

Approximation results

Ideally, the NN set $NN_H(D)$ should be chosen large enough to approximate both the solution of the PDE and its derivatives.

Prop: [Density of NN in Hölder spaces, De Ryck et al (2021), Doumèche et al (2023)]

For $H \geq 2$, $\boxed{NN_H \text{ is dense in the space } (\mathcal{C}^\infty(\bar\Omega, \mathbb{R}^{d_{out}}), \|\cdot\|_{C^K(\Omega)})}$ 　for any K

i.e. for any function $u \in \mathcal{C}^\infty(\bar\Omega, \mathbb{R}^{d_{out}})$, $\exists (u_p)_{p\in\mathbb{N}} \in NN_H^{\mathbb{N}}$ such that $\lim \|u_p - u\|_{C^K(\Omega)} = 0$

Theorem [Yarotsky, 2017]

For every $\varepsilon > 0$, for every $f \in \mathcal{C}^k([0,1]^d)$, there exists a tanh neural network $\hat{f}$ with $O(\varepsilon^{-d/k})$ neurons such that $\|f - \hat{f}\|_\infty < \varepsilon$.

NN structure allows for constructive proofs

Key properties

- sum or composition of NN is a NN
- For smooth activation fct σ with $\sigma'(0) \neq 0$, and $|x| \leq B$

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

Approximation of identity with 2 neurons.

Theorem [De Ryck, Lanthaler, Mishra, 2021]

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

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

$\perp$ of N.

Remark 1 also holds for Sobolev norms.

Remark 2 How to translate this in terms of PDE residuals

Example of the heat equation $\boxed{\mathscr{D}(f,x) = \partial_t f - \partial^2_{xx} f}$

when $f^* \in C^s([0,T] \times [0,1])$

① $\|\partial_t \hat{f} - \partial^2_{xx}\hat{f}\|_{L^2(\Omega)} \leq \sqrt{|\Omega|}\ \|\partial_t \hat{f} - \partial^2_{xx}\hat{f}\|_{\infty / C^0}$

② $$\|\partial_t \hat{f} - \partial^2_{xx}\hat{f}\|_{C^0} = \|\partial_t \hat{f} \underbrace{- \partial_t f^* + \partial^2_{xx} f^*}_{=0} - \partial^2_{xx}\hat{f}\|_{C^0}$$

$$\leq \|\partial_t \hat{f} - \partial_t f^*\|_{C^0} + \|\partial^2_{xx}\hat{f} - \partial^2_{xx} f^*\|_{C^0}$$

$$\leq \|\hat{f} - f^*\|_{C^1} + \|\hat{f} - f^*\|_{C^2}$$

$$\leq 2\|\hat{f} - f^*\|_{C^2}$$

Therefore, there is a NN with 2 hidden layers and N^2 neurons s.t.

$$\|\partial_t \hat{f} - \partial^2_{xx}\hat{f}\|_{L^2} \leq 2\sqrt{|\Omega|}\ \|\hat{f} - f^*\|_{C^2} \leq C\left(\ln(cN)\right)^2 N^{-s+2}.$$

Remark 3: in practice, the PINNs are of size.

$NN_4(50)$ Krishnapriyan et al. "Characterizing failure modes in PINNs"

$NN_5(100)$ Xu et al. "How NN extrapolate: from FFNN to GNN".

$NN_{10}(100)$ Arzani et al. "Blood flow".

Back to learning

- In traditional supervised learning, we have the statistical model

$$Y = f^*(X) + \text{noise} \qquad (\blacktriangle)$$

such that $\mathbb{E}[\text{noise} \mid X] = 0$.

target function

- Introduce the quadratic risk $\quad \text{Risk}(f) = \mathbb{E}\left[(Y - f(X))^2\right]$

- This risk is minimal for $f : x \mapsto \mathbb{E}[Y \mid X = x]$.

$$f = f^*. \qquad \text{(Bayes predictor)}$$

- To estimate f^*, we use n training samples (X_i, Y_i) $i = 1 \ldots n$ iid as (X, Y) distributed as $(\blacktriangle)$, and construct an estimator $\hat{f}_n$.

- The first statistical requirement is to hope for risk consistency

$$\lim_{n \to \infty} \text{Risk}(\hat{f}_n) = \text{Risk}(f^*)$$

- How to adapt this notion in the case of PINNs?

Overfitting

- Study of limitations of PINNs
- To this end, we introduce a particular risk function.

$$\mathcal{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n}\|f(X_i)-Y_i\|_2^2 + \frac{\lambda_r}{|\Omega|}\int_\Omega \|\mathcal{D}(f,x)\|^2 dx + \lambda_e \mathbb{E}\left[\|f(X^{(e)})-h(X^{(e)})\|_2^2\right]$$

expectation w.r.t. to the uniform measure on Ω

expectation w.r.t. μ_E.

Regime where

$$\begin{cases} n \text{ is fixed} & \text{(costly physical measurements)} \\ n_e, n_r \longrightarrow +\infty & (X_j^{(e)}, X_j^{(r)} \text{ can be freely sampled up to computational resources}) \end{cases}$$

$$R_{n,n_e,n_r}(f) = \frac{1}{n}\sum_{i=1}^{n}\|f(X_i)-Y_i\|_2^2 + \frac{\lambda_r}{n_r}\sum_{i=1}^{n_r}\|\mathcal{D}(f,X_i^{(r)})\|_2^2 + \frac{\lambda_e}{n_e}\sum_{i=1}^{n_e}\|f(X_i^{(e)})-h(X_i^{(e)})\|_2^2$$

- Consider $\hat{\theta}(p, n_e, n_r, D)$ a minimizing sequence of R_{n,n_e,n_r}

$$\lim_{p\to+\infty} R_{n,n_e,n_r}\left(f_{\hat{\theta}(p,n_e,n_r,D)}\right) = \inf_{\theta\in\Theta_{H,D}} R_{n,n_e,n_r}(f_\theta)$$

- Focus on ERM-type estimators (no implicit bias here)

Definition: Physical risk consistency

$$\boxed{\lim_{n_e,n_r\to+\infty}\ \lim_{p\to+\infty} \mathcal{R}_n\left(f_{\hat{\theta}(p,n_e,n_r,D)}\right) = \inf_{f\in NN_H(D)} \mathcal{R}_n(f)}$$

We will show that PINNs can dramatically fail to be risk-consistent.

- Hybrid modelling with friction

$$\Omega =]0,T[$$

$$\forall f \in \mathcal{C}^2(\bar{\Omega}, \mathbb{R}) \qquad \mathscr{D}(u,x) = m u''(x) + \eta u'(x)$$

Dynamics of an object of mass m subjected to a fluid force of friction coefficient $\eta > 0$.

In this case, the empirical risk reads as

$$R_{n,n_r}(f_\theta) = \frac{1}{n}\sum_{i=1}^{n} |f_\theta(X_i) - Y_i|^2 + \frac{\lambda_r}{n_r}\sum_{\ell=1}^{n_r}\left(m f_\theta''(X_\ell^{(r)}) + \eta f_\theta'(X_\ell^{(r)})\right)^2$$

Prop: provided that $\exists\, Y_i \neq Y_j$, whenever $D \geq n-1$, for any n_r, for all $(X_\ell^{(r)})_{\ell=1..n_r}$, there exists a minimizing sequence $(f_{\hat{\theta}(p,n,n_r,D)})$ such that

$$\begin{cases} \lim\limits_{p\to+\infty} R_{n,n_r}(f_{\hat{\theta}(p,n,n_r,D)}) = 0 & \text{(interpolation)} \\ \lim\limits_{p\to+\infty} \mathscr{R}_n(f_{\hat{\theta}(p,n,n_r,D)}) = +\infty. & \text{(PDE term exploding)} \end{cases}$$

The PINN estimator is not (physical) risk consistent.

This phenomenon is explained by the existence of piecewise constant functions interpolating the observations $X_1, \ldots, X_n$ whose derivatives are null at the points $X_1^{(r)}, \dots, X_{n_r}^{(r)}$ but diverge between these points.

proof: Consider the following NN

$$f_{\hat{\theta}(p,n,n_r,D)}(x) = Y_{(1)} + \sum_{i=1}^{n} \frac{Y_{(i+1)} - Y_{(i)}}{2} \left[\tanh_p^{\circ H}\left(x - X_{(i)} - \frac{\delta(n,n_r)}{2}\right) + 1 \right]$$

where the observations $(X_1, Y_1) \dots (X_n, Y_n)$ are reordered into $(X_{(1)}, Y_{(1)}) \dots (X_{(n)}, Y_{(n)})$ such that $X_{(1)} \leq X_{(2)} \leq X_{(3)} \leq \dots \leq X_{(n)}$

$\delta(n, n_r)$ is the minimal distance between 2 distinct points in the input observations (X_i, Y_i) and the collocation points $(X_\ell^{(r)})_\ell$

$\tanh_p(\cdot) = \tanh(p\,\cdot)$

$\tanh_p^{\circ H}$ composition of $\tanh_p$ H times

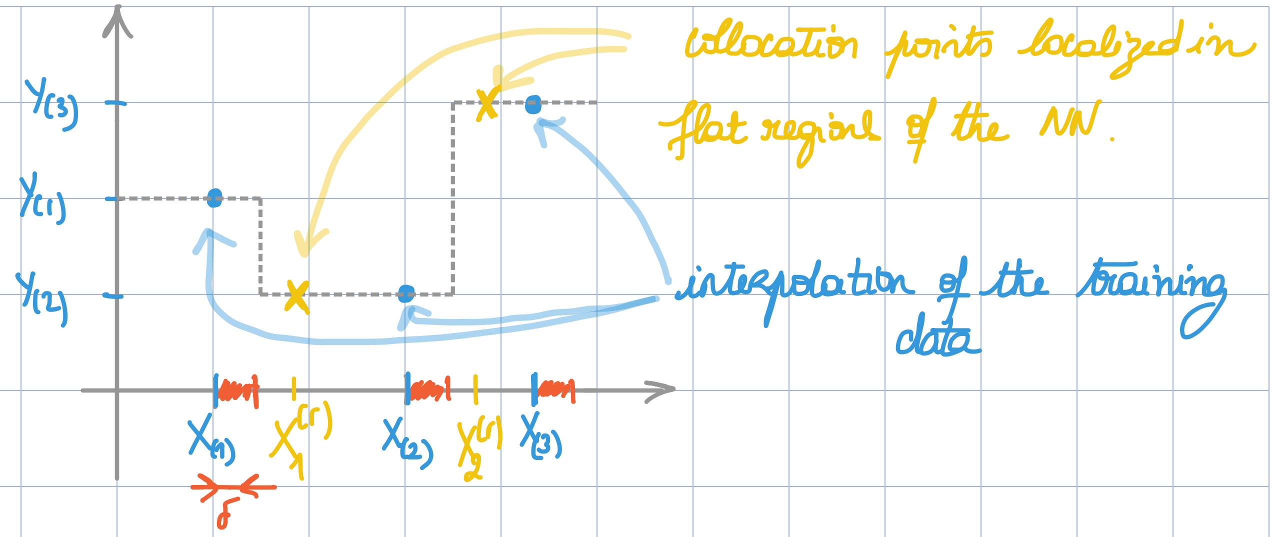
This NN has the following properties

(1) Interpolation of the training points

$$\boxed{\lim_{p \to +\infty} f_{\hat{\theta}(p,n,n_r,D)}(X_i) = Y_i}$$

Indeed for $\delta > 0$ $\lim \| \tanh_p^{\circ H} - \operatorname{sgn} \|_{C^K(\mathbb{R} \setminus [-\delta;\delta])} = 0$

Therefore $\frac{1}{2}\left[\tanh_p^{\circ H}\left(x - X_{(i)} - \frac{\delta(n,n_r)}{2}\right) + 1 \right] \xrightarrow[p \to +\infty]{}$ Heaviside of x centered at $\left(X_{(i)} + \frac{\delta}{2}\right)$



② All derivatives higher than order 1 vanish almost everywhere.

$$\mathcal{D}(f_{\hat\theta_p}, x_\ell^{(r)}) = \Big(m \underbrace{f''_{\hat\theta_p}(x_\ell^{(r)})}_{=0} + \eta \underbrace{f'_{\hat\theta_p}(x_\ell^{(r)})}_{=0}\Big)^2 = 0 \qquad \forall x_\ell^{(r)}.$$

③ But explode locally.

For any $0<\varepsilon<\delta(n,n_r)/4$,

$$R_n(f_{\hat\theta(p,n_r,D)}) \geq \frac{1}{T}\int_{[0,T]} \mathcal{D}(f_{\hat\theta(p,n_r,D)}, x)^2\,dx$$

$$\geq \frac{1}{T}\sum_{i=1}^{n} \int_{X_{(i)}+\delta/2-\varepsilon}^{X_{(i)}+\delta/2+\varepsilon} \mathcal{D}(f_{\hat\theta(p,n_r,D)}, x)^2\,dx.$$

put the square outside the integral + fundamental thm of integration.

$$\geq \frac{1}{T}\sum_{i=1}^{n}\frac{1}{2\varepsilon}\Big[m\underbrace{\big(f'_{\hat\theta}(X_{(i)}+\delta/2+\varepsilon) - f'_{\hat\theta}(X_{(i)}+\delta/2-\varepsilon)\big)}_{\xrightarrow{p\to+\infty} 0} + \eta\underbrace{\big(f_{\hat\theta}(X_{(i)}+\delta/2+\varepsilon) - f_{\hat\theta}(..)\big)}_{\xrightarrow{p\to+\infty} Y_{(i+1)}-Y_{(i)}}\Big]^2$$

$$\geq \eta^2 \cdot \frac{1}{2\varepsilon T}\sum_{i=1}^{n}(Y_{(i+1)}-Y_{(i)})^2.$$

Finally, $\lim\limits_{p\to+\infty} R_n(f_{\hat\theta(p,n_r,D)}) = +\infty$

Remarks:

① This pathological sequence of NN is such that $\boxed{\|\theta\|_2 \to +\infty}$.

② The proof does not depend on the geometry of the collocation points $(X_\ell^{(r)})_\ell$ or the condition points $(X_\ell^{(e)})_\ell$ (holds with grids or quasi Monte Carlo methods)

③ Holds for any PDE with derivatives ≥ 1 involved.

- Similar behaviour in the case of PDE solver

Consider
$$\Omega =]-1;1[\times]0,T[$$
$$\mathcal{D}(f,x) = \partial_t f - \partial^2_{xx} f \quad \text{Heat equation}$$
Boundary conditions $f(-1,t) = f(1,t) = 0$

Initial condition
$$f(x,0) = \tanh^{oH}(x+0.5) - \tanh^{oH}(x-0.5) + \tanh^{oH}(0.5) - \tanh^{oH}(1.5)$$
Bell function

Empirical risk function
$$R_{n_r,n_e}(f) = \frac{\lambda_e}{n_e} \sum_{j=1}^{n_e} |f(X_j^{(e)}) - h(X_j^{(e)})|^2 + \frac{1}{n_r} \sum_{\ell=1}^{n} \mathcal{D}(f, X_\ell^{(r)})^2$$

Physics ~~theoretical~~ risk function
$$\mathcal{R}(f) = \lambda_e \mathbb{E}\left[|f(X^{(e)}) - h(X^{(e)})|^2 \right] + \frac{1}{|\Omega|} \int_\Omega \mathcal{D}(f,x)^2 dx$$

Prop: When $D \geq 4$, $\forall (n_e, n_r)$ $\forall (X_1^{(r)}, \dots, X_{n_r}^{(r)})$ $\forall (X_1^{(e)}, \dots X_{n_e}^{(e)})$

$\exists$ a minimizing sequence $\left(f_{\hat{\theta}(p, n_e, n_r, D)}\right)_p$ s.t

$$\lim_{p \to +\infty} R_{n_e, n_r}\left(f_{\hat{\theta}(p, n_e, n_r, D)}\right) = 0$$

$$\lim_{p \to +\infty} \mathcal{R}\left(f_{\hat{\theta}(p, n_e, n_r, D)}\right) = +\infty$$

This PINN estimator is not physics risk consistent.

↳ Similarly, this estimator corresponds to a function that equals zero on Ω (and thus satisfies the linear PDE) while satisfying the initial condition on $\partial\Omega$

↳ Again this corresponds to a limit of neural networks f_θ such that $\|\theta\|_2 \longrightarrow +\infty$.

Fighting overfitting

- Connection between L^2-norm of the weights and the NN regularity.

  Prop: The following holds: $\forall \theta \in \Theta_{H,D}$
  $$\|f_\theta\|_{C^K(\mathbb{R}^{d_{in}})} \leqslant C_{K,H}\,(D+1)^{HK+1}\left(1+\|\theta\|_2\right)^{HK}\|\theta\|_2 .$$

  Remark: this bound is tight in the sense that there exists a sequence $(\theta_p) \in \Theta_{H,D}$ such that

  (i) $\lim_{p\to+\infty} \|\theta_p\|_2 = +\infty$

  (ii) $\|f_{\theta_p}\|_{C^K(\mathbb{R}^{d_{in}})} \geqslant \bar{C}_{KH}\,\|\theta_p\|_2^{HK+1}$.

- Idea: use a ridge-type / tikhonov-type regularization to avoid pathological minimizing sequences.

- Def: Ridge risk function

  hyperparameter
  $$R^{(ridge)}_{n,n_r,n_e}(f_\theta) = R_{n,n_r,n_e}(f_\theta) + \lambda_{(ridge)}\,\|\theta\|_2^2$$

- Denote $\left(\hat{\theta}^{(ridge)}_{(p,n_e,n_r,D)}\right)_p$ a minimizing sequence of the ridge risk function
  $$\lim_{p\to+\infty} R^{(ridge)}_{n,n_e,n_r}\left(f_{\hat{\theta}^{(ridge)}_{(p,n_e,n_r,D)}}\right) = \inf_{NN_H(D)} R^{(ridge)}_{n,n_e,n_r}.$$
  $\theta \in \Theta_{H,D}$.

• Note that ridge regularization is implementable in most deep learning libraries (eg. keras, pytorch, ...), implemented via "weight decay".

• Class of polynomial operators

bounded Lipschitz domain of $\mathbb{R}^3$

ex: Navier_Stokes equation on $\Omega = \Omega_1 \times]0,T[$, $f = (f_x, f_y, f_3, p)$

(3rd eqt) $$\mathcal{D}(f,x) = \partial_t f_3 - f_3 \partial_3 f_3 - \eta \partial^2_{33} f_3 + \rho^{-1} \partial_3 p + g(x).$$

It can be seen as a polynomial P of $f_3, \partial_3 f_3, \partial_t f_3, \partial^2_{33} f_3, \partial_3 p$

where P is given by $P(z_1, z_2, z_3, z_4, z_5) = z_3 - z_1 z_2 - \eta z_4 + \rho^{-1} z_5 + g$.

$\deg P = 2$ but $\deg \mathcal{D} = 3$ (counting the degree of the polynomial combined with the order of differentiation).

Encompass linear + nonlinear PDEs.

Theorem: $(H \geq 2)$ Assume the condition fct h is Lipschitz and that $\mathcal{D}$ is a polynomial operator.

By choosing $\lambda_{(ridge)} = \dfrac{1}{(\min(n_e, n_r))^{K}}$ with $K = \dfrac{1}{12 + 4H(1 + 2H \deg \mathcal{D})}$

$$\lim_{n_e n_r \to +\infty} \lim_{p \to +\infty} R_n\left(f_{\hat{\theta}_p^{(ridge)}}\right) = \inf_{NN_H(D)} R_n$$

- The choice of the hyperparameter $\lambda_{(ridge)}$ which vanishes when $n_e, n_r \longrightarrow +\infty$ makes the ridge bias vanish.
- This choice depends on known quantities (NN depth, degree of the differential operator, nb of discretization points).
- Extension of this result to remove any approximation bias introduced by the class $NN_H(D)$:

$$\lim_{D\to+\infty} \lim_{n_e,n_r\to+\infty} \lim_{p\to+\infty} R_n\left(f_{\hat\theta^{(ridge)}_{(p,n_e,n_r,D)}}\right) = \inf_{\mathcal{C}^\infty(\bar\Omega,\mathbb{R}^{d_{out}})} R_n$$

One can make D depend on n_e, n_r as well

$\left(D = \min(n_e,n_r)^{\xi} \text{ with } \xi \text{ function of } H \text{ and } \deg \mathcal{D}\right).$

In such a case

$$\lim_{n_e,n_r\to+\infty} \lim_{p\to+\infty} R_n\left(f_{\hat\theta^{(ridge)}_{(p,n_e,n_r,D(n_e,n_r))}}\right) = \inf_{\mathcal{C}^\infty(\bar\Omega,\mathbb{R}^{d_{out}})} R_n$$

Proof: In the case of the 1D heat equation <u>only</u>

$\partial_t f - \partial^2_{xx} f = 0$ meaning that $\mathcal{D}(f,\cdot) = \partial_t f - \partial^2_{xx} f$

① Let $f_0 = 0 \in NN_H(D)$ the NN with parameters all 0.

Trivially $R^{(ridge)}_{n,n_e,n_r}(f_0) = R_{n,n_e,n_r}(f_0)$

Therefore,

$$R_{n,n_e,n_r}(f_\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(f_\theta(X_i)-Y_i\right)^2 + \frac{\lambda_r}{n_r}\sum_{i=1}^{n_r}\left[\mathcal{D}(f_\theta, X_i^{(r)})\right]^2$$

$$\leqslant \frac{1}{n}\sum_{i=1}^{n}|Y_i|^2 + \lambda_r \, 2 \quad =: UB$$

does not depend on n_r (n_e if any) and λ^{ridge}

2 "coefficients" in the polynomial operator of sup norm 1.

$$\mathcal{D}(f,\cdot) = \phi_1 \, \partial_t f - \phi_2 \, \partial^2_{xx} f$$

$$\|\phi_1\|_{\infty,\bar\Omega} = \|\phi_2\|_{\infty,\bar\Omega} = 1.$$

② Let $\hat\theta^{(ridge)}_{p,n_e,n_r,D}$ be any minimizing sequence of the empirical ridge risk : $\lim_{p\to+\infty} R^{(ridge)}_{n,n_e,n_r}\left(f_{\hat\theta^{(ridge)}_{p,n_e,n_r}}\right) = \inf_{\theta\in\Theta_{H,D}} R_{n,n_e,n_r}(f_\theta)$

Call $n_{r,e} = \min(n_e, n_r)$

Define
$$E_1^{large}(n_{r,e}) = \left\{\theta\in\Theta_{H,D},\ \|\theta\|_2 \geqslant n_{r,e}^{K}\right\}$$
$$E_2^{inter}(n_{r,e}) = \left\{\theta\in\Theta_{H,D},\ n_{r,e}^{K/4} \leqslant \|\theta\|_2 \leqslant n_{r,e}^{K}\right\}$$
$$E_3^{small}(n_{r,e}) = \left\{\theta\in\Theta_{H,D},\ \|\theta\|_2 < n_{r,e}^{K/4}\right\}$$

Remark that $\Theta_{H,D} = E_1^{large} \cup E_2^{inter} \cup E_3^{small}$.

Summary of what's next :

θ small $\equiv$ no problem.

in-between $\equiv$ be careful.

θ large $\equiv$ does not happen

Idea of the proof: show that almost surely, given any n_r and n_e, for p large enough, $\hat{\theta}^{(ridge)}_{(p,n_r,D)} \in E_2^{inter} \cup E_3^{small}$.
Moreover on $E_2^{inter} \cup E_3^{small}$, the empirical risk function $R^{(ridge)}_{n,n_r}$ is close to the theoretical risk $\mathcal{R}_n$.

③ ⓐ $\forall \theta \in E_1^{large}$ $\quad R^{(ridge)}_{n,n_r}(\theta) \geqslant \lambda_{ridge} \|\theta\|_2^2 \geqslant n_{r,e}^K$.

Once $n_{r,e} \geqslant (UB+1)^{1/K}$,

$$\begin{aligned}\inf_{\theta \in E_3^{small}} R^{(ridge)}_{n,n_r}(f_\theta) + 1 &\leqslant R^{(ridge)}_{n,n_r}(f_0) + 1 \\ &\leqslant UB + 1 \\ &\leqslant n_{r,e}^K \\ &\leqslant \inf_{\theta \in E_1^{large}} R^{(ridge)}_{n,n_r}(f_\theta).\end{aligned}$$

This shows that for n_r large enough, $\boxed{\hat{\theta}^{(ridge)}_{(p,n_r,D)} \notin E_1^{large}.}$

ⓑ We can show that

(*) $$\sup_{\theta \in E_2^{inter} \cup E_3^{small}} \left| \frac{1}{n_r} \sum_{\ell=1}^{n_r} \mathscr{D}(f_\theta, X_\ell^{(r)})^2 - \frac{1}{|\Omega|} \int_\Omega \mathscr{D}(f_\theta, x)^2 dx \right| \leqslant \log^2(n_r)\, n_r^{\beta - 1/2}$$

with $\beta = (d + H(1 + (d+H)\deg \mathscr{D}))$. $\quad$ (Admitted).

Uniform control of stochastic process (Hoeffding / Dudley / Mc Diarmid)

Thus almost surely for all n_r large enough and $\theta \in E_2(n_{r,e})$

$$R^{(ridge)}_{n,n_r}(f_\theta) \geq \mathcal{R}_n(f_\theta) + \underbrace{\lambda_{(ridge)} \|\theta\|_2^2 - \cancel{2}\,\log^2(n_{r,e})\, n_r^{\beta-1/2}}_{\geq 0\,?}$$

$\cancel{1}$

~~coming from the absolute value~~.

Yes! Because

for all $\theta \in E_2^{inter}$, $\quad \underset{\substack{\parallel \\ n_{e,r}^{-K}}}{\lambda_{(ridge)}} \underset{\substack{\geq \\ n^{K/4}}}{\|\theta\|^2} \geq n_{r,e}^{-K/2}$

and $-K/2 \geq \beta - 1/2$

Then almost surely for $n_{r/e}$ large enough and for all $\theta \in E_2^{inter}(n_{r,e})$

$$\boxed{R^{(ridge)}_{n,n_r}(f_\theta) \geq \mathcal{R}_n(f_\theta)}$$

There, on E_2^{inter}, the empirical risk controls the theoretical one

(c) $\forall \theta \in E_3^{small}$ $\quad \lambda_{(ridge)} \|\theta\|^2 \leq n_{e,r}^{-K/2}$

Using (*), we deduce that for all n_r large enough for all $\theta \in E_3^{small}$

$$\left| R^{(ridge)}_{n,n_r}(f_\theta) - \mathcal{R}_n(f_\theta) \right| \leq \underset{\uparrow}{2}\, \log^2(n_{r,e})\, n_{r,e}^{-K/2}$$

one coming from integral approx.
one coming from the ridge reg.

(d) Fix $\varepsilon > 0$

Call $(\theta_p)_p$ any minimizing sequence of the theoretical risk $\mathcal{R}_n$

$$\lim_{p \to +\infty} \mathcal{R}_n(f_{\theta_p}) = \inf_{\theta \in \Theta_{H,D}} \mathcal{R}_n(f_\theta)$$

By def$^\circ$ $\exists p_\varepsilon$ such that $\left|\mathcal{R}_n(f_{\theta_{p_\varepsilon}}) - \inf_{\theta \in \Theta_{H,D}} \mathcal{R}_n(f_\theta)\right| \leq \varepsilon$.

↳ For fixed n_r, according to (a), for p large enough

$\hat{\theta}^{(ridge)}_{p,n_r,D} \in E_2(n_r) \cup E_3(n_r)$.

↳ According to (b) & (c)

$$\mathcal{R}_n\left(f_{\hat{\theta}^{(ridge)}_{(p,n_r,D)}}\right) \leq R^{(ridge)}_{n,n_r}\left(f_{\hat{\theta}^{(ridge)}_{p,n_r,D}}\right) + 2\log^2(n_r)\, n_r^{-K/2}$$

↳ By def$^\circ$ of the minimizing sequence $\hat{\theta}^{(ridge)}_{p,n_r,D}$, for p large enough,

$$R^{(ridge)}_{n,n_r}\left(f_{\hat{\theta}^{(ridge)}_{p,n_r,D}}\right) \leq \inf_{\Theta_{H,D}} R^{(ridge)}_{n,n_r} + \varepsilon$$

↳ According to (c)

$$\inf_{\theta \in E_2^{large}(n_r) \cup E_3^{small}(n_r)} R^{(ridge)}_{n,n_r}(f_\theta) \leq \inf_{\theta \in E_3^{small}(n_r)} R^{(ridge)}_{n,n_r}(f_\theta)$$

$$\leq \inf_{\theta \in E_3^{small}(n_r)} \mathcal{R}_n(f_\theta) + 2\log^2(n_r)\, n_r^{-K/2}$$

↳ For all n_r large enough $\theta_{p_\varepsilon} \in E_3^{small}(n_r)$

therefore $\inf_{\theta \in E_3^{small}(n_r)} \mathcal{R}_n(f_\theta) \leq \mathcal{R}_n(f_{\theta_{p_\varepsilon}})$

Combining all, almost surely for n_r and p large enough

$$\mathcal{R}_n\left(f_{\hat{\theta}^{(ridge)}_{p,n_r,D}}\right) \leq \inf_{\Theta_{H,D}} \mathcal{R}_n(f_\theta) + 3\varepsilon$$

As ε is arbitrary, we get $\lim_{n_r \to +\infty} \lim_{p \to +\infty} \mathcal{R}_n\left(f_{\hat{\theta}^{(ridge)}_{(p,n_r,D)}}\right) = \inf_{\theta \in \Theta_{H,D}} \mathcal{R}_n(f_\theta)$.

$\blacksquare$

Open question:

- Implicit regularization in PINN training.
- Spectral bias.

Strong convergence of PINNs for linear PDE

- Good CV properties in terms of risk $\mathcal{R}_n$ do not imply good CV properties in terms of function, for instance

$$\hat{f}_{PINN} \xrightarrow{?} f^* \text{ in } L^2\text{-sense}(?)$$

- A first example in hybrid modeling setting

$d_{in} = 2 \qquad d_{out} = 1 \qquad \Omega =]0,1[\times]0,T[$

$\mathscr{D}(f,\cdot) = \partial_x f + \partial_t f$ (advection equation)

$h(x,0) = 1 \ \forall x$ (initial condition)

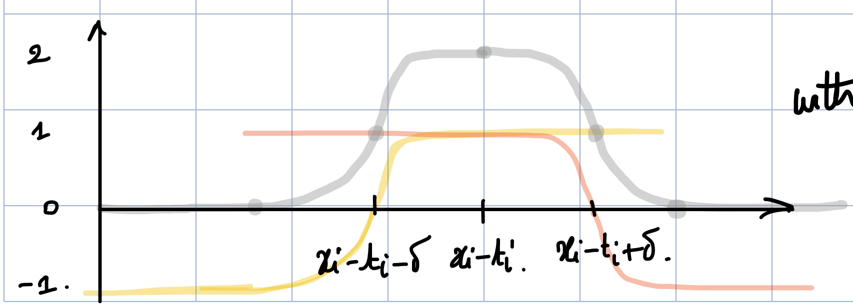
$h(0,t) = 1 \ \forall t$ (boundary condition)

Prior: the regression function f^* should be close to 1.

Consider the sequence of NN:

$$f_{\theta_p}(x,t) = 1 + \sum_{i=1}^{n} \frac{(Y_i - 1)}{2} \Big(\underbrace{\tanh_p^{oH}(x - t - x_i + t_i + \delta)}_{\to\, \mathrm{sgn}(x-t-x_i+t_i+\delta)} - \underbrace{\tanh_p^{oH}(x - t - x_i + t_i - \delta)}_{\to\, \mathrm{sgn}(x-t-x_i+t_i-\delta)} \Big)$$

$X_i = (x_i, t_i)$



with $\delta \leq \frac{1}{2} \min_{i \neq j} |x_i - x_j + t_i - t_j|$

Indeed, note that by the step shape, we get

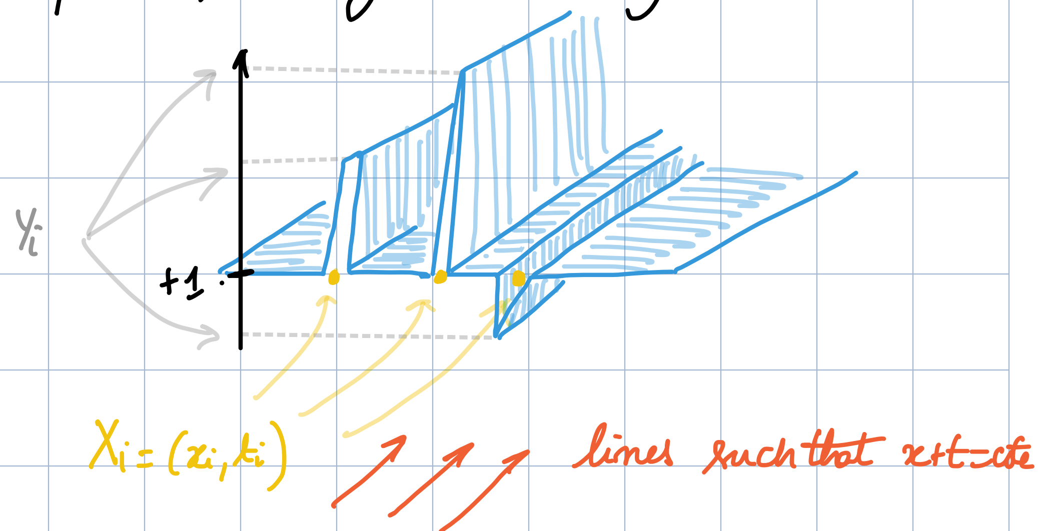
$$|f(x+\delta) - f(x)| = 1 = \left| \int_x^{x+\delta} f'(u)\,du \right| = \left| \langle f', \mathbb{1}_{[x,x+\delta]} \rangle_{L^2} \right|$$

$$\underset{\text{by Cauchy-Schwarz}}{\leq} \sqrt{\int_x^{x+\delta} (f')^2(u)\,du}\ \sqrt{\delta}.$$

Therefore $\|f'\|_{L^2(\Omega)} \geq 1/\sqrt{\delta}$. $\square$

⊕ $f_{\delta,p}(x_i, t_i) \xrightarrow[p \to +\infty]{} y_i$.

⊕ $f_{\delta,p}$ satisfies the equation, as being constant along the lines $x + t = cst$.



Therefore $\lim_{p \to +\infty} \mathcal{R}(f_{\delta,p}) = 0$ by ⊕ and ⊕.

If $D \geq 2n$, $\inf_{f \in NN_H(D)} \mathcal{R}_n(f) = 0 \implies f_{\hat{\theta}(\text{ridge})} \xrightarrow[\text{p.n.c.r.}]{L^2(\Omega)} \mathbb{1}_\Omega$

Fine if the model is accurate independently of n and f^*!!!

problematic if the model is not exact.

• A second example in PDE slices

$\Omega =]-1;1[$

$$\begin{cases} h(1) = 1 \\ \mathcal{D}(f,x) = x f'(x) \end{cases}$$

Clearly $f^*(x) = 1$ is the only strong sol° of the PDE

Consider the NN sequence to be $\hat{f}_p = \tanh_p \circ \tanh^{(H-1)}$

$\lim_{p\to+\infty} \mathcal{R}(\hat{f}_p) = \mathcal{R}(f^*) = 0$ (as $\hat{f}_p$ satisfies the ODE and $\hat{f}_p(1)=1$)

$(\hat{f}_p)_p$ is a minimizing sequence of $\mathcal{R}$.

But $\hat{f}_\infty(x) = \mathrm{sgn}(x) \neq f^*$.

• To retrieve strong convergence properties, one can use Sobolev regularization

for the sake of presentation.

• Assumption [Linear PDEs]

linear

Assume that $\mathcal{D}$ is an affine operator (of order K) of f and its derivatives

$$\mathcal{D}(f,x) = \mathcal{D}^{(lin)}(f,x) + B(x)$$

$$\mathcal{D}^{(lin)}(f,x) = \sum_{|\alpha| \leq K} \langle A_\alpha(x), \partial^\alpha f(x) \rangle$$

of order K.

$A_\alpha \in \mathcal{C}^\infty(\bar{\Omega}, \mathbb{R}^{d_{out}})$

in the paper, we can handle a 2nd member.

ex: advection, heat, wave, Maxwell equations.

• **Regularized theoretical risk**

$$\boxed{R_n^{(reg)}(f) = R_n(f) + \lambda_{(sob)} \|f\|^2_{H^{m+1}(\Omega)}}$$

with $\|f\|^2_{H^{m+1}(\Omega)} = \sum_{|\alpha| \leq m+1} \|\partial^\alpha f\|_{L^2(\Omega)}$.

• **Regularized empirical risk**

$$\boxed{R^{(reg)}_{n,n_e,n_r}(f) = R_{n,n_e,n_r}(f) + \lambda_{(ridge)} \|\theta\|_2^2 + \frac{\lambda_{(sob)}}{n_e} \sum_{\ell=1}^{n_e} \sum_{|\alpha| \leq m+1} \|\partial^\alpha f(X_\ell^{(r)})\|_2^2}$$

✓ Can be straightforwardly implemented in the usual PINN framework by considering the Sobolev penalty as additional PDE.

In what follows, for the sake of presentation, we discard the boundary condition (cf. gray parts).

Proposition [Characterization of the unique minimizer of $\mathcal{R}_n^{(reg)}$]

Assume $\mathcal{D}$ to be of order K and to encode a linear PDE.

$$m \geq \max\left(\lfloor d_{in}/2 \rfloor, K\right)$$

Then, $\mathcal{R}_n^{(reg)}$ has a unique minimizer $\hat{f}_n$ over $H^{m+1}(\Omega)$, characterized to be the unique element of $H^{m+1}(\Omega)$ such that

$$\forall v \in H^{m+1}(\Omega) \quad \boxed{\mathcal{A}_n(\hat{f}_n, v) = \mathcal{B}_n(v)}$$

Weak formulation of PDE on $H^{m}(\Omega, \mathbb{R}^{d_{out}})$

with
$$\mathcal{A}_n(\hat{f}_n, v) = \frac{1}{n}\sum_{i=1}^{n} \langle \Pi(\hat{f}_n)(X_i), \Pi(v)(X_i)\rangle$$
$$+ \lambda_e \mathbb{E}\langle \Pi(\hat{f}_n)(X^{(e)}), \Pi(v)(X^{(e)})\rangle$$
$$+ \frac{\lambda_r}{|\Omega|}\int_\Omega \mathcal{D}^{(lin)}(\hat{f}_n, x)\, \mathcal{D}^{(lin)}(v, x)\, dx$$
$$+ \frac{\lambda_{(reg)}}{|\Omega|}\sum_{|\alpha| \leq m+1} \int_\Omega \langle \partial^\alpha \hat{f}_n(x), \partial^\alpha v(x)\rangle\, dx$$

$$\mathcal{B}_n(v) = \frac{1}{n}\sum_{i=1}^{n} \langle Y_i, \Pi(v)(X_i)\rangle +$$
$$+ \lambda_e \mathbb{E}\left[\langle \Pi(v)(X^{e}), h(X^{(e)})\rangle\right]$$
$$- \frac{\lambda_r}{|\Omega|}\int_\Omega B(x)\, \mathcal{D}^{(lin)}(v, x)\, dx.$$

$$\Pi : H^{m+1}(\Omega, \mathbb{R}^{d_{out}}) \to C^0(\Omega, \mathbb{R}^{d_{out}})$$

so-called Sobolev embedding

such that $\Pi(f)$ is the unique continuous fct that coincides with f almost everywhere.

↳ Lax-Milgram based result.

↳ Remark that

$$R_n^{(reg)}(f) = \mathcal{A}_n(f,f) - 2\mathcal{B}_n(f) + \frac{1}{n}\sum_{i=1}^{n} \|Y_i\|^2 + \mu\, \mathbb{E}\|h(X^{(e)})\|_2^2 + \frac{1}{|\Omega|}\int_\Omega B^2(x)\,dx.$$

proof: For the sake of simplicity, in the proof, we discard the boundary conditions & 2nd member in the PDE

Remark that

$$① \quad \mathcal{A}_n(f,f) - 2\mathcal{B}_n(f) = R_n^{(reg)}(f) - \frac{1}{n}\sum_{i=1}^{n}\|Y_i\|^2$$

$$② \quad \mathcal{A}_n(f,f) \geq \lambda_{(sob)} \|f\|^2_{H^{m+1}(\Omega)}$$

meaning ~~that~~ $\mathcal{A}_n$ is coercive on the normed space $H^{m+1}(\Omega, \|.\|_{H^{m+1}(\Omega)})$

Since $m+1 > \max(d_1/2, K)$, one has

$$|\mathcal{A}_n(f,g)| \leq \left[C_\Omega^2 + \Big(\sum_{|\alpha|\leq K} \|A_{k,\alpha}\|_{\infty,\Omega}\Big)^2 + \lambda_{sob}\right] \|f\|^2_{H^{m+1}(\Omega)} \|g\|^2_{H^{m+1}(\Omega)}$$

and

$$|\mathcal{B}_n(f)| \leq C_\Omega \frac{1}{n}\sum_{i=1}^{n}\|Y_i\|_2 \quad \|f\|^2_{H^{m+1}(\Omega)}$$

therefore, $\mathcal{A}_n$ and $\mathcal{B}_n$ are continuous

By Lax-Milgram theorem [Brezis, 2010, Corollary 5.8]

$\exists!\ \hat{f} \in H^{m+1}(\Omega, \mathbb{R}^{d_{out}})$ such that

$$\mathcal{A}_n(\hat{f},\hat{f}) - 2\mathcal{B}_n(\hat{f}) = \min_{f\in H^{m+1}(\Omega)} \mathcal{A}_n(f,f) - 2\mathcal{B}_n(f)$$

$\hat{f}$ is the unique minimizer of $\mathcal{R}_n^{(reg)}$ over $H^{m+1}(\Omega, \mathbb{R}^{d_{out}})$

Again Lax_Milgram thus gives that $\hat{f}$ is the unique element of $H^{m+1}(\Omega)$ s.t $\forall g \in H^{m+1}(\Omega)$ $\mathcal{A}_n(\hat{f}, g) = \mathcal{B}_n(g)$.

$\blacksquare$

The former proposition ensures that the Sobolev regularization makes the PINN problem well_posed, i.e., the theoretical risk function admits a unique minimizer.

? Next question : what about CV to this unique minimizer?

Prop [From regularized risk consistency to strong convergence]

Assume that $m > \max(\lfloor d_{in}/2 \rfloor, K)$

Let $(f_p)_p \in C^\infty(\bar{\Omega}, \mathbb{R}^{d_{out}})$ — minimizing sequence of $\mathcal{R}_n^{(reg)}$

such that $\lim_{p\to+\infty} \mathcal{R}_n^{(reg)}(f_p) = \inf_{f \in C^\infty(\bar{\Omega}, \mathbb{R}^{d_{out}})} \mathcal{R}_n^{(reg)}(f)$

theoretical risk

Then

$$\lim_{p\to+\infty} \| f_p - \hat{f}_n \|_{H^m(\Omega)} = 0$$

unique minimizer of $\mathcal{R}_n^{(reg)}$ over $H^{m+1}(\Omega, \mathbb{R}^{d_{out}})$

Proof : admitted

Message : minimizing sqces of $\mathcal{R}_n^{(reg)}$ converge to the unique minimizer of $H^{m+1}(\Omega)$

Thm: [Strong convergence of regularized PINNs]

Assume that
- $\mathscr{D}$ is an affine operator
- $m \geq \max\left[\frac{d_{in}}{2}, K\right]$
- the condition function h is Lipschitz

Let $\left(\hat{\theta}^{(reg)}(p, n_e, n_r, D)\right)_p$ be a min. sequence of the regularized empirical risk function

$$R^{(reg)}_{n, n_e, n_r}(f_\theta) = R_{n, n_e, n_r}(f_\theta) + \lambda_{(ridge)} \|\theta\|_2^2 + \frac{\lambda_{(sob)}}{n_r} \sum_{l=1}^{n} \sum_{|\alpha| \leq m+1} \|\partial^\alpha f_\theta(X_e^{(r)})\|_2^2$$

over the class of $NN_H(D)$.

Then, choosing $\lambda_{(ridge)} = \min(n_e, n_r)^{-K}$ with $K = \frac{1}{12 + 4H(1 + (2+H)(m+2))}$

one has

$$\lim_{D \to +\infty} \lim_{n_e, n_r \to +\infty} \lim_{p \to +\infty} \| f_{\hat{\theta}^{(reg)}_{(p, n_e, n_r, D)}} - \hat{f}_n \|_{H^m(\Omega)} = 0 .$$

unique minimizer of $R_n^{(reg)}$ over $H^{m+1}(\Omega, \mathbb{R}^{d_{out}})$

Message: minimizing seqs of the empirical regularized risk converge to the unique minimizer of the regularized theoretical risk

✓ The PINN $f_{\hat{\theta}^{(reg)}}$ converges to the unique minimizer $\hat{f}_n$ of the theoretical risk

? What can we say in terms of CV to f^*?

Theu [Strong CV of linear PDE solvers].

Same assumptions as before.

+ the PDE admits a unique solution $f^\star$ in $H^{m+1}(\Omega, \mathbb{R}^{d_{out}})$

Then,

$$\lim_{\lambda^{(ridge)} \to 0} \lim_{D \to 0} \lim_{n_e, n_r \to +\infty} \lim_{p \to +\infty} \| f_{\hat{\theta}^{reg}_{(p, n_e, n_r, D, \lambda_{ridge})}} - f^\star \|_{H^m(\Omega)} = 0$$

For the hybrid modeling framework, we need to measure the gap between $f^\star$ and the physical prior.

Def: [physical inconsistency]

For any $f \in H^{m+1}(\Omega, \mathbb{R}^{d_{out}})$,

$$PI(f) = \lambda_e \mathbb{E} \| \Pi(f)(X^{(e)}) - h(X^{(e)}) \|^2 + \frac{\lambda_r}{|\Omega|} \int \mathcal{D}^2(u, x) dx$$

↳ PI measures the modeling error.

Better the model, lower $PI(f^\star)$

↳ Remark that $\mathcal{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \|\Pi(f)(X_i) - Y_i\|_2^2 + PI(f)$

Choose λ_{ridge} as before w.r.t n_r (and n_e) such that $\lambda_{ridge} \xrightarrow[n_e, n_r \to +\infty]{} 0$.

Prop: Under the same assumptions as before. Assume that $f^* \in H^{m+1}(\Omega, \mathbb{R}^{d_{out}})$, let

$$\lambda_{e/r} = \frac{\log n}{\sqrt{n}} \qquad \lambda_{(sob)} = \frac{1}{\sqrt{n}}$$

Then

STATISTICAL ACCURACY.

$$\lim_{D \to +\infty} \lim_{n_e, n_r \to +\infty} \lim_{p \to +\infty} \mathbb{E} \int_\Omega \| f^{(n)}_{\hat{\theta}^{(reg)}_{(p, n_e, n_r, D)}} - f^* \|_2^2 \, d\mu_X \lesssim \frac{\log^2(n)}{\sqrt{n}}$$

PHYSICAL CONSISTENCY.

$$\lim_{D \to +\infty} \lim_{n_e, n_r \to +\infty} \lim_{p \to +\infty} \mathbb{E}\left[PI\left(f^{(n)}_{\hat{\theta}^{(reg)}_{(p, n_e, n_r, D)}} \right) \right] \leqslant PI(f^*) + \underset{n \to +\infty}{o}(1).$$

In the lecture: $\lambda_r = \lambda_e = \frac{\log n}{\sqrt{n}}$ $\quad \lambda_{sob} = \frac{1}{\sqrt{n}}$ $\quad \lambda_{data} = 1$

Choice of λ_{ridge} unaffected by the rescaling since we take the asymptotic in n_e, n_r first.

*Choice in the article $\lambda_r = 1$ $\lambda_e = 1$ $\lambda_{sob} = 1/\log n$ $\lambda_{data} = \sqrt{n}/\log n$