Convergence and error analysis of PINNs

Nathan Doumèche, Gérard Biau, Claire Boyer

CONVERGENCE AND ERROR ANALYSIS OF PINNS

BY NATHAN DOUMÈCHE[1,a], GÉRARD BIAU[1,b] AND CLAIRE BOYER[1,c]

[1]*Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université,* [a]*nathan.doumeche@sorbonne-universite.fr;* [b]*gerard.biau@sorbonne-universite.fr;* [c]*claire.boyer@sorbonne-universite.fr*

Physics-informed neural networks (PINNs) are a promising approach that combines the power of neural networks with the interpretability of physical modeling. PINNs have shown good practical performance in solving partial differential equations (PDEs) and in hybrid modeling scenarios, where physical models enhance data-driven approaches. However, it is essential to establish their theoretical properties in order to fully understand their capabilities and limitations. In this study, we highlight that classical training of PINNs can suffer from systematic overfitting. This problem can be addressed by adding a ridge regularization to the empirical risk, which ensures that the resulting estimator is risk-consistent for both linear and nonlinear PDE systems. However, the strong convergence of PINNs to a solution satisfying the physical constraints requires a more involved analysis using tools from functional analysis and calculus of variations. In particular, for linear PDE systems, an implementable Sobolev-type regularization allows to reconstruct a solution that not only achieves statistical accuracy but also maintains consistency with the underlying physics.

1. Introduction.

Physics-informed machine learning. Advances in machine learning and deep learning have led to significant breakthroughs in almost all areas of science and technology. However, despite remarkable achievements, modern machine learning models are difficult to interpret and do not necessarily obey the fundamental governing laws of physical systems [Linardatos et al., 2021]. Moreover, they often fail to extrapolate scenarios beyond those on which they were trained [Xu et al., 2021]. On the contrary, numerical or pure physical methods struggle to capture nonlinear relationships in complex and high-dimensional systems, while lacking flexibility and being prone to computational problems. This state of affairs has led to a growing consensus that data-driven machine learning methods need to be coupled with prior scientific knowledge based on physics. This emerging field, often called physics-informed machine learning [Raissi et al., 2019], seeks to combine the predictive power of machine learning techniques with the interpretability and robustness of physical modeling. The literature in this field has is still disorganized, with a somewhat unstable nomenclature. In particular, the terms physics-informed, physics-based, physics-guided, and theory-guided are used interchangeably. For a comprehensive account, we refer to the reviews by Rai and Sahu [2020], Karniadakis et al. [2021], Cuomo et al. [2022], and Hao et al. [2022], which survey some of the prevailing trends in embedding physical knowledge in machine learning, present some of the current challenges, and discuss various applications.

Vocabulary and use cases. Depending on the nature of the interaction between machine learning and physics, physics-informed machine learning is usually achieved by preprocessing the features [Rai and Sahu, 2020], by designing innovative network architectures that incorporate the physics of the problem [Karniadakis et al., 2021], or by forcing physics infusion

MSC2020 subject classifications: Primary 62G08; secondary 68T07.
Keywords and phrases: Physics-informed neural networks, Hybrid modeling, PDE solver, Consistency.

into the loss function [Cuomo et al., 2022]. It is this latter approach, which is most often referred to as physics regularization [Rai and Sahu, 2020], to which our article is devoted. Note that other names are possible, including physics consistency penalty [Wang et al., 2020a], knowledge-based loss term [von Rueden et al., 2023], and physics-guided neural networks [Cunha et al., 2022]. In the following, we will focus more specifically on neural networks incorporating a physical regularization, called PINNs (for physics-informed neural networks, Raissi et al. 2019). Such models have been successfully applied to (i) model hybrid learning tasks, where the data-driven loss is regularized to satisfy a physical prior, and (ii) design efficient solvers of partial differential equations (PDEs). A significant advantage of PINNs is that they are easy to implement compared to other PDE solvers, and that they rely on the backpropagation algorithm, resulting in reasonable computational cost. Although (i) and (ii) are different facets of the same mathematical problem, they differ in their geometry and the nature of the data on which they are based, as we will see later.

Related work and contributions. Despite a rapidly growing literature highlighting the capabilities of PINNs in various real-world applications, there are still few theoretical guarantees regarding the overfitting, consistency, and error analysis of the approach. Most existing theoretical work focuses either on intractable modifications of PINNs [Cuomo et al., 2022] or on negative results, such as in Krishnapriyan et al. [2021] and Wang et al. [2022].

Our goal in the present article is to provide a comprehensive theoretical analysis of the mathematical forces driving PINNs, in both the hybrid modeling and PDE solver settings, with the constant concern to provide approaches that can be implemented in practice. Our results complement those of Shin [2020], Shin et al. [2020], Mishra and Molinaro [2023], De Ryck and Mishra [2022], Wu et al. [2022], and Qian et al. [2023] for the PDE solver problem. Shin [2020] and Wu et al. [2022] focus on modifications of PINNs using the Hölder norm of the neural network in the loss function, which is unfortunately intractable in practice. In the context of linear PDEs, Shin et al. [2020] analyze the expected generalization error of PINNs using the Rademacher complexity of the image of the neural network class by a differential operator. However, this Rademacher complexity does not obviously vanish with increasing sample size. Similarly, Mishra and Molinaro [2023] bound the generalization error by a quadrature rule depending on the Hölder norm of the neural network, which does not necessarily tend to zero as the number of training points tends to infinity. De Ryck and Mishra [2022] derive bounds on the expectation of the L^2 error, provided that the weights of the neural networks are bounded. In contrast to this series of works, we consider models and assumptions that can be practically verified or implemented. Moreover, our approach includes hybrid modeling, for which, as pointed out by Karniadakis et al. [2021], no theoretical guarantees have been given so far. Preliminary interesting results on the statistical consistency of a regression function penalized by a PDE are reported in Arnone et al. [2022]. The original point of our approach lies in the use of a mix of statistical and functional analysis arguments [Evans, 2010] to characterize the PINN problem.

Overview. After correctly defining the PINN problem in Section 2, we show in Section 3 that an additional regularization term is needed in the loss, otherwise PINNs can overfit. This first important result is consistent with the approach of Shin [2020], which penalizes PINNs by Hölder norms to ensure their convergence, and with the experiments of Nabian and Meidani [2020], which improve performance by adding an extra-regularization term. In Section 4, we establish the consistency of ridge PINNs by proving in Theorem 4.6 that a slowly vanishing ridge penalty is sufficient to prevent overfitting. Finally, in Section 5, we show that an additional level of regularization is sufficient in order to guarantee the strong convergence of PINNs (Theorem 5.7). We also prove that an adapted tuning of the hyperparameters allows to reconstruct the solution in the PDE solver setting (Theorem 5.8), as well

as to ensure both statistical and physics consistency in the hybrid modeling setting (Theorems 5.13). All proofs are postponed to the appendices. The code of all the numerical experiments can be found at https://github.com/NathanDoumeche/Convergence_and_error_analysis_of_PINNs.

2. The PINN framework. In its most general formulation, the PINN method can be described as an empirical risk minimization problem, penalized by a PDE system.

Notation. Throughout this article, the symbol $\mathbb{E}$ denotes expectation and $\|\cdot\|_2$ (resp., $\langle\cdot,\cdot\rangle$) denotes the Euclidean norm (resp., scalar product) in $\mathbb{R}^d$, where d may vary depending on the context. Let $\Omega \subset \mathbb{R}^{d_1}$ be a bounded Lipschitz domain with boundary $\partial\Omega$ and closure $\bar{\Omega}$, and let $(\mathbf{X}, Y) \in \Omega \times \mathbb{R}^{d_2}$ be a pair of random variables. Recall that Lipschitz domains are a general category of open sets that includes bounded convex domains (such as $]0,1[^{d_1}$) and usual manifolds with C^1 boundaries (see Appendix A). This level of generality with respect to the domain Ω is necessary to encompass most of the physical problems, such as those presented in Arzani et al. [2021], which use non-trivial (but Lipschitz) geometries. For $K \in \mathbb{N}$, the space of functions from Ω to $\mathbb{R}^{d_2}$ that are K times continuously differentiable is denoted by $C^K(\Omega, \mathbb{R}^{d_2})$.

Let $C^\infty(\Omega, \mathbb{R}^{d_2}) = \cap_{K\geqslant 0} C^K(\Omega, \mathbb{R}^{d_2})$ be the space of infinitely differentiable functions. The space $C^K(\Omega, \mathbb{R}^{d_2})$ is endowed with the Hölder norm $\|\cdot\|_{C^K(\Omega)}$, defined for any u by $\|u\|_{C^K(\Omega)} = \max_{|\alpha|\leqslant K} \|\partial^\alpha u\|_{\infty,\Omega}$. The space $C^\infty(\bar{\Omega}, \mathbb{R}^{d_2})$ of smooth functions is defined as the subspace of continuous functions $u : \bar{\Omega} \to \mathbb{R}^{d_2}$ satisfying $u|_\Omega \in C^\infty(\Omega, \mathbb{R}^{d_2})$ and, for all $K \in \mathbb{N}$, $\|u\|_{C^K(\Omega)} < \infty$. A differential operator $\mathscr{F} : C^\infty(\Omega, \mathbb{R}^{d_2}) \times \Omega \to \mathbb{R}$ is said to be of order K if it can be expressed as a function over the partial derivatives of order less than or equal to K. For example, the operator $\mathscr{F}(u, \mathbf{x}) = \partial_1 u(\mathbf{x}) \partial^2_{1,2} u(\mathbf{x}) + u(\mathbf{x}) \sin(\mathbf{x})$ has order 2. A summary of the mathematical notation used in this paper is to be found in Appendix A.

Hybrid modeling. As in classical regression analysis, we are interested in estimating the unknown regression function $u^\star$ such that $Y = u^\star(\mathbf{X}) + \varepsilon$, for some random noise ε that satisfies $\mathbb{E}(\varepsilon|\mathbf{X}) = 0$. What makes the problem original is that the function $u^\star$ is assumed to satisfy (at least approximately) a collection of $M \geqslant 1$ PDE-type constraints of order at most K, denoted in a standard form by $\mathscr{F}_k(u^\star, \mathbf{x}) \simeq 0$ for $1 \leqslant k \leqslant M$. It is therefore assumed that $u^\star$ can be derived K times. Moreover, there exists some subset $E \subseteq \partial\Omega$ and an initial/boundary condition function $h : E \to \mathbb{R}^{d_2}$ such that, for all $\mathbf{x} \in E$, $u^\star(\mathbf{x}) \simeq h(\mathbf{x})$. We stress that E can be strictly included in Ω, as shown in Example 2.2 for a spatio-temporal domain Ω. The specific case $E = \partial\Omega$ corresponds to Dirichlet boundary conditions.

These constraints model some a priori physical information about $u^\star$. However, this knowledge may be incomplete (e.g., the PDE system may be ill-posed and have no or multiple solutions) and/or imperfect (i.e., there is some modeling error, that is, $\mathscr{F}_k(u^\star, \mathbf{x}) \neq 0$ and $u^\star|_E \neq h$). This again emphasizes that $u^\star$ is not necessarily a solution of the system of differential equations.

EXAMPLE 2.1 (Maxwell equations). Let $\mathbf{x} = (x, y, z, t) \in \mathbb{R}^3 \times \mathbb{R}_+$, and consider Maxwell equations describing the evolution of an electro-magnetic field $u^\star = (E^\star, B^\star)$ in vacuum, defined by

$$\begin{cases} \mathscr{F}_1(u^\star, \mathbf{x}) = \mathrm{div} E^\star(\mathbf{x}) \\ \mathscr{F}_2(u^\star, \mathbf{x}) = \mathrm{div} B^\star(\mathbf{x}) \\ (\mathscr{F}_3,\ \mathscr{F}_4,\ \mathscr{F}_5)(u^\star, \mathbf{x}) = \partial_t E^\star(\mathbf{x}) - \mathrm{curl} B^\star(\mathbf{x}) \\ (\mathscr{F}_6,\ \mathscr{F}_7,\ \mathscr{F}_8)(u^\star, \mathbf{x}) = \partial_t B^\star(\mathbf{x}) + \mathrm{curl} E^\star(\mathbf{x}), \end{cases}$$

where $E^\star \in C^1(\mathbb{R}^4, \mathbb{R}^3)$ is the electric field, $B^\star \in C^1(\mathbb{R}^4, \mathbb{R}^3)$ the magnetic field, and the div and curl operators are respectively defined for $F = (F_x, F_y, F_z) \in C^1(\mathbb{R}^4, \mathbb{R}^3)$ by

$$\mathrm{div} F = \partial_x F_x + \partial_y F_y + \partial_z F_z \quad \text{and} \quad \mathrm{curl} F = (\partial_y F_z - \partial_z F_y,\ \partial_z F_x - \partial_x F_z,\ \partial_x F_y - \partial_y F_x).$$

In this case, $d_1 = 4$, $d_2 = 6$, and $M = 8$. □

EXAMPLE 2.2 (Spatio-temporal condition function). Assume that the domain $\Omega \subseteq \mathbb{R}^{d_1}$ is of the form $\Omega = \Omega_1 \times]0, T[$, where $\Omega_1 \subseteq \mathbb{R}^{d_1 - 1}$ is a bounded Lipschitz domain and $T \geqslant 0$ is a finite time horizon. The spatio-temporal PDE system admits (spatial) boundary conditions specified by a function $f : \partial\Omega_1 \to \mathbb{R}^{d_2}$, i.e.,

$$\forall x \in \partial\Omega_1,\ \forall t \in [0, T], \quad u^\star(x, t) = f(x),$$

and a (temporal) initial condition specified by a function $g : \partial\Omega_1 \to \mathbb{R}^{d_2}$, that is

$$\forall x \in \Omega_1, \quad u^\star(x, 0) = g(x).$$

The set on which the boundary and initial conditions are defined is $E = (\Omega_1 \times \{0\}) \cup (\partial\Omega_1 \times [0, T])$, and the associated condition function $h : E \to \mathbb{R}^{d_2}$ is

$$h(\mathbf{x}) = \begin{cases} f(x) \text{ if } \mathbf{x} = (x, t) \in \partial\Omega_1 \times [0, T] \\ g(x) \text{ if } \mathbf{x} = (x, t) \in \Omega_1 \times \{0\}. \end{cases}$$

Notice that $E \subsetneq \partial\Omega$. □

In order to estimate $u^\star$, we assume to have at hand three sets of data:

(i) A collection of i.i.d. random variables $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ distributed as $(\mathbf{X}, Y) \in \Omega \times \mathbb{R}^{d_2}$, the distribution of which is *unknown*;

(ii) A collection of i.i.d. random variables $\mathbf{X}_1^{(e)}, \ldots, \mathbf{X}_{n_e}^{(e)}$ distributed according to some *known* distribution μ_E on E;

(iii) A sample of i.i.d. random variables $\mathbf{X}_1^{(r)}, \ldots, \mathbf{X}_{n_r}^{(r)}$ *uniformly distributed* on Ω.

The function $u^\star$ is then estimated by minimizing the empirical risk function

$$R_{n,n_e,n_r}(u_\theta) = \frac{\lambda_d}{n} \sum_{i=1}^{n} \|u_\theta(\mathbf{X}_i) - Y_i\|_2^2 + \frac{\lambda_e}{n_e} \sum_{j=1}^{n_e} \|u_\theta(\mathbf{X}_j^{(e)}) - h(\mathbf{X}_j^{(e)})\|_2^2$$

$$+ \frac{1}{n_r} \sum_{k=1}^{M} \sum_{\ell=1}^{n_r} \mathscr{F}_k(u_\theta, \mathbf{X}_\ell^{(r)})^2 \tag{1}$$

over the class $\mathrm{NN}_H(D) := \{u_\theta, \theta \in \Theta_{H,D}\}$ of feedforward neural networks with H hidden layers of common width D (see below for a precise definition), where $(\lambda_d, \lambda_e) \in \mathbb{R}_+^2 \backslash (0, 0)$ are hyperparameters that establish a tradeoff between the three terms. In practice, one often encounters the case where $\lambda_e = 0$ (data + PDEs). Another situation of interest is when $\lambda_d = 0$ (PDEs + initial/boundary conditions), which corresponds to the special case of a PDE solver. Setting (1) is more general as it includes all the combinations data + PDEs + initial/boundary conditions. Since a minimizer of the empirical risk function (1) does not necessarily exist, we denote by $(\hat{\theta}(p, n_e, n_r, D))_{p \in \mathbb{N}} \in \Theta_{H,D}^{\mathbb{N}}$ any minimizing sequence, i.e.,

$$\lim_{p \to \infty} R_{n,n_e,n_r}(u_{\hat{\theta}(p,n_e,n_r,D)}) = \inf_{\theta \in \Theta_{H,D}} R_{n,n_e,n_r}(u_\theta).$$

In practice, such a sequence is usually obtained by implementing some optimization procedure, the exact description of which is not important for our purpose.

On the practical side, simulations using hybrid modeling have been successfully applied to model image denoising [Wang et al., 2020a], turbulence [Wang et al., 2020b], blood streams [Arzani et al., 2021], wave propagation [Davini et al., 2021], and ocean streams [de Wolff et al., 2021]. Experiments with real data have been performed to assess the sea temperature [de Bézenac et al., 2019], subsurface transport [He et al., 2020], fused filament fabrication [Kapusuzoglu and Mahadevan, 2020], seismic response [Zhang et al., 2020], glacier dynamic [Riel et al., 2021], lake temperature [Daw et al., 2022], thermal modeling of buildings [Gokhale et al., 2022], blasts [Pannell et al., 2022], and heat transfers [Ramezankhani et al., 2022]. The generality and flexibility of the empirical risk function (1) allows it to encompass most PINN-like problems. For example, the case $M \geqslant 2$ is considered in de Bézenac et al. [2019] and Riel et al. [2021], while Zhang et al. [2020] and Wang et al. [2020b] assume that $d_1 = d_2 = 3$. Importantly, the situation where $\lambda_d > 0$ *and* $\lambda_e > 0$ (data + boundary conditions + PDEs) is also interesting from a physical point of view. This is, for example, the approach advocated by Arzani et al. [2021], which uses both data and boundary conditions (see also Cuomo et al., 2022, and Hao et al., 2022).

The PDE solver case. The particular case $\lambda_d = 0$ deserves a special comment. In this setting, without physical measures $(\mathbf{X}_i, Y_i)$, the function $u^\star$ is viewed as the unknown solution of the system of PDEs $\mathscr{F}_1, \dots, \mathscr{F}_M$ with initial/boundary conditions h. The goal is to estimate the solution $u^\star$ of the PDE problem

$$\begin{cases} \forall k,\, \forall \mathbf{x} \in \Omega,\, \mathscr{F}_k(u^\star, \mathbf{x}) = 0 \\ \forall \mathbf{x} \in E, \qquad\quad u^\star(\mathbf{x}) = h(\mathbf{x}), \end{cases}$$

with neural networks from $\mathrm{NN}_H(D)$. In this case, the empirical risk function (1) becomes

$$R_{n_e,n_r}(u_\theta) = \frac{\lambda_e}{n_e} \sum_{j=1}^{n_e} \|u_\theta(\mathbf{X}_j^{(e)}) - h(\mathbf{X}_j^{(e)})\|_2^2 + \frac{1}{n_r} \sum_{k=1}^{M} \sum_{\ell=1}^{n_r} \mathscr{F}_k(u_\theta, \mathbf{X}_\ell^{(r)})^2,$$

where the boundary and initial conditions $(\mathbf{X}_1^{(e)}, h(\mathbf{X}_1^{(e)})), \dots, (\mathbf{X}_{n_e}^{(e)}, h(\mathbf{X}_{n_e}^{(e)}))$ are sampled on $E \times \mathbb{R}^{d_2}$ according to some known distribution μ_E, and $(\mathbf{X}_1^{(r)}, \dots, \mathbf{X}_{n_r}^{(r)})$ are uniformly distributed on Ω. Note that, for simplicity, we write $R_{n_e,n_r}(u_\theta)$ instead of $R_{n,n_e,n_r}(u_\theta)$ because no $\mathbf{X}_i$ is involved in this context. Since no confusion is possible, the same convention is used for all subsequent risk functions throughout the paper. The first term of $R_{n_e,n_r}(u_\theta)$ measures the gap between the network u_θ and the condition function h on E, while the second term forces u_θ to obey the PDE in a discretized way. Since both the condition function h and the distribution μ_E are known, it is reasonable to think of n_e and n_r as large (up to the computational resources). In this scientific computing perspective, PINNs have been successfully applied to solve a wide variety of linear and nonlinear problems, including motion, advection, heat, Euler, high-frequency Helmholtz, Schrödinger, Blasius, Burgers, and Navier-Stokes equations, covering various fields ranging from classical (mechanics, fluid dynamics, thermodynamics, and electromagnetism) to quantum physics [e.g., Cuomo et al., 2022, Li et al., 2023].

The class of neural networks. A fully-connected feedforward neural network with $H \in \mathbb{N}^\star$ hidden layers of sizes $(L_1, \dots, L_H) := (D, \dots, D) \in (\mathbb{N}^\star)^H$ and activation $\tanh$, is a function from $\mathbb{R}^{d_1}$ to $\mathbb{R}^{d_2}$, defined by

$$u_\theta = \mathcal{A}_{H+1} \circ (\tanh \circ \mathcal{A}_H) \circ \cdots \circ (\tanh \circ \mathcal{A}_1),$$

where the hyperbolic tangent function $\tanh$ is applied element-wise. Each $\mathcal{A}_k : \mathbb{R}^{L_{k-1}} \to \mathbb{R}^{L_k}$ is an affine function of the form $\mathcal{A}_k(\mathbf{x}) = W_k \mathbf{x} + b_k$, with W_k a $(L_{k-1} \times L_k)$-matrix, $b_k \in \mathbb{R}^{L_k}$ a vector, $L_0 = d_1$, and $L_{H+1} = d_2$. The neural network u_θ is parameterized by

$\theta = (W_1, b_1, \dots, W_{H+1}, b_{H+1}) \in \Theta_{H,D}$, where $\Theta_{H,D} = \mathbb{R}^{\sum_{i=0}^{H}(L_i+1)\times L_{i+1}}$. Throughout, we let $\mathrm{NN}_H(D) = \{u_\theta,\ \theta \in \Theta_{H,D}\}$. We emphasize that the $\tanh$ function is the most common activation in PINNs [see, e.g., Cuomo et al., 2022]. It is preferable to the classical ReLU $(x) = \max(x, 0)$ activation. In fact, since ReLU neural networks are a subset of piecewise linear functions, their high derivatives vanish and therefore cannot be captured by the penalty term $\frac{1}{n_r}\sum_{k=1}^{M}\sum_{\ell=1}^{n_r}\mathscr{F}_k(u_\theta, \mathbf{X}_\ell^{(r)})^2$.

The parameter space $\mathrm{NN}_H(D)$ must be chosen large enough to approximate both the solutions of the PDEs and their derivatives. This property is encapsulated in Proposition 2.3, which shows that for any number $H \geqslant 2$ of hidden layers, the set $\mathrm{NN}_H := \cup_D \mathrm{NN}_H(D)$ is dense in the space $(C^\infty(\bar{\Omega}, \mathbb{R}^{d_2}), \|\cdot\|_{C^K(\Omega)})$. This generalizes Theorem 5.1 in De Ryck et al. [2021] which states that NN_2 is dense in $(C^\infty([0,1]^{d_1}, \mathbb{R}), \|\cdot\|_{C^K(]0,1[^{d_1})})$ for all $d_1 \geqslant 1$ and $K \in \mathbb{N}$.

PROPOSITION 2.3 (Density of neural networks in Hölder spaces). *Let $K \in \mathbb{N}$, $H \geqslant 2$, and $\Omega \subseteq \mathbb{R}^{d_1}$ be a bounded Lipschitz domain. Then $\mathrm{NN}_H := \cup_D \mathrm{NN}_H(D)$ is dense in $(C^\infty(\bar{\Omega}, \mathbb{R}^{d_2}), \|\cdot\|_{C^K(\Omega)})$, i.e., for any function $u \in C^\infty(\bar{\Omega}, \mathbb{R}^{d_2})$, there exists a sequence $(u_p)_{p\in\mathbb{N}} \in \mathrm{NN}_H^{\mathbb{N}}$ such that $\lim_{p\to\infty}\|u - u_p\|_{C^K(\Omega)} = 0$.*

In the remainder of the article, the number H of hidden layers is considered to be fixed. Krishnapriyan et al. [2021] use $\mathrm{NN}_4(50)$, Xu et al. [2021] take $\mathrm{NN}_5(100)$, whereas Arzani et al. [2021] employ $\mathrm{NN}_{10}(100)$. It is worth noting that in this series of papers the width D is much larger than H, as in Proposition 2.3.

3. PINNs can overfit. Our goal in this section is to show through two examples how learning with standard PINNs can lead to severe overfitting problems. This weakness has already been noted in Costabal et al. [2020], Nabian and Meidani [2020], Chandrajit et al. [2023], and Esfahani [2023], which propose to improve the performance of their models by resorting to an additional regularization strategy. The pathological cases that we highlight both rely on neural networks with exploding derivatives.

The theoretical risk function is defined by

(2)
$$\mathscr{R}_n(u) = \frac{\lambda_d}{n}\sum_{i=1}^{n}\|u(\mathbf{X}_i) - Y_i\|_2^2 + \lambda_e\mathbb{E}\|u(\mathbf{X}^{(e)}) - h(\mathbf{X}^{(e)})\|_2^2 + \frac{1}{|\Omega|}\sum_{k=1}^{M}\int_\Omega \mathscr{F}_k(u, \mathbf{x})^2 d\mathbf{x}.$$

Observe that in $\mathscr{R}_n(u)$ we take expectation with respect to μ_E (for the initial/boundary condition part) and integrate with respect to the uniform measure on Ω (for the PDE part), but keep the term $\sum_{i=1}^{n}\|u_\theta(\mathbf{X}_i) - Y_i\|_2^2$ intact. This regime corresponds to the limit of the empirical risk function (1), holding n fixed and letting $n_e, n_r \to \infty$. The rationale is that while the random samples $(\mathbf{X}_i, Y_i)$ may be limited in number (e.g., because their acquisition is more delicate and require physical measurements), this is not the case for $\mathbf{X}_j^{(e)}$ or $\mathbf{X}_j^{(r)}$, which can be freely sampled (up to computational resources). Note however that in the PDE solver setting, the first term is not included.

Given any minimizing sequence $(\hat{\theta}(p, n_e, n_r, D))_{p\in\mathbb{N}}$ of the empirical risk, satisfying

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

a natural requirement, called risk-consistency, is that

$$\lim_{n_e,n_r\to\infty}\lim_{p\to\infty}\mathscr{R}_n(u_{\hat{\theta}(p,n_e,n_r,D)}) = \inf_{u\in\mathrm{NN}_H(D)}\mathscr{R}_n(u).$$

We show below that standard PINNs can dramatically fail to be risk-consistent, through two counterexamples, one in the hybrid modeling context and one in the specific PDE solver setting.

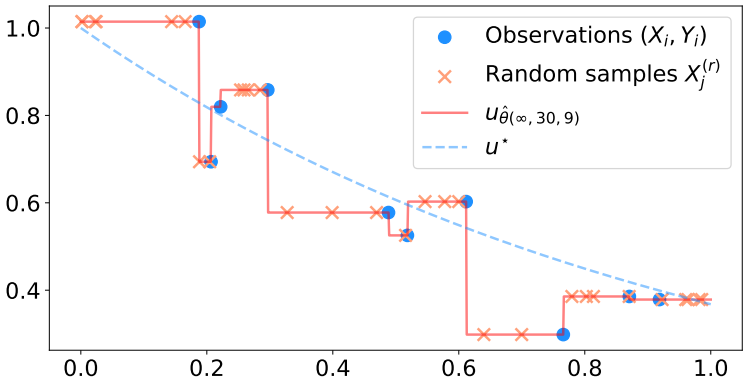

FIG 1. *Example of an inconsistent PINN estimator in hybrid modeling with* $m = \gamma = 1$, $\varepsilon \sim \mathcal{N}(0, 10^{-2})$, *and* $n = 10$.

The case of dynamics with friction. Consider the following ordinary differential constraint, defined on the domain $\Omega =]0, T[$ (with closure $\bar{\Omega} = [0, T]$) by

$$\forall u \in C^2(\bar{\Omega}, \mathbb{R}),\ \forall \mathbf{x} \in \Omega, \quad \mathscr{F}(u, \mathbf{x}) = mu''(\mathbf{x}) + \gamma u'(\mathbf{x}). \tag{3}$$

This models the dynamics of an object of mass $m > 0$, subjected to a fluid force of friction coefficient $\gamma > 0$. The goal is to reconstruct the real trajectory $u^\star$ by taking advantage of the model $\mathscr{F}$ and the noisy observations Y_i at the $\mathbf{X}_i$. This is an example where the modeling is perfect, i.e., $\mathscr{F}(u^\star, \cdot) = 0$, but the challenge is that the physical model is incomplete because the boundary conditions are unknown. Following the hybrid modeling framework, the trajectory $u^\star$ is estimated by minimizing over the space $\mathrm{NN}_H(D)$ the empirical risk function

$$R_{n,n_r}(u_\theta) = \frac{\lambda_d}{n} \sum_{i=1}^{n} |u_\theta(\mathbf{X}_i) - Y_i|^2 + \frac{1}{n_r} \sum_{\ell=1}^{n_r} \mathscr{F}(u_\theta, \mathbf{X}_\ell^{(r)})^2.$$

PROPOSITION 3.1 (Overfitting). *Consider the dynamics with friction model* (3)*, and assume that there are two observations such that* $Y_i \neq Y_j$. *Then, whenever* $D \geqslant n - 1$, *for any integer* n_r, *for all* $\mathbf{X}_1^{(r)}, \ldots, \mathbf{X}_{n_r}^{(r)}$, *there exists a minimizing sequence* $(u_{\hat{\theta}(p,n_r,D)})_{p\in\mathbb{N}} \in \mathrm{NN}_H(D)^{\mathbb{N}}$ *such that* $\lim_{p\to\infty} R_{n,n_r}(u_{\hat{\theta}(p,n_r,D)}) = 0$ *but* $\lim_{p\to\infty} \mathscr{R}_n(u_{\hat{\theta}(p,n_r,D)}) = \infty$. *So, this PINN estimator is not consistent.*

Proposition 3.1 illustrates how fitting a PINN by minimizing the empirical risk alone can lead to a catastrophic situation, where the empirical risk of the minimizing sequence is (close to) zero, while its theoretical risk is infinite. This phenomenon is explained by the existence of piecewise constant functions interpolating the observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$, whose derivatives are null at the points $\mathbf{X}_1^{(r)}, \ldots, \mathbf{X}_{n_r}^{(r)}$, but diverge between these points (see Figure 1). These functions correspond to neural networks u_θ such that $\|\theta\|_2 \to \infty$.

PDE solver: The heat propagation case. Consider the heat propagation differential operator defined on the domain $\Omega =]-1, 1[\times]0, T[$ (with closure $\bar{\Omega} = [-1, 1] \times [0, T]$) by

$$\forall u \in C^2(\bar{\Omega}, \mathbb{R}),\ \forall \mathbf{x} \in \Omega, \quad \mathscr{F}(u, \mathbf{x}) = \partial_t u(\mathbf{x}) - \partial^2_{x,x} u(\mathbf{x}), \tag{4}$$

associated with the boundary conditions

$$\forall t \in [0, T], \quad u(-1, t) = u(1, t) = 0,$$

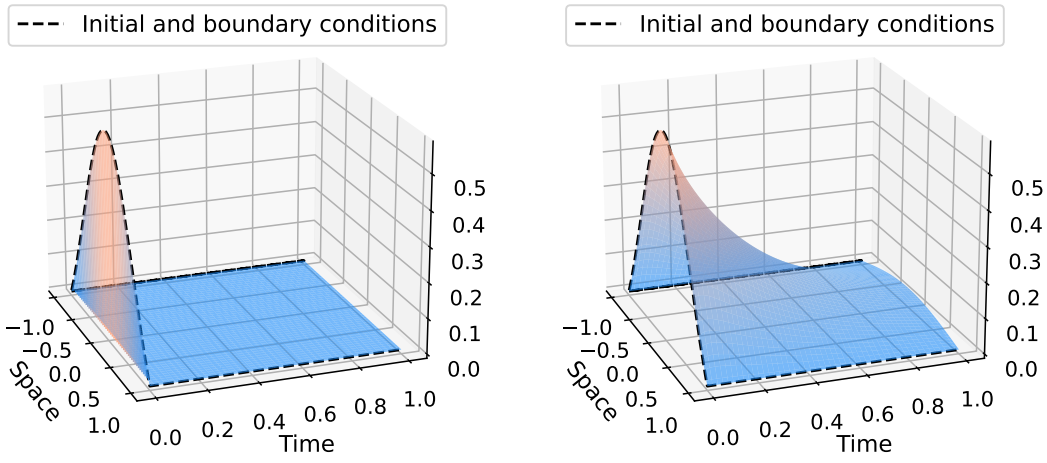

FIG 2. *Inconsistent PINN (left) compared to the solution $u^\star$ of the PDE (right) for the heat propagation case.*

and the initial condition defined, for all $x \in [-1, 1]$, by

$$u(x, 0) = \tanh^{\circ H}(x + 0.5) - \tanh^{\circ H}(x - 0.5) + \tanh^{\circ H}(0.5) - \tanh^{\circ H}(1.5).$$

The notation $\tanh^{\circ k}$ stands for the function recursively defined by $\tanh^{\circ 1} = \tanh$ and $\tanh^{\circ(k+1)} = \tanh \circ \tanh^{\circ k}$. The unique solution $u^\star$ of the PDE is shown in Figure 2 (right). It models the time evolution of the temperature of a wire, whose extremities at $x = -1$ and $x = 1$ are maintained at zero temperature. Note that the initial condition corresponds to a bell-shaped function, which belongs to $\mathrm{NN}_H(2)$. However, the setting can be extended to arbitrary initial conditions that take the form of a neural network function, given the boundary condition $u(\partial\Omega \times [0, T]) = \{0\}$.

To solve the PDE (4), we use n_e i.i.d. samples $\mathbf{X}_1^{(e)}, \dots, \mathbf{X}_{n_e}^{(e)}$ on $E = ([-1, 1] \times \{0\}) \cup (\{-1, 1\} \times [0, T])$, distributed according to μ_E, together with n_r i.i.d. samples $\mathbf{X}_1^{(r)}, \dots, \mathbf{X}_{n_r}^{(r)}$, uniformly distributed on Ω. Let $(\hat{\theta}(p, n_e, n_r, D))_{p \in \mathbb{N}}$ be a sequence of parameters minimizing the empirical risk function

$$R_{n_e, n_r}(u_\theta) = \frac{\lambda_e}{n_e} \sum_{j=1}^{n_e} |u_\theta(\mathbf{X}_j^{(e)}) - h(\mathbf{X}_j^{(e)})|^2 + \frac{1}{n_r} \sum_{\ell=1}^{n_r} \mathscr{F}(u_\theta, \mathbf{X}_\ell^{(r)})^2,$$

over the space $\mathrm{NN}_H(D)$. The theoretical counterpart of this empirical risk is

$$\mathscr{R}(u) = \lambda_e \mathbb{E}|u(\mathbf{X}^{(e)}) - h(\mathbf{X}^{(e)})|^2 + \frac{1}{|\Omega|} \int_\Omega \mathscr{F}(u, \mathbf{x})^2 d\mathbf{x}.$$

PROPOSITION 3.2 (PDE solver overfitting). *Consider the heat propagation model (4). Then, whenever $D \geqslant 4$, for any pair (n_e, n_r), for all $\mathbf{X}_1^{(e)}, \dots, \mathbf{X}_{n_e}^{(e)}$ and for all $\mathbf{X}_1^{(r)}, \dots, \mathbf{X}_{n_r}^{(r)}$, there exists a minimizing sequence $(u_{\hat{\theta}(p, n_e, n_r, D)})_{p \in \mathbb{N}} \in \mathrm{NN}_H(D)^{\mathbb{N}}$ such that $\lim_{p \to \infty} R_{n_e, n_r}(u_{\hat{\theta}(p, n_e, n_r, D)}) = 0$ but $\lim_{p \to \infty} \mathscr{R}(u_{\hat{\theta}(p, n_e, n_r, D)}) = \infty$. So, this PINN estimator is not consistent.*

Figure 2 (left) shows an example of an inconsistent PINN estimator. Such an estimator corresponds to a function that equals zero on Ω (and thus satisfies the linear PDE), while satisfying the initial condition on $\partial\Omega$. This function corresponds to a limit of neural networks u_θ such that $\|\theta\|_2 \to \infty$.

The proof strategy of Propositions 3.1 and 3.2 does not depend on the geometry of the points $\mathbf{X}^{(r)}$ and the points $\mathbf{X}^{(e)}$, which could therefore be sampled along a grid, or by any quasi Monte Carlo method. We emphasize that the two negative examples of Propositions 3.1 and 3.2 are no exceptions. In fact, their proofs can be easily generalized to differential operators $\mathscr{F}$ such that the following property holds: for all $\mathbf{x} \in \Omega$, for all $u \in C^\infty(\Omega, \mathbb{R}^{d_2})$, if ∇u vanishes on an open set containing $\mathbf{x}$, then $\mathscr{F}(u, \mathbf{x}) = 0$. This property is satisfied in the case of motion with friction, advection, heat, wave propagation, Schrödinger, Maxwell and Navier-Stokes equations, which are so as many cases that will suffer from overfitting.

4. Consistency of regularized PINNs for linear and nonlinear PDE systems. Training PINNs can be tricky because it can lead to the type of pathological situations highlighted in Section 3. To avoid such an overfitting behavior, a standard approach in machine learning is to resort to ridge regularization, where the empirical risk to be minimized is penalized by the L^2 norm of the parameters θ. This technique has been shown to improve not only the optimization convergence during the training phase, but also the generalization ability of the resulting predictor [Krogh and Hertz, 1991, Guo et al., 2017]. Ridge regularization is available in most deep learning libraries (e.g., `pytorch` or `keras`), where it is implemented using the so-called weight decay [Loshchilov and Hutter, 2019]. Interestingly, the ridge regularization of a slight modification of PINNs, using adaptive activation functions, has been studied in Jagtap et al. [2020], which shows that gradient descent algorithms manage to generate an effective minimizing sequence of the penalized empirical risk. In this section, we formalize ridge PINNs and study their risk-consistency.

DEFINITION 4.1 (Ridge PINNs). The ridge risk function is defined by

$$R^{(\mathrm{ridge})}_{n,n_e,n_r}(u_\theta) = R_{n,n_e,n_r}(u_\theta) + \lambda_{(\mathrm{ridge})}\|\theta\|_2^2, \tag{5}$$

where $\lambda_{(\mathrm{ridge})} > 0$ is the ridge hyperparameter. We denote by $(\hat{\theta}^{(\mathrm{ridge})}_{(p,n_e,n_r,D)})_{p\in\mathbb{N}}$ a minimizing sequence of this risk, i.e.,

$$\lim_{p\to\infty} R^{(\mathrm{ridge})}_{n,n_e,n_r}(u_{\hat{\theta}^{(\mathrm{ridge})}_{(p,n_e,n_r,D)}}) = \inf_{\theta\in\Theta} R^{(\mathrm{ridge})}_{n,n_e,n_r}(u_\theta).$$

Our next Proposition 4.2 states that the L^2 norm of the parameters θ bounds the Hölder norm of the neural network u_θ. This result is interesting in itself because it establishes a connection between the L^2 norm of a fully connected neural network and its regularity. In the present paper it plays a key role in the risk-consistency analysis.

PROPOSITION 4.2 (Bounding the norm of a neural network by the norm of its parameter). *Consider the class* $\mathrm{NN}_H(D) = \{u_\theta, \theta \in \Theta_{H,D}\}$. *Let* $K \in \mathbb{N}$. *Then there exists a constant* $C_{K,H} > 0$, *depending only on* K *and* H, *such that, for all* $\theta \in \Theta_{H,D}$,

$$\|u_\theta\|_{C^K(\mathbb{R}^{d_1})} \leqslant C_{K,H}(D+1)^{HK+1}(1+\|\theta\|_2)^{HK}\|\theta\|_2.$$

Moreover, this bound is tight with respect to $\|\theta\|_2$, *in the sense that, for all* $H, D \geqslant 1$ *and all* $K \in \mathbb{N}$, *there exists a sequence* $(\theta_p)_{p\in\mathbb{N}} \in \mathrm{NN}_H(D)$ *and a constant* $\bar{C}_{K,H} > 0$ *such that* (i) $\lim_{p\to\infty}\|\theta_p\|_2 = \infty$ *and* (ii) $\|u_{\theta_p}\|_{C^K(\mathbb{R}^{d_1})} \geqslant \bar{C}_{K,H}\|\theta_p\|_2^{HK+1}$.

In order to study the generalization capabilities of regularized PINNs, we need to restrict the PDEs to a class of smooth differential operators, which we call polynomial operators (Definition 4.4 below). This class includes the most common PDE systems, as shown in the following example with the Navier-Stokes equations.

EXAMPLE 4.3 (Navier-Stokes equations). Let $\Omega = \Omega_1 \times]0, T[$, where $\Omega_1 \subseteq \mathbb{R}^3$ is a bounded Lipschitz domain and $T \geqslant 0$ is a finite time horizon. The incompressible Navier-Stokes system of equations is defined for all $u = (u_x, u_y, u_z, p) \in C^2(\bar{\Omega}, \mathbb{R}^4)$ by

$$\forall \mathbf{x} = (x, y, z, t) \in \Omega, \begin{cases} \mathscr{F}_1(u, \mathbf{x}) = \partial_t u_x - u_x \partial_x u_x - \eta \partial^2_{x,x} u_x + \rho^{-1} \partial_x p \\ \mathscr{F}_2(u, \mathbf{x}) = \partial_t u_y - u_y \partial_y u_y - \eta \partial^2_{y,y} u_y + \rho^{-1} \partial_y p \\ \mathscr{F}_3(u, \mathbf{x}) = \partial_t u_z - u_z \partial_z u_z - \eta \partial^2_{z,z} u_z + \rho^{-1} \partial_z p + g(\mathbf{x}) \\ \mathscr{F}_4(u, \mathbf{x}) = \partial_x u_x + \partial_y u_y + \partial_z u_z, \end{cases}$$

where $\eta, \rho > 0$ and $g \in C^\infty(\bar{\Omega}, \mathbb{R})$. Observe that $\mathscr{F}_1, \mathscr{F}_2, \mathscr{F}_3$, and $\mathscr{F}_4$ are polynomials in u and its derivatives, with coefficients in $C^\infty(\bar{\Omega}, \mathbb{R})$. For example, $\mathscr{F}_3(u, \mathbf{x}) = P_3(u_z, \partial_z u_z, \partial_t u_z, \partial^2_{z,z} u_z, \partial_z p)(\mathbf{x})$, where the polynomial $P_3 \in C^\infty(\bar{\Omega}, \mathbb{R})[Z_1, Z_2, Z_3, Z_4]$ is defined by $P_3(Z_1, Z_2, Z_3, Z_4) = Z_3 - Z_1 Z_2 - \eta Z_4 + \rho^{-1} Z_5 + g$. □

The above example can be generalized with the following definition.

DEFINITION 4.4 (Polynomial operator). An operator $\mathscr{F} : C^K(\bar{\Omega}, \mathbb{R}^{d_2}) \times \Omega \to \mathbb{R}$ is a polynomial operator of order $K \in \mathbb{N}$ if there exists an integer $s \in \mathbb{N}$ and multi-indexes $(\alpha_{i,j})_{1 \leqslant i \leqslant d_2, 1 \leqslant j \leqslant s} \in (\mathbb{N}^{d_1})^{s d_2}$ such that

$$\forall u = (u_1, \ldots, u_{d_2}) \in C^K(\bar{\Omega}, \mathbb{R}^{d_2}), \quad \mathscr{F}(u, \cdot) = P((\partial^{\alpha_{i,j}} u_i)_{1 \leqslant i \leqslant d_2, 1 \leqslant j \leqslant s}),$$

where $P \in C^\infty(\bar{\Omega}, \mathbb{R})[Z_{1,1}, \ldots, Z_{d_2,s}]$ is a polynomial with smooth coefficients.

In other words, $\mathscr{F}$ is a polynomial operator if it is of the form

$$\mathscr{F}(u, \mathbf{x}) = \sum_{k=1}^{N(P)} \phi_k \times \prod_{i=1}^{d_2} \prod_{j=1}^{s} (\partial^{\alpha_{i,j}} u_i(\mathbf{x}))^{I(i,j,k)},$$

where $N(P) \in \mathbb{N}^\star$, $\phi_k \in C^\infty(\bar{\Omega}, \mathbb{R})$, and $I(i, j, k) \in \mathbb{N}$. The associated polynomial is $P(Z_{1,1}, \ldots, Z_{d_2,s}) = \sum_{k=1}^{N(P)} \phi_k \times \prod_{i=1}^{d_2} \prod_{j=1}^{s} Z_{i,j}^{I(i,j,k)}$ (recall that $\partial^\alpha u_i = u_i$ when $\alpha = 0$).

DEFINITION 4.5 (Degree). The degree of the polynomial operator $\mathscr{F}$ is

$$\deg(\mathscr{F}) = \max_{1 \leqslant k \leqslant N(P)} \sum_{i=1}^{d_2} \sum_{j=1}^{s} (1 + |\alpha_{i,j}|) I(i, j, k).$$

As an illustration, in Example 4.3, one has $\deg(\mathscr{F}_3) = 3$, and this degree is reached in both the terms $u_z \partial_z u_z$ and $\partial^2_{z,z} u_z$. Note that $\deg(P_3) = 2$ but $\deg(\mathscr{F}_3) = 3$. To compute $\deg(\mathscr{F}_3)$, we first count the number of terms in each monomial ($u_z \partial_z u_z$ has two terms while $\partial^2_{z,z} u_z$ has one term), which is $\sum_{i=1}^{d_2} \sum_{j=1}^{s} I(i, j, k)$ for the kth monomial, and add the number of derivatives involved in the product ($u_z \partial_z u_z$ contains a single ∂_z operator while $\partial^2_{z,z} u_z$ contains two derivatives in ∂_z), which corresponds to $\sum_{i=1}^{d_2} \sum_{j=1}^{s} |\alpha_{i,j}| I(i, j, k)$ for the kth monomial. Thus, for each monomial k, the total sum is $\sum_{i=1}^{d_2} \sum_{j=1}^{s} (1 + |\alpha_{i,j}|) I(i, j, k)$.

We emphasize that this class includes a large number of PDEs, such as linear PDEs (e.g., advection, heat, and Maxwell equations), as well as some nonlinear PDEs (e.g., Blasius, Burger's, and Navier-Stokes equations). Proposition 4.2 is a key ingredient to uniformly bound the risk of PINNs involving polynomial PDE operators (see Appendix F). This in turn can be used to establish the risk-consistency of these PINNs when n_e and n_r tend to ∞, as follows.

THEOREM 4.6 (Risk-consistency of ridge PINNs). *Consider the ridge PINN problem (5), over the class* $\mathrm{NN}_H(D) = \{u_\theta, \theta \in \Theta_{H,D}\}$, *where* $H \geqslant 2$. *Assume that the condition function* h *is Lipschitz and that* $\mathscr{F}_1, \dots, \mathscr{F}_M$ *are polynomial operators. Assume, in addition, that the ridge parameter is of the form*

$$\lambda_{(\mathrm{ridge})} = \min(n_e, n_r)^{-\kappa}, \quad \textit{where} \quad \kappa = \frac{1}{12 + 4H(1 + (2+H)\max_k \deg(\mathscr{F}_k))}.$$

Then, almost surely,

$$\lim_{n_e, n_r \to \infty} \lim_{p \to \infty} \mathscr{R}_n(u_{\hat{\theta}^{(\mathrm{ridge})}(p, n_e, n_r, D)}) = \inf_{u \in \mathrm{NN}_H(D)} \mathscr{R}_n(u).$$

Thus, minimizing the ridge empirical risk (5) over $\Theta_{H,D}$ amounts to minimizing the theoretical risk (2) over $\Theta_{H,D}$ in the asymptotic regime $n_e, n_r \to \infty$. This fundamental result is complemented by the following one, which resorts to another asymptotics in the width D. This ensures that the choice of the neural architecture $\mathrm{NN}_H \subseteq C^\infty(\bar{\Omega}, \mathbb{R}^{d_2})$ does not introduce any asymptotic bias.

THEOREM 4.7 (The ridge PINN is asymptotically unbiased). *Under the same assumptions as in Theorem 4.6, one has, almost surely,*

$$\lim_{D \to \infty} \lim_{n_e, n_r \to \infty} \lim_{p \to \infty} \mathscr{R}_n(u_{\hat{\theta}^{(\mathrm{ridge})}(p, n_e, n_r, D)}) = \inf_{u \in C^\infty(\bar{\Omega}, \mathbb{R}^{d_2})} \mathscr{R}_n(u).$$

In other words, minimizing the ridge empirical risk over $\Theta_{H,D}$ and letting $D, n_e, n_r \to \infty$ amounts to minimizing the theoretical risk (2) over the entire class $C^\infty(\bar{\Omega}, \mathbb{R}^{d_2})$. We emphasize that these two theorems hold independently of the values of the hyperparameters $\lambda_d, \lambda_e \geqslant 0$. Therefore, our results cover the general hybrid modeling framework (1), which includes the PDE solver. To the best of our knowledge, these are the first results that provide theoretical guarantees for PINNs regularized with a standard penalty. They complement the state-of-the-art approaches of Shin [2020], Shin et al. [2020], Mishra and Molinaro [2023], and Wu et al. [2022], which consider regularization strategies that are unfortunately not feasible in practice.

It is worth noting that Theorem 4.7 still holds by choosing D as a function of n_e and n_r. In fact, an easy modification of the proofs reveals that one can take $D(n_e, n_r) = \min(n_e, n_r)^\xi$, where ξ is a constant ξ depending only on H and $\max_k \deg(\mathscr{F}_k)$. Thus, in this setting,

$$\lim_{n_e, n_r \to \infty} \lim_{p \to \infty} \mathscr{R}_n(u_{\hat{\theta}^{(\mathrm{ridge})}(p, n_e, n_r, D(n_e, n_r))}) = \inf_{u \in C^\infty(\bar{\Omega}, \mathbb{R}^{d_2})} \mathscr{R}_n(u).$$

REMARK 4.8 (Dirichlet boundary conditions). Theorems 4.6 and 4.7 can be easily adapted to PINNs with Von Neumann conditions instead of Dirichlet boundary conditions. This is achieved by substituting the term $n_e^{-1} \sum_{j=1}^{n_e} \|u_\theta(\mathbf{X}_j^{(e)}) - h(\mathbf{X}_j^{(e)})\|_2^2$ in the PINN definition (1) by $n_e^{-1} \sum_{j=1}^{n_e} \|\partial_{\overrightarrow{n}} u_\theta(\mathbf{X}_j^{(e)})\|_2^2$, where $\overrightarrow{n}$ is the normal to $\partial\Omega$.

Practical considerations. The decay rate of $\lambda_{(\mathrm{ridge})} = \min(n_e, n_r)^{-\kappa}$ does not depend on the dimension d_1 of Ω. This is consistent with the results of Karniadakis et al. [2021] and De Ryck and Mishra [2022], which suggest that PINNs can overcome the curse of dimensionality, opening up interesting perspectives for efficient solvers of high-dimensional PDEs. We also emphasize that $\lambda_{(\mathrm{ridge})}$ depends only on the degree of the polynomial PDE operator, the depth H, and the sample sizes n_e and n_r. All these quantities are known, which makes this hyperparameter immediately useful for practical applications. For example, in Navier-Stokes equations of Example 4.3, one has $\max_k \deg(\mathscr{F}_k) = 3$. Thus, for a neural network

of depth, say $H = 2$, the ridge hyperparameter $\lambda_{(\text{ridge})} = \min(n_e, n_r)^{-1/116}$ is sufficient to ensure consistency. It is also interesting to note that the bound on $\lambda_{(\text{ridge})}$ in the theorems deteriorates with increasing depth H. This confirms the preferential use of shallow neural networks in the experimental works of Arzani et al. [2021], Karniadakis et al. [2021], and Xu et al. [2021]. The bound also deteriorates as $\max_k \deg \mathscr{F}_k$ increases. This is in line with the empirical results of Davini et al. [2021], which was able to improve the performance of PINNs by reformulating their polynomial differential equation of degree 3 as a system of two polynomial differential equations of degree 2.

It is also interesting to note that Theorems 4.6 and 4.7 hold for any ridge hyperparameter $\lambda_{(\text{ridge})} \geqslant \min(n_e, n_r)^{-\kappa}$ such that $\lim_{n_e,n_r\to\infty} \lambda_{(\text{ridge})} = 0$. However, if n_e and n_r are fixed, choosing too large a $\lambda_{(\text{ridge})}$ will lead to a bias toward parameters of $\Theta_{H,D}$ with a low L^2 norm. Therefore, there is a trade-off between taking $\lambda_{(\text{ridge})}$ as small as possible to reduce this bias, but large enough to avoid overfitting, as illustrated in Section 3. Moreover, our choice of $\lambda_{(\text{ridge})}$ may be suboptimal, since these results rely on inequalities involving a general class of polynomial operators. When studying a particular PDE, the consistency results of Theorems 4.6 and 4.7 should eventually hold with a smaller $\lambda_{(\text{ridge})}$. To tune $\lambda_{(\text{ridge})}$ in practice, one could, for example, monitor the overfitting gap $\text{OG}_{n,n_e,n_r} = |R_{n,n_e,n_r} - \mathscr{R}_n|$ for a ridge estimator $\hat{\theta}^{(\text{ridge})}(p, n_e, n_r, D)$, by standard validation strategy (e.g., by sampling $\tilde{n}_r$ and $\tilde{n}_e$ new points to estimate $\mathscr{R}_n(u_{\hat{\theta}^{(\text{ridge})}(p,n_e,n_r,D)})$ at a $\min(\tilde{n}_r, \tilde{n}_e)^{-1/2}$-rate given by the central limit theorem), and then choose the smallest parameter $\lambda_{(\text{ridge})}$ to introduce as little bias as possible. More information about the relevance of OG_{n,n_e,n_r} is given in Appendix C.

5. Strong convergence of PINNs for linear PDE systems. Beyond risk-consistency concerns, the ultimate goal of PINNs is to learn a physics-informed regression function $u^\star$, or, in the PDE solver setting, to strongly approximate the unique solution $u^\star$ of a PDE system. Thus, what we want is to have guarantees regarding the convergence of $u_{\hat{\theta}^{(\text{ridge})}(p,n_e,n_r,D)}$ to $u^\star$ for an adapted norm. This requirement is called strong convergence in the functional analysis literature. This is however not guaranteed under the sole convergence of the theoretical risk $(\mathscr{R}_n(u_{\hat{\theta}^{(\text{ridge})}(p,n_e,n_r,D)}))_{p,n_e,n_r,D\in\mathbb{N}}$, as shown in the following two examples.

EXAMPLE 5.1 (Lack of data incorporation in the hybrid modeling setting). Suppose $M = 1$, $d_1 = 2$, $d_2 = 1$, $\Omega =]0,1[\times]0,T[$, $h(x,0) = 1$ and $h(0,t) = 1$, and let $\mathscr{F}(u, \mathbf{x}) = \partial_x u(\mathbf{x}) + \partial_t u(\mathbf{x})$. This corresponds to the assumption that the solution should approximately follow the advection equation and that it should be close to 1. For any $\delta > 0$, let the sequence $(u_{\delta,p})_{p\in\mathbb{N}} \in \text{NN}_H(2n)^{\mathbb{N}}$ be defined by

$$u_{\delta,p}(x,t) = 1 + \sum_{i=1}^{n} \frac{Y_i}{2}\big(\tanh_p^{\circ H}(x - t - x_i + t_i + \delta) - \tanh_p^{\circ H}(x - t - x_i + t_i - \delta)\big),$$

where $\tanh_p := \tanh(p\,\cdot)$, and $\mathbf{X}_i = (x_i, t_i)$. Then, as soon as $\delta \leqslant \frac{1}{2}\min_{i\neq j} |x_i - x_j + t_j - t_i|$, $\lim_{p\to\infty} \mathscr{R}_n(u_{\delta,p}) = 0$. However, the limit of $u_{\delta,p}$ in $L^2(\Omega)$ as $p \to \infty$ and $\delta \to 0$ equals 1, independently of n and the function $u^\star$. □

Consequently, no matter how large the number n of data samples, the PINN solution of Example 5.1 is always 1 in $L^2(\Omega)$ and thus fails to learn $u^\star$. Note that the pathological sequence $(u_{\delta,p})_{p\in\mathbb{N}}$ satisfies that, for all $\delta > 0$, $\lim_{p\to\infty} \|\nabla u_{\delta,p}\|_{L^2(\Omega)} = \infty$.

In the PDE solver setting, one can consider the a priori favorable case where the PDE system admits a unique (strong) solution $u^\star$ in $C^K(\bar{\Omega}, \mathbb{R}^{d_2})$ (where K is the maximum order of the differential operators $\mathscr{F}_1, \ldots, \mathscr{F}_M$). Note that $u^\star$ is the unique minimizer of $\mathscr{R}$ over $C^K(\bar{\Omega}, \mathbb{R}^{d_2})$, with $\mathscr{R}(u^\star) = 0$ (and $\mathscr{R}(u) = 0$ if and only if u satisfies the initial conditions,

the boundary conditions, and the system of differential equations). However, we describe below a situation where a minimizing sequence of $\mathscr{R}$ does not converge to the unique strong solution $u^\star$ of the PDE in question.

EXAMPLE 5.2 (Divergence in the PDE solver setting). Suppose $M = 1$, $d_1 = d_2 = 1$, $\Omega =]-1, 1[$, $h(1) = 1$, $\lambda_e > 0$, and let the polynomial operator be $\mathscr{F}(u, \mathbf{x}) = \mathbf{x}u'(\mathbf{x})$. Clearly, $u^\star(\mathbf{x}) = 1$ is the only strong solution of the PDE $\mathbf{x}u'(\mathbf{x}) = 0$ with $u(1) = 1$. Let the sequence $(u_p)_{p\in\mathbb{N}} \in \mathrm{NN}_H(D)^{\mathbb{N}}$ be defined by $u_p = \tanh_p \circ \tanh^{\circ(H-1)}$. According to Appendix C, $\lim_{p\to\infty} \mathscr{R}(u_p) = \mathscr{R}(u^\star) = 0$. However, the minimizing sequence $(u_p)_{p\in\mathbb{N}}$ does not converge to $u^\star$, since $u_\infty(\mathbf{x}) := \lim_{p\to\infty} u_p(\mathbf{x}) = \mathbf{1}_{\mathbf{x}>0} - \mathbf{1}_{\mathbf{x}<0}$. □

We have therefore exhibited a sequence $(u_p)_{p\in\mathbb{N}}$ of neural networks that minimizes $\mathscr{R}$ and such that $(u_p)_{p\in\mathbb{N}}$ converges pointwise. However, its limit u_∞ is not the unique strong solution of the PDE. In fact, u_∞ is not differentiable at 0, which is incompatible with the differential operators $\mathscr{F}$ used in $\mathscr{R}(u_\infty)$. Interestingly, the Cauchy-Schwarz inequality states that the pathological sequence $(u_p)_{p\in\mathbb{N}}$ satisfies $\lim_{p\to\infty} \|u_p'\|^2_{L^2(\Omega)} = \infty$, as in Example 5.1.

5.1. *Sobolev regularization.* The two examples above illustrate how the convergence of the theoretical risk $\mathscr{R}_n(u_{\hat{\theta}^{(\mathrm{ridge})}(p,n_e,n_r,D)})$ to $\inf_{u\in C^\infty(\bar{\Omega},\mathbb{R}^{d_2})} \mathscr{R}_n(u)$ (for any n) is not sufficient to guarantee the strong convergence to a PDE or hybrid modeling solution. To ensure such a convergence, a different analysis is needed, mobilizing tools from functional analysis. In the sequel, we build upon the regression estimation penalized by PDEs of Azzimonti et al. [2015], Sangalli [2021], Arnone et al. [2022], and Ferraccioli et al. [2022], and make use of the calculus of variations [e.g., Evans, 2010, Theorems 1-4, Chapter 8]. In the former references, the minimizer of $\mathscr{R}_n$ does not satisfy the PDE system injected in the PINN penalty, but another PDE system, known as the Euler-Lagrange equations. Although interesting, the mathematical framework is different from ours. First, the authors do not study the convergence of neural networks, but rather methods in which the boundary conditions are hard-coded, such as the finite element method. Second, these frameworks are limited to special cases of theoretical risks. Indeed, only second-order PDEs with $\lambda_e = \infty$ are considered in Azzimonti et al. [2015], while Evans [2010] deal with first-order PDEs, echoing the case of $\lambda_d = 0$ and $\lambda_e = \infty$.

It is worthwhile mentioning that the results of Azzimonti et al. [2015] rely on an important property of the theoretical risk function $\mathscr{R}_n$, called coercivity. This is a common assumption of the calculus of variations [Evans, 2010]. The operator $\mathscr{R}_n$ is said to be coercive if there exist $K \in \mathbb{N}$ and $\lambda_t > 0$ such that, for all $u \in H^K(\Omega, \mathbb{R}^{d_2})$, $\mathscr{R}_n(u) \geqslant \lambda_t \|u\|^2_{H^K(\Omega)}$ (the notation $H^K(\Omega, \mathbb{R}^{d_2})$ stands for the usual Sobolev space of order K—see Appendix A. It turns out that the failures of Examples 5.1 and 5.2 are due to a lack of coercivity, since, in both cases, $\lim_{p\to\infty} \|u_p\|_{H^1(\Omega)} = \infty$ but $\lim_{p\to\infty} \mathscr{R}_n(u_p) \leqslant \mathscr{R}_n(u^\star)$. There are two ways to correct this problem: either one can restrict the study to coercive operators only, or one can resort to an explicit regularization of the risk to enforce its coercivity. We choose the latter, since most PDEs used in the practice of PINNs are not coercive.

In the following, we restrict ourselves to affine operators, which exactly correspond to linear PDE systems, including the advection, heat, wave, and Maxwell equations.

DEFINITION 5.3 (Affine operator). The operator $\mathscr{F}$ is affine of order K if there exists $A_\alpha \in C^\infty(\bar{\Omega}, \mathbb{R}^{d_2})$ and $B \in C^\infty(\bar{\Omega}, \mathbb{R})$ such that, for all $\mathbf{x} \in \Omega$ and all $u \in H^K(\Omega, \mathbb{R}^{d_2})$,

$$\mathscr{F}(u, \mathbf{x}) = \mathscr{F}^{(\mathrm{lin})}(u, \mathbf{x}) + B(\mathbf{x}),$$

where $\mathscr{F}^{(\mathrm{lin})}(u, \mathbf{x}) = \sum_{|\alpha|\leqslant K} \langle A_\alpha(\mathbf{x}), \partial^\alpha u(\mathbf{x})\rangle$ is linear.

The source term B is important, as it makes it possible to model a large variety of applied physical problems, as illustrated in Song et al. [2021]. Note also that affine operators of order K are in fact polynomial operators of degree $K+1$ (Definitions 4.4 and 4.5) that are extended from smooth functions to the whole Sobolev space $H^K(\Omega, \mathbb{R}^{d_2})$.

DEFINITION 5.4 (Regularized PINNs). The regularized theoretical risk function is

$$\mathscr{R}_n^{(\mathrm{reg})}(u) = \mathscr{R}_n(u) + \lambda_t \|u\|^2_{H^{m+1}(\Omega)}, \tag{6}$$

where $\mathscr{R}_n$ is the original theoretical risk as defined in (2), and $m \in \mathbb{N}$. The corresponding regularized empirical risk function is

$$R^{(\mathrm{reg})}_{n,n_e,n_r}(u_\theta) = R_{n,n_e,n_r}(u_\theta) + \lambda_{(\mathrm{ridge})}\|\theta\|_2^2 + \frac{\lambda_t}{n_\ell}\sum_{\ell=1}^{n_\ell}\sum_{|\alpha|\leqslant m+1} \|\partial^\alpha u_\theta(\mathbf{X}_\ell^{(r)})\|_2^2.$$

It is noteworthy that $R^{(\mathrm{reg})}_{n,n_e,n_r}$ can be straightforwardly implemented in the usual PINN framework and benefit from the computational scalability of the backpropagation algorithm, by encoding the regularization as supplementary PDE-type constraints $\mathscr{F}_\alpha(u, \mathbf{x}) = \partial^\alpha u(\mathbf{x}) = 0$. Note also that the Sobolev regularization has been shown to avoid overfitting in machine learning, yet in different contexts [e.g., Fischer and Steinwart, 2020].

The following proposition shows that the unique mimimizer of (6) can be interpreted as the unique minimizer of an optimization problem involving a weak formulation of the differential terms included in the risk. Its proof is based on the Lax-Milgram theorem [e.g., Brezis, 2010, Corollary 5.8].

PROPOSITION 5.5 (Characterization of the unique minimizer of $\mathscr{R}_n^{(\mathrm{reg})}$). *Assume that $\mathscr{F}_1, \ldots, \mathscr{F}_M$ are affine operators of order K. Assume, in addition, that $\lambda_t > 0$ and $m \geqslant \max(\lfloor d_1/2 \rfloor, K)$. Then the regularized theoretical risk $\mathscr{R}_n^{(\mathrm{reg})}$ has a unique minimizer $\hat{u}_n$ over $H^{m+1}(\Omega, \mathbb{R}^{d_2})$. This minimizer $\hat{u}_n$ is the unique element of $H^{m+1}(\Omega, \mathbb{R}^{d_2})$ that satisfies*

$$\forall v \in H^{m+1}(\Omega, \mathbb{R}^{d_2}), \quad \mathcal{A}_n(\hat{u}_n, v) = \mathcal{B}_n(v),$$

where

$$\begin{aligned}
\mathcal{A}_n(\hat{u}_n, v) = {} & \frac{\lambda_d}{n}\sum_{i=1}^n \langle \tilde{\Pi}(\hat{u}_n)(\mathbf{X}_i), \tilde{\Pi}(v)(\mathbf{X}_i)\rangle + \lambda_e \mathbb{E}\langle \tilde{\Pi}(\hat{u}_n)(\mathbf{X}^{(e)}), \tilde{\Pi}(v)(\mathbf{X}^{(e)})\rangle \\
& + \frac{1}{|\Omega|}\sum_{k=1}^M \int_\Omega \mathscr{F}_k^{(\mathrm{lin})}(\hat{u}_n, \mathbf{x}) \mathscr{F}_k^{(\mathrm{lin})}(v, \mathbf{x}) d\mathbf{x} \\
& + \frac{\lambda_t}{|\Omega|}\sum_{|\alpha|\leqslant m+1} \int_\Omega \langle \partial^\alpha \hat{u}_n(\mathbf{x}), \partial^\alpha v(\mathbf{x})\rangle d\mathbf{x}, \\
\mathcal{B}_n(v) = {} & \frac{\lambda_d}{n}\sum_{i=1}^n \langle Y_i, \tilde{\Pi}(v)(\mathbf{X}_i)\rangle + \lambda_e \mathbb{E}\langle \tilde{\Pi}(v)(\mathbf{X}^{(e)}), h(\mathbf{X}^{(e)})\rangle \\
& - \frac{1}{|\Omega|}\sum_{k=1}^M \int_\Omega B_k(\mathbf{x}) \mathscr{F}_k^{(\mathrm{lin})}(v, \mathbf{x}) d\mathbf{x},
\end{aligned}$$

and where $\tilde{\Pi} : H^{m+1}(\Omega, \mathbb{R}^{d_2}) \to C^0(\Omega, \mathbb{R}^{d_2})$ is the so-called Sobolev embedding, such that $\tilde{\Pi}(u)$ is the unique continuous function that coincides with u almost everywhere.

The Sobolev embedding $\tilde{\Pi}$ is essential in order to give a precise meaning to the pointwise evaluation at the points $\mathbf{X}_i$ of a function $u \in H^{m+1}(\Omega, \mathbb{R}^{d_2}) \subseteq L^2(\Omega, \mathbb{R}^{d_2})$, which is defined only almost everywhere. The rationale behind Proposition 5.5 is that

$$\mathscr{R}_n^{(\mathrm{reg})}(u) = \mathcal{A}_n(u,u) - 2\mathcal{B}_n(u) + \frac{\lambda_d}{n}\sum_{i=1}^{n}\|Y_i\|^2 + \lambda_e \mathbb{E}\|h(\mathbf{X}^{(e)})\|_2^2 + \frac{1}{|\Omega|}\sum_{k=1}^{M}\int_\Omega B_k(\mathbf{x})^2 d\mathbf{x}.$$

Therefore, minimizing $\mathscr{R}_n^{(\mathrm{reg})}$ amounts to minimizing $\mathcal{A}_n - 2\mathcal{B}_n$. It is also interesting to note that the weak formulation $\mathcal{A}_n(\hat{u}, v) = \mathcal{B}_n(v)$ can be interpreted as a weak PDE on $H^{m+1}(\Omega, \mathbb{R}^{d_2})$. In particular, if $\hat{u}_n \in H^{2(m+1)}(\Omega, \mathbb{R}^{d_2})$, then one has, almost everywhere,

$$\sum_{k=1}^{M}(\mathscr{F}_k^{(\mathrm{lin})})^* \mathscr{F}_k(\hat{u}_n, \mathbf{x}) + \lambda_t \sum_{|\alpha| \leqslant m+1} (-1)^{|\alpha|}(\partial^\alpha)^2 \hat{u}_n(\mathbf{x}) = 0.$$

$(\mathscr{F}_k^{(\mathrm{lin})})^*$ is the adjoint operator of $\mathscr{F}_k^{(\mathrm{lin})}$ such that, for all $u, v \in C^\infty(\Omega, \mathbb{R})$ with $v|_{\partial\Omega} = 0$,

$$\int_\Omega u \mathscr{F}^{(\mathrm{lin})}(v, \mathbf{x}) d\mathbf{x} = \int_\Omega (\mathscr{F}_k^{(\mathrm{lin})})^*(u, \mathbf{x}) v d\mathbf{x}.$$

Thus, even in the regime $\lambda_t \to 0$ (i.e., when the regularization becomes negligible), the solution of the PINN problem does not satisfy the constraints $\mathscr{F}_k(u, \mathbf{x}) = 0$, but the slightly different ones $\sum_{k=1}^{M}(\mathscr{F}_k^{(\mathrm{lin})})^* \mathscr{F}_k(u, \mathbf{x}) = 0$. (Notice that, in the PDE solver setting, since $u^\star$ satisfies all the constraints, it satisfies in particular the constraint $\sum_{k=1}^{M}(\mathscr{F}_k^{(\mathrm{lin})})^* \mathscr{F}_k(u^\star, \mathbf{x}) = 0$.) For instance, the advection equation constraint $\mathscr{F}(u, \mathbf{x}) = (\partial_x + \partial_t)u(\mathbf{x})$ of Example 5.1 becomes $\mathscr{F}^* \mathscr{F}(u, \mathbf{x}) = -(\partial_x + \partial_t)^2 u(\mathbf{x})$, and the constraint $\mathscr{F}(u, \mathbf{x}) = \mathbf{x}u'(\mathbf{x})$ of Example 5.2 becomes $\mathscr{F}^* \mathscr{F}(u, \mathbf{x}) = -2\mathbf{x}u'(\mathbf{x}) - \mathbf{x}^2 u''(\mathbf{x})$.

Proposition 5.5 shows that the regularization in λ_t is sufficient to make the PINN problem well-posed, i.e., to ensure that the theoretical risk function (6) admits a unique minimizer. The next natural requirement is that the regularized PINN estimator obtained by minimizing the regularized empirical risk function converges to this unique minimizer $\hat{u}_n$. Proposition 5.6 and Theorem 5.7 show that this is true for linear PDE systems.

PROPOSITION 5.6 (From risk-consistency to strong convergence). *Assume that $\lambda_t > 0$ and $m \geqslant \max(\lfloor d_1/2 \rfloor, K)$. Let $(u_p)_{p\in\mathbb{N}} \in C^\infty(\bar{\Omega}, \mathbb{R}^{d_2})$ be a sequence of smooth functions such that $\lim_{p\to\infty} \mathscr{R}_n^{(\mathrm{reg})}(u_p) = \inf_{u \in C^\infty(\bar{\Omega}, \mathbb{R}^{d_2})} \mathscr{R}_n^{(\mathrm{reg})}$. Then $\lim_{p\to\infty} \|u_p - \hat{u}_n\|_{H^m(\Omega)} = 0$, where $\hat{u}_n$ is the unique minimizer of $\mathscr{R}_n^{(\mathrm{reg})}$ over $H^{m+1}(\Omega, \mathbb{R}^{d_2})$.*

The next theorem follows from Theorem 4.7 and Proposition 5.6, by simply observing that the Sobolev regularization is just an ordinary PINN regularization, taking the form of a polynomial operator of degree $(m+2)$.

THEOREM 5.7 (Strong convergence of regularized PINNs). *Assume that $\mathscr{F}_1, \ldots, \mathscr{F}_M$ are affine operators of order K. Assume, in addition, that $\lambda_t > 0$, $m \geqslant \max(\lfloor d_1/2 \rfloor, K)$, and the condition function h is Lipschitz. Let $(\hat{\theta}^{(\mathrm{reg})}(p, n_e, n_r, D))_{p\in\mathbb{N}}$ be a minimizing sequence of the regularized empirical risk function*

$$R_{n,n_e,n_r}^{(\mathrm{reg})}(u_\theta) = R_{n,n_e,n_r}(u_\theta) + \lambda_{(\mathrm{ridge})}\|\theta\|_2^2 + \frac{\lambda_t}{n_\ell}\sum_{\ell=1}^{n_\ell}\sum_{|\alpha|\leqslant m+1}\|\partial^\alpha u_\theta(\mathbf{X}_\ell^{(r)})\|_2^2$$

over the class $\mathrm{NN}_H(D) = \{u_\theta, \theta \in \Theta_{H,D}\}$*, where* $H \geqslant 2$*. Then, with the choice*

$$\lambda_{(\mathrm{ridge})} = \min(n_e, n_r)^{-\kappa}, \quad \textit{where} \quad \kappa = \frac{1}{12 + 4H(1 + (2+H)(m+2))},$$

one has, almost surely,

$$\lim_{D\to\infty} \lim_{n_e,n_r\to\infty} \lim_{p\to\infty} \|u_{\hat{\theta}^{(\mathrm{reg})}(p,n_e,n_r,D)} - \hat{u}_n\|_{H^m(\Omega)} = 0,$$

where $\hat{u}_n$ *is the unique minimizer of* $\mathscr{R}_n^{(\mathrm{reg})}$ *over* $H^{m+1}(\Omega, \mathbb{R}^{d_2})$*.*

Theorem 5.7 ensures that the sequence $u_{\hat{\theta}^{(\mathrm{reg})}(p,n_e,n_r,D)}$ of PINNs converges to the unique minimizer $\hat{u}_n$ of the regularized theoretical risk function (6), provided that the ridge hyperparameter $\lambda_{(\mathrm{ridge})}$ vanishes slowly enough. However, it does not provide any information about the proximity between $u_{\hat{\theta}^{(\mathrm{reg})}(p,n_e,n_r,D)}$ and $u^\star$. On the other hand, since the regularized theoretical risk function is a small perturbation of the theoretical risk function (2), it is reasonable to think that its minimizer $\hat{u}_n$ should in some way converge to $u^\star$ as $\lambda_t \to 0$. This is encapsulated in Theorem 5.8 for the PDE solver setting and in Theorem 5.13 for the more general hybrid modeling setting.

5.2. *The PDE solver case.*

THEOREM 5.8 (Strong convergence of linear PDE solvers). *Assume that* $\mathscr{F}_1, \ldots, \mathscr{F}_M$ *are affine operators of order* K*. Consider the PDE solver setting (i.e.,* $\lambda_e > 0$ *and* $\lambda_d = 0$*) and assume that the condition function* h *is Lipschitz. Assume, in addition, that the PDE system admits a unique solution* $u^\star$ *in* $H^{m+1}(\Omega, \mathbb{R}^{d_2})$ *for some* $m \geqslant \max(\lfloor d_1/2 \rfloor, K)$*. Let* $(\hat{\theta}^{(\mathrm{reg})}(p, n_e, n_r, D, \lambda_t))_{p\in\mathbb{N}}$ *be a minimizing sequence of the regularized empirical risk function*

$$R^{(\mathrm{reg})}_{n_e,n_r}(u_\theta) = R_{n_e,n_r}(u_\theta) + \lambda_{(\mathrm{ridge})}\|\theta\|_2^2 + \frac{\lambda_t}{n_\ell}\sum_{\ell=1}^{n_\ell} \sum_{|\alpha|\leqslant m+1} \|\partial^\alpha u_\theta(\mathbf{X}_\ell^{(r)})\|_2^2$$

over the class $\mathrm{NN}_H(D) = \{u_\theta, \theta \in \Theta_{H,D}\}$*, where* $H \geqslant 2$*. Then, with the choice*

$$\lambda_{(\mathrm{ridge})} = \min(n_e, n_r)^{-\kappa}, \quad \textit{where} \quad \kappa = \frac{1}{12 + 4H(1 + (2+H)(m+2))},$$

one has, almost surely,

$$\lim_{\lambda_t\to 0} \lim_{D\to\infty} \lim_{n_e,n_r\to\infty} \lim_{p\to\infty} \|u_{\hat{\theta}^{(\mathrm{reg})}(p,n_e,n_r,D,\lambda_t)} - u^\star\|_{H^m(\Omega)} = 0.$$

Back to Example 5.2, one has $m = 1$. Recall that, in this setting, the unique minimizer of $\mathscr{R}$ over $C^0([-1,1],\mathbb{R})$ is $u^\star(\mathbf{x}) = 1$, satisfying $u^\star \in H^2(]-1,1[,\mathbb{R})$. Therefore, by letting $\lambda_t \to 0$, this theorem shows that any sequence minimizing the regularized empirical risk function converges, with respect to the $H^2(\Omega)$ norm, to the unique strong solution $u^\star$ of the PDE $\mathbf{x}u'(\mathbf{x}) = 0$ and $u(1) = 1$.

REMARK 5.9 (Dimensionless hyperparameters and lower regularity assumptions on $u^\star$). The condition $m \geqslant \lfloor d_1/2 \rfloor$ in Theorem 5.7 is necessary to make the pointwise evaluations $\tilde{\Pi}(u)(\mathbf{X}_i)$ continuous. This condition does have an impact on $\lambda_{(\mathrm{ridge})}$, which grows exponentially fast with the dimension d_1. However, in the PDE solver setting, it is possible to get rid of this dimension problem, taking $m = \max_k \deg(\mathscr{F}_k)$. To see this, just note that

there is no $\mathbf{X}_i$, and so there is no need to resort to the $\tilde{\Pi}(u)(\mathbf{X}_i)$. Indeed, the proof of Theorem 5.8 can be adapted by replacing the Sobolev inequalities in the proofs of Theorem 5.7 by the trace theorem for Lipschitz domains [e.g., Grisvard, 2011, Theorem 1.5.1.10]. In this case, it is enough to assume that $u^\star \in H^{K+1}(\Omega, \mathbb{R}^{d_2})$, which is less restrictive than $u^\star \in H^{\max(\lfloor d_1/2 \rfloor, K)+1}(\Omega, \mathbb{R}^{d_2})$. However, this comes at the price of assuming that μ_E admits a density with respect to the hypersurface measure on $\partial\Omega$ (as it is often the case in practice).

5.3. *The hybrid modeling case.* To apply Theorem 5.7 to the general hybrid modeling setting, it is necessary to measure the gap between $u^\star$ and the model specified by the constraints $\mathscr{F}_1, \dots, \mathscr{F}_M$ and the condition function h. This is encapsulated in the next definition.

DEFINITION 5.10 (Physics inconsistency). For any $u \in H^{m+1}(\Omega, \mathbb{R}^{d_2})$, the physics inconsistency of u is defined by

$$\mathrm{PI}(u) = \lambda_e \mathbb{E}\|\tilde{\Pi}(u)(\mathbf{X}^{(e)}) - h(\mathbf{X}^{(e)})\|_2^2 + \frac{1}{|\Omega|}\sum_{k=1}^{M}\int_\Omega \mathscr{F}_k(u, \mathbf{x})^2 d\mathbf{x}.$$

Observe that $\mathscr{R}_n(u) = \frac{\lambda_d}{n}\sum_{i=1}^n \|\tilde{\Pi}(u)(\mathbf{X}_i) - Y_i\|_2^2 + \mathrm{PI}(u)$. In short, the quantity $\mathrm{PI}(u)$ measures how well the initial/boundary conditions, encoded by h, and the PDE system, encoded by the $\mathscr{F}_k$, describe the function u [see also Willard et al., 2023]. In particular, $\mathrm{PI}(u^\star)$ measures the modeling error—the better the model, the lower $\mathrm{PI}(u^\star)$.

PROPOSITION 5.11 (Strong convergence of hybrid modeling). *Assume that the conditions of Theorem 5.7 are satisfied. Then $\hat{u}_n \equiv \hat{u}_n(\mathbf{X}_1, \dots, \mathbf{X}_n, \varepsilon_1, \dots, \varepsilon_n)$ is a random variable such that $\mathbb{E}\|\hat{u}_n\|_{H^{m+1}(\Omega)}^2 < \infty$.*

Suppose, in addition, that $u^\star \in H^{m+1}(\Omega, \mathbb{R}^{d_2})$, that the noise ε is independent from $\mathbf{X}$, and that ε has the same distribution as $-\varepsilon$. Then there exists a constant $C_\Omega > 0$, depending only on Ω, such that

$$\begin{aligned}\mathbb{E}\int_\Omega \|\tilde{\Pi}(\hat{u}_n - u^\star)\|_2^2 d\mu_{\mathbf{X}} \leqslant{}& \frac{1}{\lambda_d}\big(\mathrm{PI}(u^\star) + \lambda_t \|u^\star\|_{H^{m+1}(\Omega)}^2\big) \\ &+ \frac{C_\Omega d_2^{1/2}}{n^{1/2}}\Big(2\|u^\star\|_{H^{m+1}(\Omega)}^2 + \frac{\mathrm{PI}(u^\star)}{\lambda_t}\Big) \\ &+ \frac{8\mathbb{E}\|\varepsilon\|_2^2}{n}\Big(1 + C_\Omega d_2^{3/2}\Big(\frac{\lambda_d}{\lambda_t} + \frac{\lambda_d^2}{\lambda_t^2 n^{1/2}}\Big)\Big).\end{aligned}$$

In particular, with the choice $\lambda_e = 1$, $\lambda_t = (\log n)^{-1}$, and $\lambda_d = n^{1/2}/(\log n)$, one has

$$\mathbb{E}\int_\Omega \|\tilde{\Pi}(\hat{u}_n - u^\star)\|_2^2 d\mu_{\mathbf{X}} \leqslant \frac{\Lambda \log^2(n)}{n^{1/2}},$$

where $\Lambda = 24 d_2^{3/2} C_\Omega(\mathrm{PI}(u^\star) + \|u^\star\|_{H^{m+1}(\Omega)} + \mathbb{E}\|\varepsilon\|_2^2)$.

This (nonasymptotic) proposition provides an insight into the scaling of the PINN hyperparameters. Indeed, the term $\frac{1}{\lambda_d}(\mathrm{PI}(u^\star) + \lambda_t\|u^\star\|_{H^{m+1}(\Omega)})$ encapsulates the modeling error, damped by the weight λ_d. However, λ_d cannot be arbitrarily large because of the term $\frac{8\mathbb{E}\|\varepsilon\|_2^2}{n}\big(1 + C_\Omega d_2^{3/2}\big(\frac{\lambda_d}{\lambda_t} + \frac{\lambda_d^2}{\lambda_t^2 n^{1/2}}\big)\big)$. So, there is a trade-off between the modeling error and the random variation in the data. Note also the other trade-off in the regularization hyperparameter λ_t, which should not converge to 0 too quickly because of the term $\frac{C_\Omega d_2^{1/2}}{n^{1/2}}\big(2\|u^\star\|_{H^{m+1}(\Omega)}^2 + \frac{\mathrm{PI}(u^\star)}{\lambda_t}\big)$.

PROPOSITION 5.12 (Physics consistency of hybrid modeling). *Under the conditions of Proposition 5.11, if* $\lim_{n\to\infty} \frac{\lambda_d^2}{n\lambda_t} = 0$ *and* $\lim_{n\to\infty} \lambda_t = 0$*, one has*

$$\mathbb{E}(\mathrm{PI}(\hat{u}_n)) \leqslant \mathrm{PI}(u^\star) + o_{n\to\infty}(1).$$

(Note that the conditions are satisfied with $\lambda_e = 1$*,* $\lambda_t = (\log n)^{-1}$*, and* $\lambda_d = n^{1/2}/(\log n)$*.)*

As usual, we let $(u^{(n)}_{\hat{\theta}^{(\mathrm{reg})}(p,n_e,n_r,D)})_{p\in\mathbb{N}} \in \mathrm{NN}_H(D)^{\mathbb{N}}$ be a minimizing sequence of $R^{(\mathrm{reg})}_{n,n_e,n_r}$, where the exponent n indicates that the sample size n is kept fixed along the sequence. Since $u^{(n)}_{\hat{\theta}^{(\mathrm{reg})}(p,n_e,n_r,D)} \in C^\infty(\bar{\Omega}, \mathbb{R}^{d_2})$, one has $\tilde{\Pi}(u^{(n)}_{\hat{\theta}^{(\mathrm{reg})}(p,n_e,n_r,D)}) = u^{(n)}_{\hat{\theta}^{(\mathrm{reg})}(p,n_e,n_r,D)}$. Thus, by combining Theorem 5.7 with Propositions 5.11 and 5.12, we obtain the following important theorem.

THEOREM 5.13 (Strong convergence of regularized PINNs). *Under the same assumptions as in Theorem 5.7 and Proposition 5.11, with the choice* $\lambda_e = 1$*,* $\lambda_t = (\log n)^{-1}$*, and* $\lambda_d = n^{1/2}/(\log n)$*, one has*

$$\lim_{D\to\infty} \lim_{n_e,n_r\to\infty} \lim_{p\to\infty} \mathbb{E} \int_\Omega \|u^{(n)}_{\hat{\theta}^{(\mathrm{reg})}(p,n_e,n_r,D)} - u^\star\|_2^2 d\mu_{\mathbf{X}} \leqslant \frac{\Lambda \log^2(n)}{n^{1/2}}$$

and

$$\lim_{D\to\infty} \lim_{n_e,n_r\to\infty} \lim_{p\to\infty} \mathbb{E}(\mathrm{PI}(u^{(n)}_{\hat{\theta}^{(\mathrm{reg})}(p,n_e,n_r,D)})) \leqslant \mathrm{PI}(u^\star) + o_{n\to\infty}(1).$$

The minimax regression rate over any bounded class of functions in $C^{(m+1)}(\Omega, \mathbb{R}^{d_2})$ is known to be $n^{-2(m+1)/(2(m+1)+d_1)}$ [Stone, 1982, Theorem 1]. Theorem 5.13 shows that the regularized PINN estimator achieves the rate $\log(n)/n^{1/2}$ over any *larger* class bounded in $H^{(m+1)}(\Omega, \mathbb{R}^{d_2})$. Thus, the regularized PINN estimator has the optimal rate, up to a log term, in the regime $d_1 \to \infty$ and $m = \lfloor d_1/2 \rfloor$.

Theorem 5.13 shows that a properly regularized PINN estimator is both statistically *and* physics consistent, in the sense that the error $\mathbb{E}\int_\Omega \|u^{(n)}_{\hat{\theta}^{(\mathrm{reg})}(p,n_e,n_r,D)} - u^\star\|_2^2 d\mu_{\mathbf{X}}$ converges to zero with a physics inconsistency $\mathbb{E}(\mathrm{PI}(u^{(n)}_{\hat{\theta}^{(\mathrm{reg})}(p,n_e,n_r,D)}))$ that is asymptotically no larger than $\mathrm{PI}(u^\star)$. It is also worth mentioning that in some applications, the physical measures $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are forced to be sampled in certain subset of Ω. An important application is when Ω is spatio-temporal and one wishes to extrapolate/transfer a model from a training dataset collected on $\mathrm{supp}(\mu_{\mathbf{X}}) = \Omega_1 \times]0, T_{\mathrm{train}}[$ to a test $\Omega_1 \times]T_{\mathrm{train}}, T_{\mathrm{test}}[$, using a temporal evolution PDE system to extrapolate [e.g., Cai et al., 2021]. On the other hand, the physical restriction on the data measurement can be also strictly spatial. This is for example the case in some blood modeling problems, where the blood flow measures can only be taken in a specific region of a blood vessel, as illustrated in Arzani et al. [2021]. Thus, in both these contexts, the support $\mathrm{supp}(\mu_{\mathbf{X}})$ of the distribution $\mu_{\mathbf{X}}$ is strictly contained in Ω. Of course, this is compatible with Theorem 5.13, which shows that the regularized PINN estimator consistently interpolates the function $u^\star$ on $\mathrm{supp}(\mu_{\mathbf{X}})$. Furthermore, Theorem 5.13 shows that the estimator uses the physical model to extrapolate on $\Omega \backslash \mathrm{supp}(\mu_{\mathbf{X}})$. In summary, the better the model, the lower the modeling error $\mathrm{PI}(u^\star)$, and the better the domain adaptation capabilities. This provides an interesting mathematical insight into the relevance of combining data-driven statistical models with the interpretability and extrapolation capabilities of physical modeling.

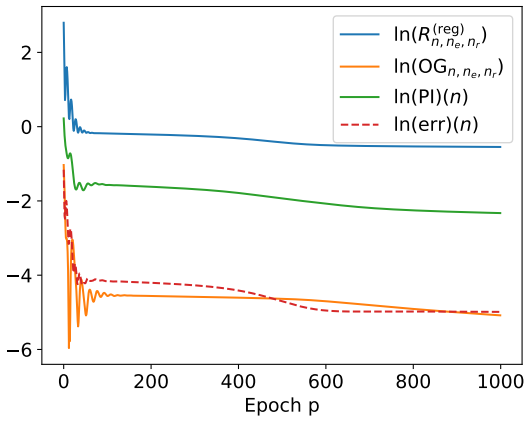

FIG 3. *Regularized empirical risk (blue) and overfitting gap* OG *(orange) with respect to the number* p *of epochs for* $n = 100$. *The physics inconsistency* $\mathrm{PI}(n)$ *(green) and the* L^2 *error* $\mathrm{err}(n)$ *(red) are also depicted.*

Numerical illustration of imperfect modeling. In the following experiments, we illustrate with a toy example the results of Theorem 5.13 and show how the Sobolev regularization can be implemented directly in the PINN framework, taking advantage of the automatic differentiation and backpropagation. Let $\Omega =]0,1[^2$ and assume that $Y = u^\star(\mathbf{X}) + \mathcal{N}(0, 10^{-2})$, where $u^\star(x,t) = \exp(t-x) + 0.1\cos(2\pi x)$. In this hybrid modeling setting, the goal is to reconstruct $u^\star$. We consider an advection model of the form $\mathscr{F}(u,\mathbf{x}) = \partial_x u(\mathbf{x}) + \partial_t u(\mathbf{x})$, with $h(x,0) = \exp(-x)$ and $h(0,t) = \exp(t)$. The unique solution of this PDE is $u_{\mathrm{model}}(x,t) = \exp(t-x)$ (Figure 5, left). Note that the function u_{model} is different from $u^\star$ (Figure 5, middle), which casts our problem in the imperfect modeling setting. This PDE prior is relevant because $\|u_{\mathrm{model}} - u^\star\|^2_{L^2(\Omega)} \simeq \exp(-5,3)$ and $\mathrm{PI}(u^\star) \simeq \exp(-1,6)$, two quantities that are negligible with respect to $\|u^\star\|^2_{L^2(\Omega)} \simeq \exp(0.3)$. We randomly sample n observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ uniformly on the rectangle $\mathrm{supp}(\mu_{\mathbf{X}}) =]0, 0.5[\times]0,1[\subsetneq \Omega$ (note that this is a strict inclusion), and let n vary from $n_{\min} = 10$ to $n_{\max} = 10^3$ (linearly in a log scale).

The architecture of the neural networks is set to $H = 2$ hidden layers with width $D = 100$, so that the total number of parameters is $10\,600 \gg n_{\max}$. We fix $n_e, n_r = 10^4 \gg n_{\max}$ and $\lambda_{(\mathrm{ridge})} = \min(n_e, n_r)^{-1/2}$. Figure 3 shows in blue the evolution of the regularized risk $R^{(\mathrm{reg})}_{n,n_e,n_r}(u^{(n)}_{\hat{\theta}^{(\mathrm{reg})}(p,n_r,n_e,D)})$ with respect to the number p of epochs in the gradient descent (for $n = 10$). For a fixed number n of observations, the number $p_{\max}$ of epochs to stop training is determined by monitoring the evolution of the risk $R^{(\mathrm{reg})}_{n,n_e,n_r}(u^{(n)}_{\hat{\theta}^{(\mathrm{reg})}(p_{\max},n_r,n_e,D)})$ (blue curve) and the overfitting gap $\mathrm{OG}_{n,n_e,n_r} = |R^{(\mathrm{reg})}_{n,n_e,n_r} - \mathscr{R}^{(\mathrm{reg})}_n|$ (orange curve). Both are required to be stable around a minimal value, so that the minimum of the risk is approximately reached, i.e., $R^{(\mathrm{reg})}_{n,n_e,n_r}(u^{(n)}_{\hat{\theta}^{(\mathrm{reg})}(p_{\max},n_r,n_e,D)}) \simeq \inf_{u\in \mathrm{NN}_H(D)} R^{(\mathrm{reg})}_{n,n_e,n_r}(u)$ and $\mathscr{R}^{(\mathrm{reg})}_n(u^{(n)}_{\hat{\theta}^{(\mathrm{reg})}(p_{\max},n_r,n_e,D)}) \simeq \inf_{u\in \mathrm{NN}_H(D)} \mathscr{R}^{(\mathrm{reg})}_n(u)$. In this overparameterized regime (D is large), one can consider that $\mathscr{R}^{(\mathrm{reg})}_n(u^{(n)}_{\hat{\theta}^{(\mathrm{reg})}(p_{\max},n_r,n_e,D)}) \simeq \inf_{u\in C^\infty(\bar{\Omega},\mathbb{R}^{d_2})} \mathscr{R}^{(\mathrm{reg})}_n(u)$ (Theorem 4.7). Keeping n_e, n_r, and λ_{ridge} fixed, the proximity between the PINN and $u^\star$ is measured by

$$\mathrm{err}(n) = 2\int_0^{0.5}\int_0^1 \|u^{(n)}_{\hat{\theta}^{(\mathrm{reg})}(p_{\max},n_r,n_e,D)}(x,t) - u^\star(x,t)\|_2^2 dxdt.$$

According to Theorem 5.13, there exists some constant $\Lambda > 0$ such that, approximately,

$$\ln\left(\mathbb{E}(\mathrm{err}(n))\right) \lesssim \ln(\Lambda) - \frac{\ln(n)}{2}.$$

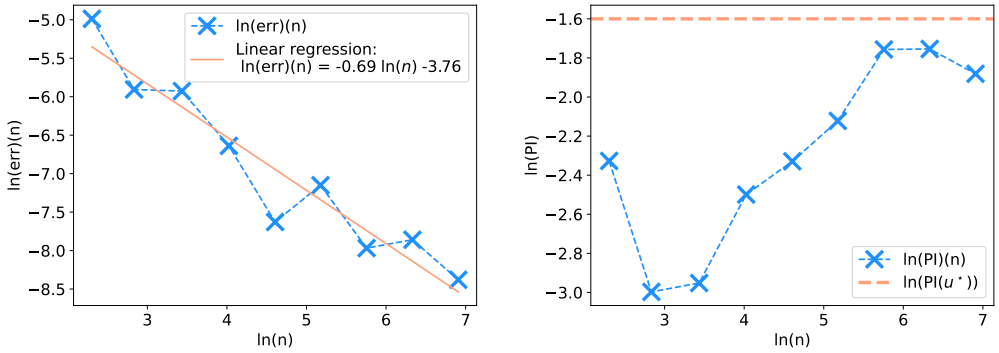

FIG 4. *Distance* $\mathrm{err}(n)$ *to* $u^\star$ *(left) and physics inconsistency* PI *(right) of the regularized PINN estimator with respect to the number* n *of observations in* log-log *scale.*

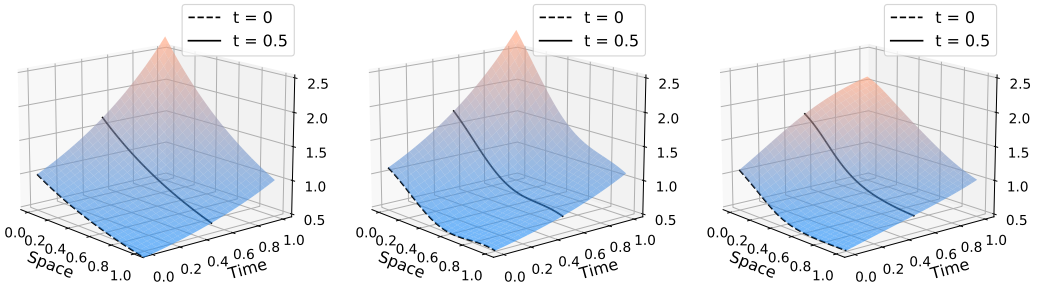

FIG 5. *Functions* u_{model} *(left),* $u^\star$ *(middle), and regularized PINN estimator with* $n = 10^3$ *(right).*

This bound is validated numerically in Figure 4, attesting a linear rate in log-log scale between $\mathrm{err}(n)$ and n of $-0.69 \leqslant -0.5$. Furthermore, the second statement of Theorem 5.13 suggests that $\ln \mathrm{PI}(n) = \ln \mathrm{PI}(u^{(n)}_{\hat{\theta}^{(\mathrm{reg})}(p_{\max}, n_r, n_e, D)}) \leqslant \ln \mathrm{PI}(u^\star) = -1.6$, which is also verified in Figure 4. Interestingly, the regularized PINN estimator quickly becomes more accurate than the initial model, since $\mathrm{err}(n)$ is less than $\int_\Omega \|u_{\mathrm{model}} - u^\star\|_2^2 d\mu_{\mathbf{X}} \simeq \exp(-5,3)$ as soon as $n \geqslant 17$.

The obtained regularized PINN estimator for $n = 10^3$ is shown in Figure 5 (right). This estimator looks globally similar to the model u_{model} (Figure 5, left) while managing to reconstruct the variation typical of the cosine perturbation of $u^\star$ (Figure 5, middle) at $t = 0$. Of course, for $t \geqslant 0.5$, the estimator cannot approximate $u^\star$ with an infinite precision, since the measurements $\mathbf{X}_i$ are only sampled for $t \leqslant 0.5$. However, the regularized PINN estimator succeeds to follow the advection equation dynamics, as it does not vary much along the lines $x - t = \mathrm{cste}$— despite some flattening effect of the Sobolev regularization for $t > 0.5$.

REFERENCES

M.S. Agranovich. Sobolev Spaces, Their Generalizations and Elliptic Problems in Smooth and Lipschitz Domains. Springer, Cham, 2015.

E. Arnone, A. Kneip, F. Nobile, and L.M. Sangalli. Some first results on the consistency of spatial regression with partial differential equation regularization. Statistica Sinica, 32:209–238, 2022.

A. Arzani, J.-X. Wang, and R.M. D'Souza. Uncovering near-wall blood flow from sparse data with physics-informed neural networks. Physics of Fluids, 33:071905, 2021.

L. Azzimonti, L.M. Sangalli, P. Secchi, M. Domanin, and F. Nobile. Blood flow velocity field estimation via spatial regression with PDE penalization. Journal of the American Statistical Association, 110:1057–1071, 2015.

H. Brezis. Functional Analysis, Sobolev Spaces and Partial Differential Equations. Springer, New York, 2010.
S. Cai, Z. Wang, S. Wang, P. Perdikaris, and G.E. Karniadakis. Physics-informed neural networks for heat transfer problems. Journal of Heat Transfer, 143(6), 2021.
B. Chandrajit, L. McLennan, T. Andeen, and A. Roy. Recipes for when physics fails: Recovering robust learning of physics informed neural networks. Machine Learning: Science and Technology, 4:015013, 2023.
L. Comtet. Advanced Combinatorics : The Art of Finite and Infinite Expansions. Springer, Dordrecht, 1974.
F.S. Costabal, Y. Yang, P. Perdikaris, D.E. Hurtado, and E. Kuhl. Physics-informed neural networks for cardiac activation mapping. Frontiers in Physics, 8:42, 2020.
B. Cunha, C. Droz, A. Zine, S. Foulard, and M. Ichchou. A review of machine learning methods applied to structural dynamics and vibroacoustic. arXiv:2204.06362, 2022.
S. Cuomo, V.S. Di Cola, F. Giampaolo, G. Rozza, M. Raissi, and F. Piccialli. Scientific machine learning through physics-informed neural networks: Where we are and what's next. Journal of Scientific Computing, 92:88, 2022.
D. Davini, B. Samineni, B. Thomas, A.H. Tran, C. Zhu, K. Ha, G. Dasika, and L. White. Using physics-informed regularization to improve extrapolation capabilities of neural networks. In Fourth Workshop on Machine Learning and the Physical Sciences (NeurIPS 2021), 2021.
A. Daw, A. Karpatne, W.D. Watkins, J.S. Read, and V. Kumar. Physics-guided neural networks (PGNN): An application in lake temperature modeling. In A. Karpatne, R. Kannan, and V. Kumar, editors, Knowledge guided machine learning: Accelerating discovery using scientific knowledge and data, pages 352–372, New York, 2022. Chapman and Hall/CRC.
E. de Bézenac, A. Pajot, and P. Gallinari. Deep learning for physical processes: Incorporating prior scientific knowledge. Journal of Statistical Mechanics: Theory and Experiment, page 124009, 2019.
T. De Ryck and S. Mishra. Error analysis for physics informed neural networks (PINNs) approximating Kolmogorov PDEs. Advances in Computational Mathematics, 48:79, 2022.
T. De Ryck, S. Lanthaler, and S. Mishra. On the approximation of functions by tanh neural networks. Neural Networks, 143:732–750, 2021.
T. de Wolff, H. Carrillo, L. Martí, and N. Sanchez-Pi. Towards optimally weighted physics-informed neural networks in ocean modelling. arXiv:2106.08747, 2021.
I.C. Esfahani. A data-driven physics-informed neural network for predicting the viscosity of nanofluids. AIP Advances, 13:025206, 2023.
L.C. Evans. Partial Differential Equations, volume 19 of Graduate Studies in Mathematics. American Mathematical Society, Providence, 2nd edition, 2010.
F. Ferraccioli, L.M. Sangalli, and L. Finos. Some first inferential tools for spatial regression with differential regularization. Journal of Multivariate Analysis, 189:104866, 2022.
S. Fischer and I. Steinwart. Sobolev norm learning rates for regularized least-squares algorithm. Journal of Machine Learning Research, 21:8464–8501, 2020.
G. Gokhale, B. Claessens, and C. Develder. Physics informed neural networks for control oriented thermal modeling of buildings. Applied Energy, 314:118852, 2022.
P. Grisvard. Elliptic Problems in Nonsmooth Domains, volume 69 of Classics in Applied Mathematics. SIAM, Philadelphia, 2011.
C. Guo, G. Pleiss, Y. Sun, and K.Q. Weinberger. On calibration of modern neural networks. In D. Precup and Y.W. Teh, editors, Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages 1321–1330. PMLR, 2017.
Z. Hao, S. Liu, Y. Zhang, C. Ying, Y. Feng, H. Su, and J. Zhu. Physics-informed machine learning: A survey on problems, methods and applications. arXiv:2211.08064, 2022.
M. Hardy. Combinatorics of partial derivatives. The Electronic Journal of Combinatorics, 13:R1, 2006.
Q. He, D. Barajas-Solano, G. Tartakovsky, and A.M. Tartakovsky. Physics-informed neural networks for multiphysics data assimilation with application to subsurface transport. Advances in Water Resources, 141:103610, 2020.
A.D. Jagtap, K. Kawaguchi, and G.E. Karniadakis. Adaptive activation functions accelerate convergence in deep and physics-informed neural networks. Journal of Computational Physics, 404:109136, 2020.
B. Kapusuzoglu and S. Mahadevan. Physics-informed and hybrid machine learning in additive manufacturing: Application to fused filament fabrication. JOM, 72:4695–4705, 2020.
G.E. Karniadakis, I.G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang. Physics-informed machine learning. Nature Reviews Physics, 3:422–440, 2021.
A. Krishnapriyan, A. Gholami, S. Zhe, R. Kirby, and M.W. Mahoney. Characterizing possible failure modes in physics-informed neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, Advances in Neural Information Processing Systems, volume 34, pages 26548–26560. Curran Associates, Inc., 2021.

A. Krogh and J. Hertz. A simple weight decay can improve generalization. In J. Moody, S. Hanson, and R.P. Lippmann, editors, Advances in Neural Information Processing Systems, volume 4, pages 950–957. Morgan-Kaufmann, 1991.

S. Li, G. Wang, Y. Di, L. Wang, H. Wang, and Q. Zhou. A physics-informed neural network framework to predict 3D temperature field without labeled data in process of laser metal deposition. Engineering Applications of Artificial Intelligence, 120:105908, 2023.

P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis. Explainable AI: A review of machine learning interpretability methods. Entropy, 23:18, 2021.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In 7th International Conference on Learning Representations, 2019.

S. Mishra and R. Molinaro. Estimates on the generalization error of physics-informed neural networks for approximating PDEs. IMA Journal of Numerical Analysis, 43:1–43, 2023.

M.A. Nabian and H. Meidani. Physics-driven regularization of deep neural networks for enhanced engineering design and analysis. Journal of Computing and Information Science in Engineering, 20:011006, 2020.

R. Nickl and B.M. Pötscher. Bracketing metric entropy rates and empirical central limit theorems for function classes of Besov- and Sobolev-type. Journal of Theoretical Probability, 20:177–199, 2007.

J.J. Pannell, S.E. Rigby, and G. Panoutsos. Physics-informed regularisation procedure in neural networks: An application in blast protection engineering. International Journal of Protective Structures, 13:555–578, 2022.

Y. Qian, Y. Zhang, Y. Huang, and S. Dong. Error analysis of physics-informed neural networks for approximating dynamic PDEs of second order in time. arxiv:2303.12245, 2023.

R. Rai and C.K. Sahu. Driven by data or derived through physics? A review of hybrid physics guided machine learning techniques with cyber-physical system (CPS) focus. IEEE Access, 8:71050–71073, 2020.

M. Raissi, P. Perdikaris, and G.E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378:686–707, 2019.

M. Ramezankhani, A. Nazemi, A. Narayan, H. Voggenreiter, M. Harandi, R. Seethaler, and A.S. Milani. A data-driven multi-fidelity physics-informed learning framework for smart manufacturing: A composites processing case study. In 2022 IEEE 5th International Conference on Industrial Cyber-Physical Systems (ICPS), pages 01–07. IEEE, 2022.

B. Riel, B. Minchew, and T. Bischoff. Data-driven inference of the mechanics of slip along glacier beds using physics-informed neural networks: Case study on Rutford Ice Stream, Antarctica. Journal of Advances in Modeling Earth Systems, 13:e2021MS002621, 2021.

L.C.G. Rogers and D. Williams. Diffusions, Markov processes and Martingales, volume 1, Foundations. Cambridge University Press, Cambridge, 2nd edition, 2000.

L.M. Sangalli. Spatial regression with partial differential equation regularisation. International Statistical Review, 89:505–531, 2021.

Y. Shin. On the convergence of physics informed neural networks for linear second-order elliptic and parabolic type PDEs. Communications in Computational Physics, 28:2042–2074, 2020.

Y. Shin, Z. Zhang, and G.E. Karniadakis. Error estimates of residual minimization using neural networks for linear PDEs. arXiv:2010.08019, 2020.

P. Shvartzman. On Sobolev extension domains in $\mathbb{R}^n$. Journal of Functional Analysis, 258:2205–2245, 2010.

C. Song, T. Alkhalifah, and U.B. Waheed. Solving the frequency-domain acoustic vti wave equation using physics-informed neural networks. Geophysical Journal International, 225:846–859, 2021.

E.M. Stein. Singular Integrals and Differentiability Properties of Functions, volume 30 of Princeton Mathematical Series. Princeton University Press, Princeton, 1970.

C.J. Stone. Optimal global rates of convergence for nonparametric regression. The Annals of Statistics, 10: 1040–1053, 1982.

R. van Handel. Probability in High Dimension. APC 550 Lecture Notes, Princeton University, 2016.

L. von Rueden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, M. Walczak, J. Pfrommer, A. Pick, R. Ramamurthy, J. Garcke, C. Bauckhage, and J. Schuecker. Informed machine learning – A taxonomy and survey of integrating prior knowledge into learning systems. IEEE Transactions on Knowledge and Data Engineering, 35:614–633, 2023.

C. Wang, E. Bentivegna, W. Zhou, L. Klein, and B. Elmegreen. Physics-informed neural network super resolution for advection-diffusion models. In Third Workshop on Machine Learning and the Physical Sciences (NeurIPS 2020), 2020a.

R. Wang, K. Kashinath, M. Mustafa, A. Albert, and R. Yu. Towards physics-informed deep learning for turbulent flow prediction. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 1457–1466. Association for Computing Machinery, 2020b.

S. Wang, X. Yu, and P. Perdikaris. When and why PINNs fail to train: A neural tangent kernel perspective. Journal of Computational Physics, 449:110768, 2022.

J. Willard, X. Jia, S. Xu, M. Steinbach, and V. Kumar. Integrating scientific knowledge with machine learning for engineering and environmental systems. ACM Computing Surveys, 55:66, 2023.
S. Wu, A. Zhu, Y. Tang, and B. Lu. Convergence of physics-informed neural networks applied to linear second-order elliptic interface problems. arXiv:2203.03407, 2022.
K. Xu, M. Zhang, J. Li, S.S. Du, K.-I. Kawarabayashi, and S. Jegelka. How neural networks extrapolate: From feedforward to graph neural networks. In International Conference on Learning Representations, 2021.
R. Zhang, Y. Liu, and H. Sun. Physics-guided convolutional neural network (PhyCNN) for data-driven seismic response modeling. Engineering Structures, 215:110704, 2020.

APPENDIX A: MATHEMATICAL DETAILS

Composition of functions. Given two functions $u, v : \mathbb{R} \to \mathbb{R}$, we denote by $u \circ v$ the function $u \circ v(x) = u(v(x))$. For all $k \in \mathbb{N}$, the function $u^{\circ k}$ is defined by induction as $u^{\circ 0}(x) = x$ and $u^{\circ(k+1)} = u^{\circ k} \circ u = u \circ u^{\circ k}$. The composition symbol is placed before the derivative, so that the kth derivative of $u^{\circ H}$ is denoted by $(u^{\circ H})^{(k)}$.

Norms. The p-norm $\|x\|_p$ of a vector $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ is defined by $\|x\|_p = (\frac{1}{d}\sum_{i=1}^d |x_i|^p)^{1/p}$. In addition, $\|x\|_\infty = \max_{1 \leqslant i \leqslant d} |x_i|$. For a function $u : \Omega \to \mathbb{R}^d$, we let $\|u\|_{L^p(\Omega)} = (\frac{1}{|\Omega|}\int_\Omega \|u\|_p^p)^{1/p}$. Similarly, $\|u\|_{\infty,\Omega} = \sup_{x\in\Omega} \|u(x)\|_\infty$. For the sake of clarity, we sometimes write $\|u\|_\infty$ instead of $\|u\|_{\infty,\Omega}$.

Multi-indices and partial derivatives. For a multi-index $\alpha = (\alpha_1, \dots, \alpha_{d_1}) \in \mathbb{N}^d$ and a differentiable function $u : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$, the α partial derivative of u is defined by

$$\partial^\alpha u = (\partial_1)^{\alpha_1} \dots (\partial_{d_1})^{\alpha_{d_1}} u.$$

The set of multi-indices of sum less than k is defined by

$$\{|\alpha| \leqslant k\} = \{(\alpha_1, \dots, \alpha_{d_1}) \in \mathbb{N}^d, \alpha_1 + \dots + \alpha_{d_1} \leqslant k\}.$$

If $\alpha = 0$, $\partial^\alpha u = u$. Given two multi-indices α and β, we write $\alpha \leqslant \beta$ when $\alpha_i \leqslant \beta_i$ for all $1 \leqslant i \leqslant d_1$. The set of multi-indices less than α is denoted by $\{\beta \leqslant \alpha\}$. For a multi-index α such that $|\alpha| \leqslant k$, both sets $\{|\beta| \leqslant k\}$ and $\{\beta \leqslant \alpha\}$ are contained in $\{0, \dots, k\}^{d_1}$ and are therefore finite.

Hölder norm. For $K \in \mathbb{N}$, the Hölder norm of order K of a function $u \in C^K(\Omega, \mathbb{R}^d)$, is defined by $\|u\|_{C^K(\Omega)} = \max_{|\alpha|\leqslant K} \|\partial^\alpha u\|_{\infty,\Omega}$. This norm allows to bound a function as well as its derivatives. The space $C^K(\Omega, \mathbb{R}^d)$ endowed with the Hölder norm $\|\cdot\|_{C^K(\Omega)}$ is a Banach space. The space $C^\infty(\bar{\Omega}, \mathbb{R}^{d_2})$ is defined as the subspace of continuous functions $u : \bar{\Omega} \to \mathbb{R}^{d_2}$ satisfying $u|_\Omega \in C^\infty(\Omega, \mathbb{R}^{d_2})$ and, for all $K \in \mathbb{N}$, $\|u\|_{C^K(\Omega)} < \infty$.

Lipschitz function. Given a normed space $(V, \|\cdot\|)$, the Lipschitz norm of a function $u : V \to \mathbb{R}^{d_1}$ is defined by

$$\|u\|_{\text{Lip}} = \sup_{x,y\in V} \frac{\|u(x) - u(y)\|_2}{\|x - y\|}.$$

A function u is Lipschitz if $\|u\|_{\text{Lip}} < \infty$. The mean value theorem implies that for all $u \in C^1(V, \mathbb{R})$, $\|u\|_{\text{Lip}} \leqslant \|u\|_{C^1(V)}$.

Lipschitz surface and domain. A surface $\Gamma \subseteq \mathbb{R}^{d_1}$ is said to be Lipschitz if locally, in a neighborhood $U(x)$ of any point $x \in \Gamma$, an appropriate rotation r_x of the coordinate system transforms Γ into the graph of a Lipschitz function ϕ_x, i.e.,

$$r_x(\Gamma \cap U_x) = \{(x_1, \dots, x_{d-1}, \phi_x(x_1, \dots, x_{d-1})), \forall (x_1, \dots, x_d) \in r_x(\Gamma \cap U_x)\}.$$

A domain $\Omega \subseteq \mathbb{R}^{d_1}$ is said to be Lipschitz if its has Lipschitz boundary and lies on one side of it, i.e., $\phi_x < 0$ or $\phi_x > 0$ on all intersections $\Omega \cap U_x$. All manifolds with C^1 boundary and all convex domains are Lipschitz domains [e.g., Agranovich, 2015].

Sobolev spaces. Let $\Omega \subseteq \mathbb{R}^{d_1}$ be an open set. A function $v \in L^2(\Omega, \mathbb{R}^{d_2})$ is said to be the αth weak derivative of $u \in L^2(\Omega, \mathbb{R}^{d_2})$ if, for any $\phi \in C^\infty(\bar{\Omega}, \mathbb{R}^{d_2})$ with compact support in Ω, one has $\int_\Omega \langle v, \phi \rangle = (-1)^{|\alpha|} \int_\Omega \langle u, \partial^\alpha \phi \rangle$. This is denoted by $v = \partial^\alpha u$. For $m \in \mathbb{N}$, the Sobolev space $H^m(\Omega, \mathbb{R}^{d_2})$ is the space of all functions $u \in L^2(\Omega, \mathbb{R}^{d_2})$ such that $\partial^\alpha u$ exists for all $|\alpha| \leqslant m$. This space is naturally endowed with the norm $\|u\|_{H^m(\Omega)} = (\sum_{|\alpha| \leqslant m} |\Omega|^{-1} \|\partial^\alpha u\|^2_{L^2(\Omega)})^{1/2}$. For example, the function $u :]-1, 1[\to \mathbb{R}$ such that $u(x) = |x|$ is not derivable on $]-1, 1[$, but it admits $u'(x) = \mathbf{1}_{x>0} - \mathbf{1}_{x<0}$ as weak derivative. Since $u' \in L^2([-1, 1], \mathbb{R})$, u belongs to the Sobolev space $H^1(]-1, 1[, \mathbb{R})$. However, u' has no weak derivative, and so $u \notin H^2(]-1, 1[, \mathbb{R})$. Of course, if a function u belongs to the Hölder space $C^K(\bar{\Omega}, \mathbb{R}^{d_2})$, then it belongs to the Sobolev space $H^K(\Omega, \mathbb{R}^{d_2})$, and its weak derivatives are the usual derivatives. For more on Sobolev spaces, we refer the reader to Evans [2010, Chapter 5].

APPENDIX B: SOME RESULTS OF FUNCTIONAL ANALYSIS ON LIPSCHITZ DOMAINS

Extension theorems. Let $\Omega \subseteq \mathbb{R}^{d_1}$ be an open set and let $K \in \mathbb{N}$ be an order of differentiation. It is not straightforward to extend a function $u \in H^K(\Omega, \mathbb{R}^{d_2})$ to a function $\tilde{u} \in H^K(\mathbb{R}^{d_1}, \mathbb{R}^{d_2})$ such that

$$\tilde{u}|_\Omega = u|_\Omega \quad \text{and} \quad \|\tilde{u}\|_{H^K(\mathbb{R}^{d_1})} \leqslant C_\Omega \|u\|_{H^K(\Omega)},$$

for some constant C_Ω independent of u. This result is known as the extension theorem in Evans [2010, Chapter 5.4] when Ω is a manifold with C^1 boundary. However, the simplest domains in PDEs take the form $]0, L[\times]0, T[$, the boundary of which is not C^1. Fortunately, Stein [1970, Theorem 5 Chapter VI.3.3] provides an extension theorem for bounded Lipschitz domains. We refer the reader to Shvartzman [2010] for a survey on extension theorems.

Example of a non-extendable domain. Let the domain $\Omega =]-1, 1[^2 \backslash (\{0\} \times [0, 1[)$ be the square $]-1, 1[^2$ from which the segment $\{0\} \times [0, 1[$ has been removed. Then the function

$$u(x, y) = \begin{cases} 0 & \text{if } x < 0 \text{ or if } y \leqslant 0 \\ \exp(-\frac{1}{y}) & \text{if } x, y > 0, \end{cases}$$

belongs to $C^\infty(\Omega, \mathbb{R})$ but cannot be extended to $\mathbb{R}^2$, since it cannot be continuously extended to the segment $\{0\} \times [0, 1[$. Notice that Ω is not a Lipschitz domain because it lies on both sides of the segment $\{0\} \times [0, 1[$, which belongs to its boundary $\partial\Omega$.

THEOREM B.1 (Sobolev inequalities). *Let $\Omega \subseteq \mathbb{R}^{d_1}$ be a bounded Lipschitz domain and let $m \in \mathbb{N}$. If $m \geqslant d_1/2$, then there exists an operator $\tilde{\Pi} : H^m(\Omega, \mathbb{R}^{d_2}) \to C^0(\Omega, \mathbb{R}^{d_2})$ such that, for any $u \in H^m(\Omega, \mathbb{R}^{d_2})$, $\tilde{\Pi}(u) = u$ almost everywhere. Moreover, there exists a constant $C_\Omega > 0$, depending only on Ω, such that, $\|\tilde{\Pi}(u)\|_{\infty,\Omega} \leqslant C_\Omega \|u\|_{H^m(\Omega)}$.*

PROOF. Since Ω is a bounded Lipschitz domain, there exists a radius $r > 0$ such that $\Omega \subseteq B(0, r)$. According to the extension theorem [Stein, 1970, Theorem 5, Chapter VI.3.3], there exists a constant $C_\Omega > 0$, depending only on Ω, such that any $u \in H^m(\Omega, \mathbb{R}^{d_2})$ can be extended to $\tilde{u} \in H^m(B(0, r), \mathbb{R}^{d_2})$, with $\|\tilde{u}\|_{H^m(B(0,r))} \leqslant C_\Omega \|u\|_{H^m(\Omega)}$. Since $m \geqslant d_1/2$, the Sobolev inequalities [e.g., Evans, 2010, Chapter 5.6, Theorem 6] state that there exists a constant $\tilde{C}_\Omega > 0$, depending only on Ω, and a linear embedding $\Pi : H^m(B(0, r), \mathbb{R}^{d_2}) \to C^0(B(0, r), \mathbb{R}^{d_2})$ such that $\|\Pi(\tilde{u})\|_\infty \leqslant \tilde{C}_\Omega \|\tilde{u}\|_{H^m(B(0,r))}$ and $\Pi(\tilde{u}) = \tilde{u}$ in $H^m(B(0, r), \mathbb{R}^{d_2})$. Therefore, $\tilde{\Pi}(u) = \Pi(\tilde{u})|_\Omega$ and $\|\tilde{\Pi}(u)\|_{\infty,\Omega} \leqslant C_\Omega \tilde{C}_\Omega \|u\|_{H^m(\Omega)}$. □

DEFINITION B.2 (Weak convergence in $L^2(\Omega)$). A sequence $(u_p)_{p\in\mathbb{N}} \in L^2(\Omega)^{\mathbb{N}}$ weakly converges to $u_\infty \in L^2(\Omega)$ if, for any $\phi \in L^2(\Omega)$, $\lim_{p\to\infty} \int_\Omega \phi u_p = \int_\Omega \phi u_\infty$. This convergence is denoted by $u_p \rightharpoonup u_\infty$.

The Cauchy-Schwarz inequality shows that the convergence with respect to the $L^2(\Omega)$ norm implies the weak convergence. However, the converse is not true. For example, the sequence of functions $u_p(x) = \cos(px)$ weakly converges to 0 in $L^2([-\pi,\pi])$, whereas $\|u_p\|_{L^2([-\pi,\pi])} = 1/2$.

DEFINITION B.3 (Weak convergence in $H^m(\Omega)$). A sequence $(u_p)_{p\in\mathbb{N}} \in H^m(\Omega)^{\mathbb{N}}$ weakly converges to $u_\infty \in H^m(\Omega)$ in $H^m(\Omega)$ if, for all $|\alpha| \leqslant m$, $\partial^\alpha u_p \rightharpoonup \partial^\alpha u_\infty$.

THEOREM B.4 (Rellich-Kondrachov). *Let $\Omega \subseteq \mathbb{R}^{d_1}$ be a bounded Lipschitz domain and let $m \in \mathbb{N}$. Let $(u_p)_{p\in\mathbb{N}} \in H^{m+1}(\Omega,\mathbb{R}^{d_2})$ be a sequence such that $(\|u_p\|_{H^{m+1}(\Omega)})_{p\in\mathbb{N}}$ is bounded. There exists a function $u_\infty \in H^{m+1}(\Omega,\mathbb{R}^{d_2})$ and a subsequence of $(u_p)_{p\in\mathbb{N}}$ that converges to u_∞ both weakly in $H^{m+1}(\Omega,\mathbb{R}^{d_2})$ and with respect to the $H^m(\Omega)$ norm.*

PROOF. Let $r > 0$ be such that $\Omega \subseteq B(0,r)$. According to the extension theorem of Stein [1970, Theorem 5, Chapter VI.3.3], there exists a constant $C_r > 0$ such that each u_p can be extended to $\tilde{u}_p \in H^{m+1}(B(0,r),\mathbb{R}^{d_2})$, with $\|\tilde{u}_p\|_{H^{m+1}(B(0,r))} \leqslant C_r \|u_p\|_{H^{m+1}(\Omega)}$. Observing that, for all $|\alpha| \leqslant m$, $\partial^\alpha \tilde{u}_p$ belongs to $H^1(B(0,r),\mathbb{R}^{d_2})$, the Rellich-Kondrachov compactness theorem [Evans, 2010, Theorem 1, Chapter 5.7] ensures that there exists a subsequence of $(\tilde{u}_p)_{p\in\mathbb{N}}$ that converges to an extension of u_∞ with respect to the $H^m(B(0,r))$ norm. Since the subsequence is also bounded, upon passing to another subsequence, it also weakly converges in $H^{m+1}(B(0,r),\mathbb{R}^{d_2})$ to $u_\infty \in H^{m+1}(B(0,r),\mathbb{R}^{d_2})$ [e.g., Evans, 2010, Chapter D.4]. Therefore, by considering the restrictions of all the previous functions to Ω, we deduce that there exists a subsequence of $(u_p)_{p\in\mathbb{N}}$ that converges to u_∞ both weakly in $H^{m+1}(\Omega)$ and with respect to the $H^m(\Omega)$ norm. □

APPENDIX C: SOME USEFUL LEMMAS

The nth Bell number B_n [Hardy, 2006] corresponds to the number of partitions of the set $\{1,\dots,n\}$. Bell numbers satisfy the relationship $B_0 = 1$ and

$$B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k. \tag{7}$$

For $K \geqslant 1$ and $u \in C^K(\mathbb{R}^{d_1},\mathbb{R}^{d_2})$, the Kth derivative of u is denoted by $u^{(K)}$.

LEMMA C.1 (Bounding the partial derivatives of a composition of functions). *Let $d_1, d_2 \geqslant 1$, $K \geqslant 0$, $f \in C^K(\mathbb{R}^{d_1},\mathbb{R})$, and $g \in C^K(\mathbb{R},\mathbb{R}^{d_2})$. Then*

$$\|g \circ f\|_{C^K(\mathbb{R}^{d_1})} \leqslant B_K \|g\|_{C^K(\mathbb{R})} (1 + \|f\|_{C^K(\mathbb{R}^{d_1})})^K.$$

PROOF. Let $K_1 \leqslant K$ and let $\Pi(K_1)$ be the set of all partitions of $\{1,\dots,K_1\}$. According to Hardy [2006, Proposition 1], one has, for all $h \in C^{K_1}(\mathbb{R}^{K_1+d_1},\mathbb{R})$,

$$\partial^{K_1}_{1,2,3,\dots,K_1}(g \circ h) = \sum_{P\in\Pi(K_1)} g^{(|P|)} \circ h \times \prod_{S\in P} \Big[\Big(\prod_{j\in S} \partial_j \Big) h \Big].$$

Let $\alpha = (\alpha_1, \dots, \alpha_{d_1})$ be a multi-index such that $|\alpha| = K_1$. Setting $\alpha_0 = 0$, $y_j = x_{K_1+j} + (x_{\alpha_1+\cdots+\alpha_{j-1}} + \cdots + x_{\alpha_1+\cdots+\alpha_j-1})$, and letting $h(x_1, \dots, x_{K_1+d_1}) = f(y_1, \dots, y_{d_1})$, we are led to

$$\partial^\alpha(g \circ f) = \sum_{P \in \Pi(K_1)} g^{(|P|)} \circ f \times \prod_{S \in P} \partial^{\alpha(S)} f, \tag{8}$$

where $\alpha(S) = (|\{b \in S, \quad \alpha_1 + \cdots + \alpha_{\ell-1} \leqslant b \leqslant \alpha_1 + \cdots + \alpha_\ell\}|)_{1 \leqslant \ell \leqslant d_1}$. Moreover, by definition of the Bell number, $|\Pi(K_1)| = B_{K_1}$, and, by definition of a partition, $|P| \leqslant K_1$. So,

$$\begin{aligned}\|\partial^\alpha(g \circ f)\|_\infty &\leqslant B_{K_1} \|g\|_{C^{K_1}(\mathbb{R}^{d_1})} \max_{i_1+2i_2+\cdots+K_1 i_{K_1}=K_1} \prod_{j=1}^{K_1} \|f\|^{i_j}_{C^j(\mathbb{R}^{d_1})} \\ &\leqslant B_{K_1} \|g\|_{C^{K_1}(\mathbb{R}^{d_1})} (1 + \|f\|_{C^{K_1}(\mathbb{R}^{d_1})})^{K_1}.\end{aligned}$$

Since this inequality is true for all $K_1 \leqslant K$ and for all $|\alpha| = K_1$, the lemma is proved. □

LEMMA C.2 (Bounding the partial derivatives of a changing of coordinates f). *Let $d_1, d_2 \geqslant 1$, $K \geqslant 0$, $f \in C^K(\mathbb{R}, \mathbb{R})$, and $g \in C^K(\mathbb{R}^{d_1}, \mathbb{R}^{d_2})$. Let $v \in C^K(\mathbb{R}^{d_1}, \mathbb{R}^{d_1})$ be defined by $v(\mathbf{x}) = (f(x_1), \dots, f(x_{d_1}))$. Then*

$$\|g \circ v\|_{C^K(\mathbb{R}^{d_1})} \leqslant B_K \times \|g\|_{C^K(\mathbb{R}^{d_1})} \times (1 + \|f\|_{C^K(\mathbb{R})})^K.$$

PROOF. Let $\alpha = (\alpha_1, \dots, \alpha_{d_1})$ be a multi-index such that $|\alpha| = K$. For $\mathbf{x} = (x_1, \dots, x_{d_1})$ and a fixed $i \in \{1, \dots, d_1\}$, we let $h(t) = g(f(x_1), \dots, f(x_{i-1}), t, f(x_{i+1}), \dots, f(x_{d_1}))$. Clearly, $(h \circ f)^{(\alpha_i)}(x_i) = (\partial_i)^{\alpha_i}(g \circ v)(\mathbf{x})$. Thus, according to Lemma C.1,

$$(h \circ f)^{(\alpha_i)} = \sum_{P_i \in \Pi(\alpha_i)} h^{(|P_i|)} \circ f \times \prod_{S_i \in P_i} f^{(|S_i|)}.$$

Therefore,

$$(\partial_i)^{\alpha_i}(g \circ v)(\mathbf{x}) = \sum_{P_i \in \Pi(\alpha_i)} (\partial_i)^{|P_i|} g \circ v(\mathbf{x}) \prod_{S_i \in P_i} f^{(|S_i|)}(x_i).$$

Letting $i = 1$ and observing that $\partial_j f^{(|S_1|)}(x_1) = 0$ for $j \neq 1$, we see that

$$\partial^\alpha(g \circ v)(\mathbf{x}) = \sum_{P_1 \in \Pi(\alpha_1)} \Big[\prod_{S_1 \in P_1} f^{(|S_1|)}(x_1) \Big] \times (\partial_2)^{\alpha_2} \dots (\partial_{d_1})^{\alpha_{d_1}} [(\partial_1)^{|P_1|} g \circ v](\mathbf{x}).$$

Repeating the same procedure for $(\partial_1)^{|P_1|} g \circ v, \dots, (\partial_1)^{|P_1|} \dots (\partial_{d_1})^{|P_{d_1}|} g \circ v$, we obtain

$$\begin{aligned}\partial^\alpha(g \circ v)(\mathbf{x}) = &\sum_{P_1 \in \Pi(\alpha_1)} \Big[\prod_{S_1 \in P_1} f^{(|S_1|)}(x_1)] \Big] \times \cdots \\ &\cdots \times \sum_{P_{d_1} \in \Pi(\alpha_{d_1})} \Big[\prod_{S_{d_1} \in P_{d_1}} f^{(|S_{d_1}|)}(x_{d_1})] \Big] \times (\partial_1)^{|P_1|} \dots (\partial_{d_1})^{|P_{d_1}|} g \circ v(\mathbf{x}).\end{aligned}$$

Since $\sum_{S_i \in P_i} |S_i| = \alpha_i$ and $\sum_{i=1}^{d_1} \alpha_i = K$, we conclude that

$$\|\partial^\alpha(g \circ v)\|_\infty \leqslant B_{\alpha_1} \times \cdots \times B_{\alpha_{d_1}} \times \|\partial^\alpha g\|_\infty (1 + \|f\|_{C^K(\mathbb{R})})^K.$$

Using the injective map $\mathcal{M} : \Pi(\alpha_1) \times \cdots \times \Pi(\alpha_{d_1}) \to \Pi(K)$ such that $\mathcal{M}(P_1, \dots, P_{d_1}) = \cup_{i=1}^{d_1} P_i$, we have $B_{\alpha_1} \times \cdots \times B_{\alpha_{d_1}} \leqslant B_K$. This concludes the proof. □

LEMMA C.3 (Bounding hyperbolic tangent and its derivatives). *For all $K \in \mathbb{N}$, one has*

$$\| \tanh^{(K)} \|_\infty \leqslant 2^{K-1}(K+2)!$$

PROOF. The $\tanh$ function is a solution of the equation $y' = 1 - y^2$. An elementary induction shows that there exists a sequence of polynomials $(P_K)_{K\in\mathbb{N}}$ such that $\tanh^{(K)} = P_K(\tanh)$, with $P_0(X) = X$ and $P_{K+1}(X) = (1-X^2) \times P_K'(X)$. Clearly, P_K is a real polynomial of degree $K+1$, of the form $P_K(X) = a_0^{(K)} + a_1^{(K)} X + \cdots + a_{K+1}^{(K)} X^{K+1}$. One verifies that $a_i^{(K+1)} = (i+1)a_{i+1}^{(K)} - (i-1)a_{i-1}^{(K)}$, with $a_{-1}^{(K)} = a_{K+2}^{(K)} = 0$. The largest coefficient $M(P_K) = \max_{0\leqslant i\leqslant K+1} |a_i^{(K)}|$ of P_K satisfies $M(P_{K+1}) \leqslant 2(K+1) \times M(P_K)$. Thus, since $M(P_1) = 1$, we see that $M(P_K) \leqslant 2^{K-1}K!$. Recalling that $0 \leqslant \tanh \leqslant 1$, we conclude that

$$\| \tanh^{(K)} \|_\infty = \|P_K(\tanh)\|_\infty \leqslant (K+2)M(P_K) \leqslant 2^{K-1}(K+2)!$$

□

In the sequel, for all $\theta \in \mathbb{R}$, we write $\tanh_\theta(x) = \tanh(\theta x)$. We define the sign function such that $\mathrm{sgn}(x) = \mathbf{1}_{x>0} - \mathbf{1}_{x<0}$.

LEMMA C.4 (Characterizing the limit of hyperbolic tangent in Hölder norm). *Let $K \in \mathbb{N}$ and $H \in \mathbb{N}^\star$. Then, for all $\varepsilon > 0$, $\lim_{\theta\to\infty} \| \tanh_\theta^{\circ H} - \mathrm{sgn}\|_{C^K(\mathbb{R}\setminus]-\varepsilon,\varepsilon[)} = 0$.*

PROOF. Fix $\varepsilon > 0$. We prove the stronger statement that, for all $m \in \mathbb{N}$, one has

$$\lim_{\theta\to\infty} \theta^m \| \tanh_\theta^{\circ H} - \mathrm{sgn}\|_{C^K(\mathbb{R}\setminus]-\varepsilon,\varepsilon[)} = 0.$$

We start with the case $H = 1$ and then prove the result by induction on H. Observe first, since $\tanh_\theta^{\circ H} - \mathrm{sgn}$ is an odd function, that

$$\| \tanh_\theta^{\circ H} - \mathrm{sgn}\|_{C^K(\mathbb{R}\setminus]-\varepsilon,\varepsilon[)} = \| \tanh_\theta^{\circ H} - \mathrm{sgn}\|_{C^K([\varepsilon,\infty[)}.$$

The case $H = 1$. Assume, to start with, that $K = 0$. For all $x \geqslant \varepsilon$, one has

$$\theta^m |\tanh_\theta(x) - 1| = \frac{2\theta^m}{1 + \exp(-2\theta x)} \leqslant \frac{2\theta^m}{1 + \exp(-2\theta\varepsilon)}.$$

Therefore, for all $m \in \mathbb{N}$,

$$\theta^m \| \tanh_\theta - \mathrm{sgn}\|_{\infty,\mathbb{R}\setminus]-\varepsilon,\varepsilon[} = \theta^m \| \tanh_\theta - \mathrm{sgn}\|_{\infty,[\varepsilon,\infty[} \leqslant \frac{2\theta^m}{1 + \exp(-2\theta\varepsilon)} \xrightarrow{\theta\to\infty} 0.$$

Next, to prove that the result if true for all $K \geqslant 1$, it is enough to show that, for all m, $\theta^m \| \tanh_\theta^{(K)} \|_{\infty,\mathbb{R}\setminus]-\varepsilon,\varepsilon[} \xrightarrow{\theta\to\infty} 0$. According to the proof of Lemma C.3, there exists a sequence of polynomials $(P_K)_{K\in\mathbb{N}}$ such that $\tanh^{(K)} = P_K(\tanh)$ and $P_{K+1}(X) = (1-X^2) \times P_K'(X)$. Since $\tanh_\theta(x) = \tanh(\theta x)$, one has

$$\begin{aligned}
\tanh_\theta^{(K)}(x) &= \theta^K \tanh^{(K)}(\theta x) \\
&= \theta^K (1 - \tanh^2(\theta x)) \times P_{K-1}'(\tanh(\theta x)) \\
&= \theta^K (1 - \tanh(\theta x))(1 + \tanh(\theta x)) \times P_{K-1}'(\tanh(\theta x)).
\end{aligned}$$

Fix $x \geqslant \varepsilon$. Then, letting $M_K = \|P'_{K-1}\|_{\infty,[-1,1]}$, we are led to

$$|\tanh_\theta^{(K)}(x)| \leqslant 2M_K\theta^K(1-\tanh(\theta x)) \leqslant 4M_K \times \frac{\theta^K}{1+\exp(2\theta x)} \leqslant 4M_K \times \frac{\theta^K}{1+\exp(2\theta\varepsilon)}.$$

This shows that $\theta^m\|\tanh_\theta^{(K)}\|_{\infty,[\varepsilon,\infty[} \leqslant 4M_K \times \frac{\theta^{K+m}}{1+\exp(2\theta\varepsilon)}$. One proves with similar arguments that the same result holds for all $x \leqslant -\varepsilon$. Thus,

$$\theta^m\|\tanh_\theta^{(K)}\|_{\infty,\mathbb{R}\setminus]-\varepsilon,\varepsilon[} \leqslant 4M_K \times \frac{\theta^{K+m}}{1+\exp(2\theta\varepsilon)} \xrightarrow{\theta\to\infty} 0,$$

and the lemma is proved for $H=1$.

Induction. Assume that that, for all K and all m,

$$\theta^m\|\tanh_\theta^{\circ H} - \mathrm{sgn}\|_{C^K(\mathbb{R}\setminus]-\varepsilon,\varepsilon[)} \xrightarrow{\theta\to\infty} 0. \tag{9}$$

Our objective is to prove that, for all K_2 and all m_2,

$$\theta^{m_2}\|\tanh_\theta^{\circ(H+1)} - \mathrm{sgn}\|_{C^{K_2}(\mathbb{R}\setminus]-\varepsilon,\varepsilon[)} \xrightarrow{\theta\to\infty} 0.$$

If $K_2 = 0$, since, for all $(x,y)\in\mathbb{R}^2$, $|\tanh_\theta(x)-\tanh_\theta(y)| \leqslant \theta|x-y| \times \|\tanh'\|_\infty \leqslant \theta|x-y|$. We deduce that

$$\theta^{m_2}\|\tanh_\theta^{\circ(H+1)} - \tanh_\theta(\mathrm{sgn})\|_{\infty,\mathbb{R}\setminus]-\varepsilon,\varepsilon[} \leqslant \theta^{m_2+1}\|\tanh_\theta^{\circ H} - \mathrm{sgn}\|_{\infty,\mathbb{R}\setminus]-\varepsilon,\varepsilon[}.$$

Therefore, according to (9), $\lim_{\theta\to\infty}\theta^{m_2}\|\tanh_\theta^{\circ(H+1)} - \tanh_\theta(\mathrm{sgn})\|_{\infty,\mathbb{R}\setminus]-\varepsilon,\varepsilon[} = 0$. Since $\tanh_\theta(\mathrm{sgn}) - \mathrm{sgn} = (\tanh(\theta)-1)\mathbf{1}_{x>0} - (\tanh(\theta)-1)\mathbf{1}_{x<0}$, we see that, for all m_2, $\lim_{\theta\to\infty}\theta^{m_2}\|\tanh_\theta(\mathrm{sgn}) - \mathrm{sgn}\|_{\infty,\mathbb{R}\setminus]-\varepsilon,\varepsilon[} = 0$. Using the triangle inequality, we conclude as desired that, for all m_2,

$$\theta^{m_2}\|\tanh_\theta^{\circ(H+1)} - \mathrm{sgn}\|_{\infty,\mathbb{R}\setminus]-\varepsilon,\varepsilon[} \xrightarrow{\theta\to\infty} 0. \tag{10}$$

Assume now that $K_2 \geqslant 1$. Since $\tanh_\theta^{\circ(H+1)} = \tanh^{\circ H}(\tanh)$, the Faà di Bruno formula [e.g., Comtet, 1974, Chapter 3.4] states that

$$(\tanh_\theta^{\circ(H+1)})^{(K_2)} = \sum_{m_1+2m_2+\cdots+K_2m_{K_2}=K_2} \frac{K_2!}{\prod_{i=1}^{K_2} m_i! \times i!^{m_i}} \times (\tanh_\theta^{\circ H})^{(m_1+\cdots+m_{K_2})}(\tanh_\theta) \times \prod_{j=1}^{K_2}(\tanh_\theta^{(j)})^{m_j}.$$

Notice that if $|x| \leq \mathrm{arctanh}(1/\sqrt{2})$, $|\tanh(x)| \geqslant \frac{|x|}{2}$ because by calling $f(x) = \tanh(x) - \frac{x}{2}$, $f(0)=0$ and $f'(x) = (1-\tanh(x)^2) - \frac{1}{2} \geqslant 0$. Therefore, if $|x| \geq \varepsilon$, $|\tanh(\theta x)| \geqslant \min(\frac{1}{\sqrt{2}}, \frac{\theta}{2}\varepsilon) \geqslant \varepsilon$ if $\theta \geqslant 2$ and $\varepsilon \geqslant \frac{1}{\sqrt{2}}$. This is why for $\theta \geqslant 2$ and $\varepsilon \leqslant 1$,

$$\|(\tanh_\theta^{\circ H})^{(m_1+\cdots+m_{K_2})}(\tanh_\theta)\|_{\infty,\mathbb{R}\setminus]-\varepsilon,\varepsilon[} \leqslant \|(\tanh_\theta^{\circ H})^{(m_1+\cdots+m_{K_2})}\|_{\infty,\mathbb{R}\setminus]-\varepsilon,\varepsilon[}.$$

Therefore, from the triangular inequality on $\|\cdot\|_{\infty,\mathbb{R}\setminus]-\varepsilon,\varepsilon[}$,

$$\|(\tanh_\theta^{\circ(H+1)})^{(K_2)}\|_{\infty,\mathbb{R}\setminus]-\varepsilon,\varepsilon[} \leqslant \sum_{m_1+2m_2+\cdots+K_2m_{K_2}=K_2} \frac{K_2!}{\prod_{i=1}^{K_2} m_i! \times i!^{m_i}} \times \|(\tanh_\theta^{\circ H})^{(m_1+\cdots+m_{K_2})}\|_{\infty,\mathbb{R}\setminus]-\varepsilon,\varepsilon[} \prod_{j=1}^{K_2}\|\tanh_\theta^{(j)}\|_{\infty,\mathbb{R}\setminus]-\varepsilon,\varepsilon[}^{m_j}.$$

According to the induction hypothesis (9), one has, for all $K \geqslant 1$ and all $m \in \mathbb{N}$,

$$\lim_{\theta\to\infty} \theta^m \|(\tanh_\theta^{\circ H})^{(K)}\|_{\infty,\mathbb{R}\setminus]-\varepsilon,\varepsilon[} = 0.$$

We deduce from the above that for all $K_2 \geqslant 1$ and all m_2,

$$\theta^{m_2} \|(\tanh_\theta^{\circ(H+1)})^{(K_2)}\|_{\infty,\mathbb{R}\setminus]-\varepsilon,\varepsilon[} \xrightarrow{\theta\to\infty} 0. \tag{11}$$

Combining (10) and (11), it comes that $\lim_{\theta\to\infty} \theta^{m_2} \|\tanh_\theta^{\circ(H+1)} - \mathrm{sgn}\|_{C^{K_2}(\mathbb{R}\setminus]-\varepsilon,\varepsilon[)} = 0$. □

COROLLARY C.5 (Bounding hyperbolic tangent compositions and their derivatives). *Let $K \in \mathbb{N}$ and $H \in \mathbb{N}^\star$. Then, for or all $\theta \in \mathbb{R}$, $\|(\tanh_\theta^{\circ H})^{(K)}\|_\infty < \infty$.*

PROOF. An induction as the one of Lemma C.4 shows that $\|(\tanh_\theta^{\circ H})^{(K)}\|_{\infty,\mathbb{R}\setminus]-\varepsilon,\varepsilon[} < \infty$. In addition, since $\tanh_\theta^{\circ H} \in C^\infty(\mathbb{R},\mathbb{R})$, $\|(\tanh_\theta^{\circ H})^{(K)}\|_{\infty,[-\varepsilon,\varepsilon]} < \infty$. □

When $d_1 = d_2 = 1$, the observations $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n) \in \mathbb{R}^2$ can be reordered as $(\mathbf{X}_{(1)}, Y_{(1)}), \dots, (\mathbf{X}_{(n)}, Y_{(n)})$ according to increasing values of the $\mathbf{X}_i$, that is, $\mathbf{X}_{(1)} \leqslant \cdots \leqslant \mathbf{X}_{(n)}$. Moreover, we let $\mathcal{G}(n, n_r) = \{(\mathbf{X}_i, Y_i), 1 \leqslant i \leqslant n\} \cup \{\mathbf{X}_j^{(r)}, 1 \leqslant j \leqslant n_r\}$, and denote by $\delta(n, n_r)$ the minimum distance between two distinct points in $\mathcal{G}(n, n_r)$, i.e.,

$$\delta(n, n_r) = \min_{\substack{z_1, z_2 \in \mathcal{G}(n,n_r) \\ z_1 \neq z_2}} |z_1 - z_2|. \tag{12}$$

LEMMA C.6 (Exact estimation with hyperbolic tangent). *Assume that $d_1 = d_2 = 1$, and let $H \geqslant 1$. Let the neural network $u_\theta \in \mathrm{NN}_H(n-1)$ be defined by*

$$u_\theta(x) = Y_{(1)} + \sum_{i=1}^{n-1} \frac{Y_{(i+1)} - Y_{(i)}}{2} \left[\tanh_\theta^{\circ H} \left(x - \mathbf{X}_{(i)} - \frac{\delta(n, n_r)}{2} \right) + 1 \right].$$

Then, for all $1 \leqslant i \leqslant n$,

$$\lim_{\theta\to\infty} u_\theta(\mathbf{X}_i) = Y_i.$$

Moreover, for all order $K \in \mathbb{N}^\star$ of differentiation and all $1 \leqslant j \leqslant n_r$,

$$\lim_{\theta\to\infty} u_\theta^{(K)}(\mathbf{X}_j^{(r)}) = 0.$$

PROOF. Applying Lemma C.4 with $\varepsilon = \delta(n,n_r)/4$ and letting

$$G = \mathbb{R}\setminus \cup_{i=1}^n]\mathbf{X}_{(i)} + \frac{1}{4}\delta(n, n_r), \mathbf{X}_{(i)} + \frac{3}{4}\delta(n, n_r)[,$$

one has, for all K, $\lim_{\theta\to\infty} \|u_\theta - u_\infty\|_{C^K(G)} = 0$, where

$$u_\infty(x) = Y_{(1)} + \sum_{i=1}^{n-1} \left[Y_{(i+1)} - Y_{(i)} \right] \times \mathbf{1}_{x > \mathbf{X}_{(i)} + \frac{\delta(n,n_r)}{2}}.$$

Clearly, for all $1 \leqslant i \leqslant n$, $u_\infty(\mathbf{X}_i) = Y_i$. Since $u_\infty'(x) = 0$ for all $x \in G$, and since $\mathbf{X}_j^{(r)} \in G$ for all $1 \leqslant j \leqslant n_r$, we deduce that $u_\infty^{(K)}(\mathbf{X}_j^{(r)}) = 0$. This concludes the proof. □

DEFINITION C.7 (Overfitting gap). For any $n, n_e, n_r \in \mathbb{N}^\star$ and $\lambda_{(\text{ridge})} \geqslant 0$, the overfitting gap operator OG_{n,n_e,n_r} is defined, for all $u \in C^\infty(\bar{\Omega}, \mathbb{R}^{d_2})$, by

$$\mathrm{OG}_{n,n_e,n_r}(u) = |R^{(\text{ridge})}_{n,n_e,n_r}(u) - \mathscr{R}_n(u)|.$$

LEMMA C.8 (Monitoring the overfitting gap). *Let $\varepsilon > 0$, $\lambda_{(\text{ridge})} \geqslant 0$, $H \geqslant 2$, and $D \in \mathbb{N}^\star$. Let $n, n_e, n_r \in \mathbb{N}^\star$. Let $\hat{\theta} \in \Theta_{H,D}$ be a parameter such that (i) $R^{(\text{ridge})}_{n,n_e,n_r}(u_{\hat{\theta}}) \leqslant \inf_{u \in \mathrm{NN}_H(D)} R^{(\text{ridge})}_{n,n_e,n_r}(u) + \varepsilon$ and (ii) $\mathrm{OG}_{n,n_e,n_r}(u_{\hat{\theta}}) \leqslant \varepsilon$. Then*

$$\mathscr{R}_n(u_{\hat{\theta}}) \leqslant \inf_{u \in \mathrm{NN}_H(D)} \mathscr{R}_n(u) + 2\varepsilon + o_{n_e,n_r \to \infty}(1).$$

PROOF. On the one hand, since $\mathscr{R}_n \leqslant R^{(\text{ridge})}_{n,n_e,n_r} + \mathrm{OG}_{n,n_e,n_r}$, assumptions (i) and (ii) imply that $\mathscr{R}_n(u_{\hat{\theta}}) \leqslant \inf_{u \in \mathrm{NN}_H(D)} R^{(\text{ridge})}_{n,n_e,n_r}(u) + 2\varepsilon$. On the other hand, $R^{(\text{ridge})}_{n,n_e,n_r} - \mathrm{OG}_{n,n_e,n_r} \leqslant \mathscr{R}_n$. The proof of Theorem 4.6 reveals that there exists a sequence $(\theta(n_e, n_r))_{n_e,n_r \in \mathbb{N}} \in \Theta^{\mathbb{N}}_{H,D}$ such that $\lim_{n_e,n_r \to \infty} \mathrm{OG}_{n,n_e,n_r}(u_{\theta(n_e,n_r)}) = 0$ and $\lim_{n_e,n_r \to \infty} \mathscr{R}_n(u_{\theta(n_e,n_r)}) = \inf_{u \in \mathrm{NN}_H(D)} \mathscr{R}_n(u)$. Thus, $\inf_{u \in \mathrm{NN}_H(D)} R^{(\text{ridge})}_{n,n_e,n_r}(u) \leqslant \inf_{\mathrm{NN}_H(D)} \mathscr{R}_n(u) + o_{n_e,n_r \to \infty}(1)$. We deduce that $\mathscr{R}_n(u_{\hat{\theta}}) \leqslant \inf_{u \in \mathrm{NN}_H(D)} \mathscr{R}_n(u) + 2\varepsilon + o_{n_e,n_r \to \infty}(1)$. □

LEMMA C.9 (Minimizing sequence of the theoretical risk.). *Let $H, D \in \mathbb{N}^\star$. Define the sequence $(v_p)_{p \in \mathbb{N}} \in \mathrm{NN}_H(D)^{\mathbb{N}}$ of neural networks by $v_p(\mathbf{x}) = \tanh_p \circ \tanh^{\circ(H-1)}(\mathbf{x})$. Then, for any $\lambda_e > 0$,*

$$\lim_{p \to \infty} \lambda_e (1 - v_p(1))^2 + \frac{1}{2} \int_{-1}^{1} \mathbf{x}^2 (v_p')^2(\mathbf{x}) d\mathbf{x} = 0.$$

PROOF. $\tanh^{\circ(H-1)}$ is an increasing C^∞ function such that $\tanh^{\circ(H-1)}(0) = 0$. Therefore, Lemma C.4 shows that $\lim_{p \to \infty} v_p(1) = 1$, so that $\lim_{p \to \infty} \lambda_e (1 - v_p(1))^2 = 0$. This shows the convergence of the left-hand term of the lemma.

To bound the right-hand term, we have, according to the chain rule,

$$|v_p'(\mathbf{x})| \leqslant p \| \tanh^{\circ(H-1)} \|_{C^1(\mathbb{R})} |\tanh'(p \tanh^{\circ(H-1)}(\mathbf{x}))|,$$

with $\| \tanh^{\circ(H-1)} \|_{C^1(\mathbb{R})} < \infty$ by Corollary C.5. Thus,

$$\int_{-1}^{1} \mathbf{x}^2 (v_p')^2(\mathbf{x}) d\mathbf{x} \leqslant \| \tanh^{\circ(H-1)} \|^2_{C^1(\mathbb{R})} \int_{-1}^{1} p^2 \mathbf{x}^2 (\tanh'(p \tanh^{\circ(H-1)}(\mathbf{x})))^2 d\mathbf{x}.$$

Notice that $\mathbf{x}^2 (\tanh'(p \tanh^{\circ(H-1)}(\mathbf{x})))^2$ is an even function, so that

$$\int_{-1}^{1} \mathbf{x}^2 (v_p')^2(\mathbf{x}) d\mathbf{x} \leqslant 2 \| \tanh^{\circ(H-1)} \|^2_{C^1(\mathbb{R})} \int_{0}^{1} p^2 \mathbf{x}^2 (\tanh'(p \tanh^{\circ(H-1)}(\mathbf{x})))^2 d\mathbf{x}.$$

Remark that $(\tanh')^2(\mathbf{x}) = (1 - \tanh(\mathbf{x}))^2 (1 + \tanh(\mathbf{x}))^2 \leqslant 16 \exp(-2\mathbf{x})$, so that

$$\int_{-1}^{1} \mathbf{x}^2 (v_p')^2(\mathbf{x}) d\mathbf{x} \leqslant 32 \| \tanh^{\circ(H-1)} \|^2_{C^1(\mathbb{R})} \int_{0}^{1} p^2 \mathbf{x}^2 \exp(-2p \tanh^{\circ(H-1)}(\mathbf{x})) d\mathbf{x}.$$

If $H = 1$, then the change of variable $\bar{\mathbf{x}} = p\mathbf{x}$ states that $\int_0^1 p^2 \mathbf{x}^2 \exp(-2p\mathbf{x}) d\mathbf{x} \leqslant p^{-1} \int_0^\infty \bar{\mathbf{x}}^2 \exp(-2\bar{\mathbf{x}}) d\bar{\mathbf{x}} \xrightarrow{p \to \infty} 0$ and the lemma is proved.

If $H \geqslant 2$, notice that $\tanh(\mathbf{x}) \geqslant \mathbf{x}\mathbf{1}_{\mathbf{x}\leqslant 1}/2 + \mathbf{1}_{\mathbf{x}\geqslant 1}/2$ for all $\mathbf{x} \geqslant 0$, so that $\tanh^{\circ(H-1)}(\mathbf{x}) \geqslant \mathbf{x}\mathbf{1}_{\mathbf{x}\leqslant 2^{H-1}}/2^H + \mathbf{1}_{\mathbf{x}\geqslant 2^{H-1}}/2^H$. Therefore, using the change of variable $\bar{\mathbf{x}} = p\mathbf{x}$,

$$\begin{aligned}\int_0^1 p^2\mathbf{x}^2 \exp(-2p\tanh^{\circ(H-1)}(\mathbf{x}))d\mathbf{x} &\leqslant \int_0^1 p^2\mathbf{x}^2\exp(-2^{H-1}p\mathbf{x})d\mathbf{x}\\ &\leqslant p^{-1}\int_0^\infty \bar{\mathbf{x}}^2\exp(-2^{H-1}\bar{\mathbf{x}})d\bar{\mathbf{x}}.\end{aligned}$$

Since this upper bound vanishes as $p \to \infty$, this concludes the proof when $H \geqslant 2$.

□

DEFINITION C.10 (Weak lower semi-continuity). A fonction $I : H^m(\Omega) \to \mathbb{R}$ is weakly lower semi-continuous on $H^m(\Omega)$ if, for any sequence $(u_p)_{p\in\mathbb{N}} \in H^m(\Omega)^{\mathbb{N}}$ that weakly converges to $u_\infty \in H^m(\Omega)$ in $H^m(\Omega)$, one has $I(u_\infty) \leqslant \liminf_{p\to\infty} I(u_p)$.

The following technical lemma will be useful for the proof of Proposition 5.6.

LEMMA C.11 (Weak lower semi-continuity with convex Lagrangians). *Let the Lagrangian $L \in C^\infty(\mathbb{R}^{\binom{d_1+m}{m}d_2} \times \cdots \times \mathbb{R}^{d_2} \times \mathbb{R}^{d_1}, \mathbb{R})$ be such that, for any $x^{(m)}, \dots, x^{(0)}$, and z, the function $x^{(m+1)} \mapsto L(x^{(m+1)}, \dots, x^{(0)}, z)$ is convex and nonnegative. Then the function $I : u \mapsto \int_\Omega L((\partial^{m+1}_{i_1,\dots,i_{m+1}} u(\mathbf{x}))_{1\leqslant i_1,\dots,i_{m+1}\leqslant d_1}, \dots, u(\mathbf{x}), \mathbf{x})d\mathbf{x}$ is lower-semi continuous for the weak topology on $H^{m+1}(\Omega, \mathbb{R}^{d_2})$.*

PROOF. This results generalizes Evans [2010, Theorem 1, Chapter 8.2], which treats the case $m = 0$. Let $(u_p)_{p\in\mathbb{N}} \in H^{m+1}(\Omega, \mathbb{R}^{d_2})^{\mathbb{N}}$ be a sequence that weakly converges to $u_\infty \in H^{m+1}(\Omega, \mathbb{R}^{d_2})$ in $H^{m+1}(\Omega, \mathbb{R}^{d_2})$. Our goal is to prove that $I(u_\infty) \leqslant \liminf_{p\to\infty} I(u_p)$. Upon passing to a subsequence, we can suppose that $\lim_{p\to\infty} I(u_p) = \liminf_{p\to\infty} I(u_p)$.

As a first step, we strengthen the convergence of $(u_p)_{p\in\mathbb{N}}$ by showing that for any $\varepsilon > 0$, there exists a subset E_ε of Ω such that $|\Omega\backslash E_\varepsilon| \leqslant \varepsilon$ (the notation $|\cdot|$ stands for the Lebesgue measure), and such that there exists a subsequence that uniformly converges on E_ε, as well as its derivatives. Recalling that a weakly convergent sequence is bounded [e.g., Evans, 2010, Chapter D.4], one has $\sup_{p\in\mathbb{N}} \|u_p\|_{H^{m+1}(\Omega)} < \infty$. Theorem B.4 ensures that a subsequence of $(u_p)_{p\in\mathbb{N}}$ converges to, say, $u_\infty \in H^{m+1}(\Omega, \mathbb{R}^{d_2})$ with respect to the $H^m(\Omega)$ norm. Upon passing again to another subsequence, we conclude that for all $|\alpha| \leqslant m$ and for almost every x in Ω, $\lim_{p\to\infty} \partial^\alpha u_p(x) = \partial^\alpha u_\infty(x)$ [see, e.g. Brezis, 2010, Theorem 4.9]. Finally, by Egorov's theorem [Evans, 2010, Chapter E.2], for any $\varepsilon > 0$, there exists a measurable set E_ε such that $|\Omega\backslash E_\varepsilon| \leqslant \varepsilon$ and such that, for all $|\alpha| \leqslant m$, $\lim_{p\to\infty} \|\partial^\alpha u_p - \partial^\alpha u_\infty\|_{L^\infty(E_\varepsilon)} = 0$.

Our next goal is to bound the function L. Let $F_\varepsilon = \{x \in \Omega, \sum_{|\alpha|\leqslant m+1} |\partial^\alpha u_\infty(x)| \leqslant \varepsilon^{-1}\}$ and $G_\varepsilon = E_\varepsilon \cap F_\varepsilon$. Observe that $\lim_{\varepsilon\to 0} |\Omega\backslash G_\varepsilon| = 0$. Since, for all $|\alpha| \leqslant m+1$, $\|\partial^\alpha u_\infty\|_{\infty,G_\varepsilon} < \infty$, and since $\lim_{p\to\infty} \|\partial^\alpha u_p - \partial^\alpha u_\infty\|_{L^\infty(G_\varepsilon)} = 0$, then, for all p large enough, $(\|\partial^\alpha u_p\|_{L^\infty(G_\varepsilon)})_{p\in\mathbb{N}}$ is bounded. For now, for the ease of notation, we write $(D^{m+1}u(z), \dots, u(z), z)$ instead of $((\partial^{m+1}_{i_1,\dots,i_{m+1}} u(z))_{1\leqslant i_1,\dots,i_{m+1}\leqslant d_1}, \dots, u(z), z)$. Therefore, since the Lagrangian L is smooth and Ω is bounded, for all p large enough, $(\|L(D^{m+1}u_p(\cdot), \dots, Du_p(\cdot), u_p(\cdot), \cdot)\|_{L^\infty(G_\varepsilon)})_{p\in\mathbb{N}}$ is bounded as well.

To conclude the proof, we take advantage of the convexity of the Lagrangian L. Let J_{m+1} be the Jacobian matrix of L along the vector $x^{(m+1)}$. The convexity of L implies

$$\begin{aligned}&L(D^{m+1}u_p(z), \dots, u_p(z), z)\\ &\quad\geqslant L(D^{m+1}u_\infty(z), D^m u_p(z) \dots, u_p(z), z)\\ &\qquad + J_{m+1}(D^{m+1}u_\infty(z), D^m u_p(z) \dots, u_p(z), z) \times (D^{m+1}u_p(z) - D^{m+1}u_\infty(z)).\end{aligned}$$

Using the fact that $L \geqslant 0$ and that $I(u_p) \geqslant \int_{G_\varepsilon} L(D^{m+1}u_p(z), \dots, u_p(z), z)dz$, we obtain

$$I(u_p) \geqslant \int_{G_\varepsilon} L(D^{m+1}u_\infty(z), D^m u_p(z), \dots, u_p(z), z) \\ + J_{m+1}(D^{m+1}u_\infty(z), D^m u_p(z), \dots, u_p(z), z) \times (D^{m+1}u_p(z) - D^{m+1}u_\infty(z))dz.$$

Since $(\|L(D^{m+1}u_p(\cdot), \dots, Du_p(\cdot), u_p(\cdot), \cdot)\|_{L^\infty(G_\varepsilon)})_{p\in\mathbb{N}}$ is bounded for p large enough, and since, for all $|\alpha| \leqslant m$, $\lim_{p\to\infty} \|\partial^\alpha u_p - \partial^\alpha u_\infty\|_{L^\infty(G_\varepsilon)} = 0$, the dominated convergence theorem ensures that

$$\lim_{p\to\infty} \int_{G_\varepsilon} L(D^{m+1}u_\infty(z), D^m u_p(z), \dots, u_p(z), z)dz = \int_{G_\varepsilon} L(D^{m+1}u_\infty(z), \dots, u_\infty(z), z)dz.$$

Using the fact that (i) L is smooth (and therefore Lipschitz on bounded domains), (ii) for all p large enough, $(\|\partial^\alpha u_p\|_{L^\infty(G_\varepsilon)})_{p\in\mathbb{N}}$ is bounded, and (iii) for all $|\alpha| \leqslant m$, $\lim_p \|\partial^\alpha u_p - \partial^\alpha u_\infty\|_{L^\infty(G_\varepsilon)} = 0$, we deduce that $\lim_{p\to\infty} \|J_{m+1}(D^{m+1}u_\infty(\cdot), D^m u_p(\cdot), \dots, u_p(\cdot), \cdot) - J_{m+1}(D^{m+1}u_\infty(\cdot), \dots, u_\infty(\cdot), \cdot)\|_{L^\infty(G_\varepsilon)} = 0$. Therefore, since $D^{m+1}u_p \rightharpoonup D^{m+1}u_\infty$,

$$\lim_{p\to\infty} \int_{G_\varepsilon} J_{m+1}(D^{m+1}u_\infty(z), D^m u_p(z), \dots, u_p(z), z) \times (D^{m+1}u_p(z) - D^{m+1}u_\infty(z))dz = 0.$$

Hence, $\lim_{p\to\infty} I(u_p) \geqslant \int_{G_\varepsilon} L(D^{m+1}u_\infty(z), \dots, u_\infty(z), z)dz$. Finally, applying the monotone convergence theorem with $\varepsilon \to 0$ shows that $\lim_{p\to\infty} I(u_p) \geqslant I(u_\infty)$, which is the desired result. □

LEMMA C.12 (Measurability of $\hat{u}_n$). *Let $\hat{u}_n = \operatorname{argmin}_{u\in H^{m+1}(\Omega,\mathbb{R}^{d_2})} \mathscr{R}_n^{(\mathrm{reg})}(u)$, where, for all $u \in H^{m+1}(\Omega, \mathbb{R}^{d_2})$,*

$$\mathscr{R}_n^{(\mathrm{reg})}(u) = \frac{\lambda_d}{n} \sum_{i=1}^n \|\tilde{\Pi}(u)(\mathbf{X}_i) - Y_i\|_2^2 + \lambda_e \mathbb{E}\|\tilde{\Pi}(u)(\mathbf{X}^{(e)}) - h(\mathbf{X}^{(e)})\|_2^2 \\ + \frac{1}{|\Omega|} \sum_{k=1}^M \|\mathscr{F}_k(u, \cdot)\|_{L^2(\Omega)} + \lambda_t \|u\|_{H^{m+1}(\Omega)}^2.$$

Then $\hat{u}_n$ is a random variable.

PROOF. Recall that

$$\mathscr{R}_n^{(\mathrm{reg})}(u) = \mathcal{A}_n(u, u) - 2\mathcal{B}_n(u) + \frac{\lambda_d}{n} \sum_{i=1}^n \|Y_i\|^2 + \lambda_e \mathbb{E}\|h(\mathbf{X}^{(e)})\|_2^2 + \frac{1}{|\Omega|} \sum_{k=1}^M \int_\Omega B_k(\mathbf{x})^2 d\mathbf{x}.$$

Throughout we use the notation $\mathcal{A}_{(\mathbf{x},e)}(u, u)$ instead of $\mathcal{A}_n(u, u)$, to make the dependence of $\mathcal{A}_n$ in the random variables $\mathbf{x} = (\mathbf{X}_1, \dots, \mathbf{X}_n)$ and $e = (\varepsilon_1, \dots, \varepsilon_n)$ more explicit. We do the same with $\mathcal{B}_n$. For a given a normed space $(F, \|\cdot\|)$, we let $\mathscr{B}(F, \|\cdot\|)$ be the Borel σ-algebra on F induced by the norm $\|\cdot\|$.

Our goal is to prove that the function

$$\hat{u}_n : (\Omega^n \times \mathbb{R}^{nd_2}, \mathscr{B}(\Omega^n \times \mathbb{R}^{nd_2}, \|\cdot\|_2)) \to (H^{m+1}(\Omega, \mathbb{R}^{d_2}), \mathscr{B}(H^{m+1}(\Omega, \mathbb{R}^{d_2}), \|\cdot\|_{H^{m+1}(\Omega)})) \\ (\mathbf{x}, e) \mapsto \operatorname*{argmin}_{u\in H^{m+1}(\Omega,\mathbb{R}^{d_2})} \mathcal{A}_{(\mathbf{x},e)}(u, u) - 2\mathcal{B}_{(\mathbf{x},e)}(u)$$

is measurable. Recall that $H^{m+1}(\Omega, \mathbb{R}^{d_2})$ is a Banach space separable with respect to its norm $\|\cdot\|_{H^{m+1}(\Omega)}$. Let $(v_q)_{q\in\mathbb{N}} \in H^{m+1}(\Omega, \mathbb{R}^{d_2})^{\mathbb{N}}$ be a sequence dense in $H^{m+1}(\Omega, \mathbb{R}^{d_2})$.

Note that, for any $\mathbf{x} \in \Omega^n$ and any $e \in \mathbb{R}^{nd_2}$, one has $\min_{u \in H^{m+1}(\Omega,\mathbb{R}^{d_2})} \mathcal{A}_{(\mathbf{x},e)}(u,u) - 2\mathcal{B}_{(\mathbf{x},e)}(u) = \inf_{q \in \mathbb{N}} \mathcal{A}_{(\mathbf{x},e)}(v_q,v_q) - 2\mathcal{B}_{(\mathbf{x},e)}(v_q)$. This identity is a consequence of the fact that the function $u \mapsto \mathcal{A}_{(\mathbf{x},e)}(u,u) - 2\mathcal{B}_{(\mathbf{x},e)}(u)$ is continuous for the $H^{m+1}(\Omega)$ norm, as shown in the proof of Proposition 5.5). Moreover, according to this proof, each function $F_q(\mathbf{x},e) := \mathcal{A}_{(\mathbf{x},e)}(u_q,u_q) - 2\mathcal{B}_{(\mathbf{x},e)}(u_q)$ is a composition of continuous functions, and is therefore measurable. Thus, the function

$$G(\mathbf{x},e) := \min_{u \in H^{m+1}(\Omega,\mathbb{R}^{d_2})} \mathcal{A}_{(\mathbf{x},e)}(u,u) - 2\mathcal{B}_{(\mathbf{x},e)}(u) = \inf_{q \in \mathbb{N}} \mathcal{A}_{(\mathbf{x},e)}(u_q,u_q) - 2\mathcal{B}_{(\mathbf{x},e)}(u_q)$$

is measurable.

Next, since Ω, $\mathbb{R}$, and $H^{m+1}(\Omega,\mathbb{R}^{d_2})$ are separable, we know that the σ-algebras $\mathscr{B}(\Omega^n \times \mathbb{R}^{nd_2} \times H^{m+1}(\Omega,\mathbb{R}^{d_2}), \|\cdot\|_\otimes)$ and $\mathscr{B}(\Omega^n \times \mathbb{R}^{nd_2}, \|\cdot\|_2) \otimes \mathscr{B}(H^{m+1}(\Omega,\mathbb{R}^{d_2}), \|\cdot\|_{H^{m+1}(\Omega)})$ are identical, where $\|(\mathbf{x},e,u)\|_\otimes = \|(\mathbf{x},e)\|_2 + \|u\|_{H^{m+1}(\Omega)}$ [see, e.g. Rogers and Williams, 2000, Chapter II.13, E13.11c]. This implies that the coordinate projections $\Pi_{\mathbf{x},e}$ and Π_u—defined for $(\mathbf{x},e) \in \Omega^n \times \mathbb{R}^{nd_2}$ and $u \in H^{m+1}(\Omega,\mathbb{R}^{d_2})$ by $\Pi_{\mathbf{x},e}(\mathbf{x},e,u) = (\mathbf{x},e)$ and $\Pi_u(\mathbf{x},e,u) = u$—are $\|\cdot\|_\otimes$ measurable. It is easy to check that, for any $(\mathbf{x},e) \in \Omega^n \times \mathbb{R}^{nd_2}$ and $u \in H^{m+1}(\Omega,\mathbb{R}^{d_2})$, if $\lim_{p\to\infty} \|(\mathbf{x}_p,e_p,u_p) - (\mathbf{x},e,u)\|_\otimes = 0$, then $\lim_{p\to\infty} \|\tilde{\Pi}(u_p) - \tilde{\Pi}(u)\|_{\infty,\Omega} = 0$ and, since $\tilde{\Pi}(u) \in C^0(\Omega,\mathbb{R}^{d_2})$, $\lim_{p\to\infty} \mathcal{A}_{\mathbf{x}_p,e_p}(u_p,u_p) - 2\mathcal{B}_{\mathbf{x}_p,e_p}(u_p) = \mathcal{A}_{\mathbf{x},e}(u,u) - 2\mathcal{B}_{\mathbf{x},e}(u)$. This proves that the function $I : (\Omega^n \times \mathbb{R}^{nd_2} \times H^{m+1}(\Omega,\mathbb{R}^{d_2}), \mathscr{B}(\Omega^n \times \mathbb{R}^{nd_2} \times H^{m+1}(\Omega,\mathbb{R}^{d_2}), \|\cdot\|_\otimes)) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ defined by

$$I(\mathbf{x},e,u) = \mathcal{A}_{(\mathbf{x},e)}(u,u) - 2\mathcal{B}_{(\mathbf{x},e)}(u)$$

is continuous with respect to $\|\cdot\|_\otimes$ and therefore measurable. According to the above, the function

$$\tilde{I}(\mathbf{x},e,u) = I(\mathbf{x},e,u) - G \circ \Pi_{x,e}(\mathbf{x},e,u)$$

is also measurable. Observe that, by definition, $\hat{u}_n = J \circ (\mathbf{X}_1,\dots,\mathbf{X}_n,\varepsilon_1,\dots,\varepsilon_n)$, where $J(\mathbf{x},e) = \Pi_u(\tilde{I}^{-1}(\{0\}) \cap (\{(\mathbf{x},e)\} \times H^{m+1}(\Omega,\mathbb{R}^{d_2})))$. For any measurable set S of $\mathscr{B}(H^{m+1}(\Omega,\mathbb{R}^{d_2}, \|\cdot\|_{H^{m+1}(\Omega)})$, $J^{-1}(S) = \Pi_{x,e}(\tilde{I}^{-1}(\{0\}) \cap (\Omega^n \times \mathbb{R}^{nd_2} \times S)) \in \mathscr{B}(\Omega^n \times \mathbb{R}^{nd_2})$. (Notice that $J^{-1}(S)$ is the collection of all pairs $(\mathbf{x},e) \in \Omega^n \times \mathbb{R}^{nd_2}$ satisfying $\operatorname{argmin}_{u \in H^{m+1}(\Omega,\mathbb{R}^{d_2})} \mathcal{A}_{(\mathbf{x},e)}(u,u) - 2\mathcal{B}_{(\mathbf{x},e)}(u) \in S$.) To see this, jut note that for any set $\tilde{S} \in \mathscr{B}(\Omega^n \times \mathbb{R}^{nd_2}, \|\cdot\|_2) \otimes \mathscr{B}(H^{m+1}(\Omega,\mathbb{R}^{d_2}), \|\cdot\|_{H^{m+1}(\Omega,\mathbb{R}^{d_2})})$, one has $\Pi_{x,e}(\tilde{S}) \in \mathscr{B}(\Omega^n \times \mathbb{R}^{nd_2}, \|\cdot\|_2)$ [see, e.g. Rogers and Williams, 2000, Lemma 11.4, Chapter II]. We conclude that the function J is measurable and so is $\hat{u}_n$. □

Let $B(1,\|\cdot\|_{H^{m+1}(\Omega)}) = \{u \in H^{m+1}(\Omega,\mathbb{R}^{d_2}), \quad \|u\|_{H^{m+1}(\Omega)} \leqslant 1\}$ be the ball of radius r centered at 0. Let $N(B(1,\|\cdot\|_{H^{m+1}(\Omega)})), \|\cdot\|_{H^{m+1}(\Omega)}, r)$ be the minimum number of balls of radius r according to the norm $\|\cdot\|_{H^{m+1}(\Omega)}$ needed to cover the space $B(1,\|\cdot\|_{H^{m+1}(\Omega)})$.

LEMMA C.13 (Entropy of $H^{m+1}(\Omega,\mathbb{R}^{d_2})$). *Let $\Omega \subseteq \mathbb{R}^{d_1}$ be a Lipschitz domain. For $m \geqslant 1$, one has*

$$\log N(B(1,\|\cdot\|_{H^{m+1}(\Omega)}), \|\cdot\|_{H^{m+1}(\Omega)}, r) = \mathcal{O}_{r\to 0}(r^{-d_1/(m+1)}).$$

PROOF. According to the extension theorem [Stein, 1970, Theorem 5, Chapter VI.3.3], there exists a constant $C_\Omega > 0$, depending only on Ω, such that any $u \in H^{m+1}(\Omega,\mathbb{R}^{d_2})$ can be extended to $\tilde{u} \in H^{m+1}(\mathbb{R}^{d_1},\mathbb{R}^{d_2})$, with $\|\tilde{u}\|_{H^{m+1}(\mathbb{R}^{d_1})} \leqslant C_\Omega \|u\|_{H^{m+1}(\Omega)}$. Let $r > 0$ be such that $\Omega \subseteq B(r,\|\cdot\|_2)$ and let $\phi \in C^\infty(\mathbb{R}^{d_1},\mathbb{R}^{d_2})$ be such that

$$\phi(\mathbf{x}) = \begin{cases} 1 \text{ for } \mathbf{x} \in \Omega \\ 0 \text{ for } \mathbf{x} \in \mathbb{R}^{d_1}, |x| \geqslant r. \end{cases}$$

Then, for any $u \in H^{m+1}(\Omega, \mathbb{R}^{d_2})$, (i) $\phi\tilde{u} \in H^{m+1}(\mathbb{R}^{d_1}, \mathbb{R}^{d_2})$, (ii) $\phi\tilde{u}|_\Omega = u$, and (iii) there exists a constant $\tilde{C}_\Omega > 0$ such that $\|\phi\tilde{u}\|_{H^{m+1}(\mathbb{R}^{d_1})} \leqslant \tilde{C}_\Omega \|u\|_{H^{m+1}(\Omega)}$. The lemma follows from Nickl and Pötscher [2007, Corollary 4]. □

LEMMA C.14 (Empirical process L^2). *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be i.i.d. random variables, with common distribution $\mu_{\mathbf{X}}$ on Ω. Then there exists a constant $C_\Omega > 0$, depending only on Ω, such that*

$$\mathbb{E}\Big(\sup_{\|u\|_{H^{m+1}(\Omega)} \leqslant 1} \mathbb{E}\|\tilde{\Pi}(u)(\mathbf{X})\|_2^2 - \frac{1}{n}\sum_{i=1}^n \|\tilde{\Pi}(u)(\mathbf{X}_i)\|_2^2 \Big) \leqslant \frac{d_2^{1/2} C_\Omega}{n^{1/2}},$$

and

$$\mathbb{E}\Big(\Big(\sup_{\|u\|_{H^{m+1}(\Omega)} \leqslant 1} \mathbb{E}\|\tilde{\Pi}(u)(\mathbf{X})\|_2^2 - \frac{1}{n}\sum_{i=1}^n \|\tilde{\Pi}(u)(\mathbf{X}_i)\|_2^2 \Big)^2 \Big) \leqslant \frac{d_2 C_\Omega}{n},$$

where $\tilde{\Pi}$ is the Sobolev embedding (see Theorem B.1).

PROOF. For any $u \in H^{m+1}(\Omega, \mathbb{R}^{d_2})$, let

$$Z_{n,u} = \mathbb{E}\|\tilde{\Pi}(u)(\mathbf{X}_i)\|_2^2 - \frac{1}{n}\sum_{j=1}^n \|\tilde{\Pi}(u)(\mathbf{X}_i)\|_2^2 \quad \text{and} \quad Z_n = \sup_{\|u\|_{H^{m+1}(\Omega)} \leqslant 1} Z_{n,u}.$$

For any $u, v \in H^{m+1}(\Omega, \mathbb{R}^{d_2})$ such that $\|u\|_{H^{m+1}(\Omega)} \leqslant 1$ and $\|v\|_{H^{m+1}(\Omega)} \leqslant 1$, we have

$$\begin{aligned}
&\Big| \frac{1}{n}(\|\tilde{\Pi}(u)(\mathbf{X}_i)\|_2^2 - \mathbb{E}\|\tilde{\Pi}(u)(\mathbf{X}_i)\|_2^2) - \frac{1}{n}(\|\tilde{\Pi}(v)(\mathbf{X}_i)\|_2^2 - \mathbb{E}\|\tilde{\Pi}(v)(\mathbf{X}_i)\|_2^2) \Big| \\
&\quad \leqslant \frac{2}{n}(\|\tilde{\Pi}(u-v)(\mathbf{X}_i)\|_2 + \mathbb{E}\|\tilde{\Pi}(u-v)(\mathbf{X}_i)\|_2) \\
&\quad \leqslant \frac{4C_\Omega}{n}\sqrt{d_2}\|u-v\|_{H^{m+1}(\Omega)} \qquad \text{(by applying Theorem B.1).}
\end{aligned}$$

Therefore, applying Hoeffding's, Azuma's and Dudley's theorem similarly as in the proof of Theorem F.2 shows that

$$\mathbb{E}(Z_n) \leqslant 24 C_\Omega d_2^{1/2} n^{-1} \int_0^\infty [\log N(B(1, \|\cdot\|_{H^{m+1}(\Omega)}), \|\cdot\|_{H^{m+1}(\Omega)}, r)]^{1/2} dr.$$

Lemma C.13 shows that there exists a constant C'_Ω, depending only on Ω, such that $\mathbb{E}(Z_n) \leqslant C'_\Omega d_2^{1/2} n^{-1/2}$. Applying McDiarmid's inequality as in the proof of Theorem F.2 shows that $\mathrm{Var}(Z_n) \leqslant 16 C_\Omega^2 d_2 n^{-1}$. Finally, since $\mathbb{E}(Z_n^2) \leqslant \mathrm{Var}(Z_n) + \mathbb{E}(Z_n)^2$, we deduce that

$$\mathbb{E}(Z_n^2) \leqslant \frac{d_2}{n}\big((C'_\Omega)^2 + 16 C_\Omega^2\big).$$

□

LEMMA C.15 (Empirical process). *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n, \varepsilon_1, \ldots, \varepsilon_n$ be independent random variables, such that $\mathbf{X}_i$ is distributed along $\mu_{\mathbf{X}}$ and ε_i is distributed along μ_ε, such that $\mathbb{E}(\varepsilon) = 0$. Then there exists a constant $C_\Omega > 0$, depending only on Ω, such that*

$$\mathbb{E}\Big(\Big(\sup_{\|u\|_{H^{m+1}(\Omega)} \leqslant 1} \frac{1}{n}\sum_{j=1}^n \langle \tilde{\Pi}(u)(\mathbf{X}_j) - \mathbb{E}(\tilde{\Pi}(u)(\mathbf{X})), \varepsilon_j \rangle \Big)^2 \Big) \leqslant \frac{d_2 \mathbb{E}\|\varepsilon\|_2^2}{n} C_\Omega,$$

where $\tilde{\Pi}$ is the Sobolev embedding.

PROOF. First note, since $H^{m+1}(\Omega,\mathbb{R}^{d_2})$ is separable and since, for all $u\in H^{m+1}(\Omega,\mathbb{R}^{d_2})$, the function $(\mathbf{x}_1,\dots,\mathbf{x}_n,e_1,\dots,e_n)\mapsto \frac{1}{n}\sum_{j=1}^n\langle\tilde{\Pi}(u)(\mathbf{x}_j)-\mathbb{E}(\tilde{\Pi}(u)(\mathbf{X})),e_j\rangle$ is continuous, that the quantity $Z=\sup_{\|u\|_{H^{m+1}(\Omega)}\leqslant 1}\frac{1}{n}\sum_{j=1}^n\langle\tilde{\Pi}(u)(\mathbf{X}_j)-\mathbb{E}(\tilde{\Pi}(u)(\mathbf{X})),\varepsilon_j\rangle$ is a random variable. Moreover, $|Z|\leqslant 2C_\Omega\sqrt{d_2}\sum_{j=1}^n\|\varepsilon_j\|_2/n$, where C_Ω is the constant of Theorem B.1. Thus, $\mathbb{E}(Z^2)<\infty$.

Define, for any $u\in H^{m+1}(\Omega,\mathbb{R}^{d_2})$,

$$Z_{n,u}=\frac{1}{n}\sum_{j=1}^n\langle\tilde{\Pi}(u)(\mathbf{X}_j)-\mathbb{E}(\tilde{\Pi}(u)(\mathbf{X})),\varepsilon_j\rangle\quad\text{and}\quad Z_n=\sup_{\|u\|_{H^{m+1}(\Omega)}\leqslant 1}Z_{n,u}.$$

For any $u,v\in H^{m+1}(\Omega,\mathbb{R}^{d_2})$, we have

$$\begin{aligned}&\Big|\frac{1}{n}\langle\tilde{\Pi}(u)(\mathbf{X}_i)-\mathbb{E}(\tilde{\Pi}(u)(\mathbf{X})),\varepsilon_i\rangle-\frac{1}{n}\langle\tilde{\Pi}(v)(\mathbf{X}_i)-\mathbb{E}(\tilde{\Pi}(u)(\mathbf{X})),\varepsilon_i\rangle\Big|\\&\quad=\frac{1}{n}|\langle\tilde{\Pi}(u-v)(\mathbf{X}_i)-\mathbb{E}(\tilde{\Pi}(u-v)(\mathbf{X})),\varepsilon_i\rangle|\\&\quad\leqslant\frac{2C_\Omega}{n}\sqrt{d_2}\|u-v\|_{H^{m+1}(\Omega)}\|\varepsilon_i\|_2\qquad\text{(by applying Theorem B.1).}\end{aligned}$$

Using that ε is independent of $\mathbf{X}$, so that the conditional expectation of Z_n is indeed a real expectation with $\varepsilon_1,\dots,\varepsilon_n$ fixed, we can apply Hoeffding's, Azuma's and Dudley's theorem similarly as in the proof of Theorem F.2 to show that

$$\begin{aligned}\mathbb{E}(Z_n\mid\varepsilon_1,\dots,\varepsilon_n)&\leqslant\frac{24C_\Omega}{n}\sqrt{d_2}\Big(\sum_{i=1}^n\|\varepsilon_i\|_2^2\Big)^{1/2}\\&\quad\times\int_0^\infty[\log N(B(1,\|\cdot\|_{H^{m+1}(\Omega)}),\|\cdot\|_{H^{m+1}(\Omega)},r)]^{1/2}dr.\end{aligned}$$

Hence, according to Lemma C.13, there exists a constant $C'_\Omega>0$, depending only on Ω, such that $\mathbb{E}(Z_n\mid\varepsilon_1,\dots,\varepsilon_n)\leqslant C'_\Omega n^{-1}\sqrt{d_2}\Big(\sum_{i=1}^n\|\varepsilon_i\|_2^2\Big)^{1/2}$. We deduce that

$$\mathbb{E}(Z_n)\leqslant C'_\Omega\sqrt{d_2}\frac{(\mathbb{E}\|\varepsilon\|_2^2)^{1/2}}{n^{1/2}},$$

and

$$\mathrm{Var}(\mathbb{E}(Z_n\mid\varepsilon_1,\dots,\varepsilon_n))\leqslant\mathbb{E}(\mathbb{E}(Z_n\mid\varepsilon_1,\dots,\varepsilon_n)^2)\leqslant(C'_\Omega)^2d_2\frac{\mathbb{E}\|\varepsilon\|_2^2}{n}.$$

Applying McDiarmid's inequality as in the proof of Theorem F.2 shows that

$$\mathrm{Var}(Z_n\mid\varepsilon_1,\dots,\varepsilon_n)\leqslant 16C_\Omega^2d_2\frac{1}{n^2}\sum_{i=1}^n\|\varepsilon_i\|_2^2.$$

The law of the total variance ensures that

$$\begin{aligned}\mathrm{Var}(Z_n)&=\mathrm{Var}(\mathbb{E}(Z_n\mid\varepsilon_1,\dots,\varepsilon_n))+\mathbb{E}(\mathrm{Var}(Z_n\mid\varepsilon_1,\dots,\varepsilon_n))\\&\leqslant\frac{d_2\mathbb{E}\|\varepsilon\|_2^2}{n}\big((C'_\Omega)^2+16C_\Omega^2\big).\end{aligned}$$

Since $\mathbb{E}(Z_n^2)\leqslant\mathrm{Var}(Z_n)+\mathbb{E}(Z_n)^2$, we deduce that

$$\mathbb{E}(Z_n^2)\leqslant\frac{d_2\mathbb{E}\|\varepsilon\|_2^2}{n}\big(2(C'_\Omega)^2+16C_\Omega^2\big).$$

□

APPENDIX D: PROOFS OF PROPOSITION 2.3

De Ryck et al. [2021, Theorem 5.1] ensures that NN_2 is dense in $(C^\infty([0,1]^{d_1},\mathbb{R}),\|\cdot\|_{C^K([0,1]^{d_1})})$ for all $d_1 \geqslant 1$ and $K \in \mathbb{N}$. Note that the authors state the result for Hölder spaces $(W^{K+1,\infty}([0,1]^{d_1}),\|\cdot\|_{W^{K,\infty}(]0,1[^{d_1})})$ [see Evans, 2010, for a definition]. Clearly, $C^\infty([0,1]^{d_1}) \subseteq W^{K+1,\infty}([0,1]^{d_1})$ and the norms $\|\cdot\|_{C^K}$ and $\|\cdot\|_{W^{K,\infty}}$ coincide on $C^\infty([0,1]^{d_1})$.

Our proof generalizes this result to any bounded Lipschitz domain Ω, to any number $H \geqslant 2$ of layers, and to any output dimension d_2. We stress that for any $U \subseteq \mathbb{R}^{d_1}$, the set $\mathrm{NN}_2 \subseteq C^\infty(\mathbb{R}^{d_1},\mathbb{R}^{d_2})$ can of course be seen as a subset of $C^\infty(U,\mathbb{R}^{d_2})$.

Generalization to any bounded Lipschitz domain Ω. In this and the next paragraph, $d_2 = 1$. Our objective is to prove that NN_2 is dense in $(C^\infty(\bar{\Omega},\mathbb{R}),\|\cdot\|_{C^K(\Omega)})$. Let $f \in C^\infty(\bar{\Omega},\mathbb{R})$. Since Ω is bounded, there exists an affine transformation $\tau : x \mapsto A_\tau x + b_\tau$, with $A_\tau \in \mathbb{R}^\star$ and $b_\tau \in \mathbb{R}^{d_1}$, such that $\tau(\Omega) \subseteq [0,1]^d$. Set $\hat{f} = f(\tau^{-1})$. According to the extension theorem for Lipschitz domains of Stein [1970, Theorem 5 Chapter VI.3.3], the function $\hat{f}$ can be extended to a function $\tilde{f} \in W^{K,\infty}([0,1]^{d_1})$ such that $\tilde{f}|_{\tau(\Omega)} = \hat{f}|_{\tau(\Omega)}$. Fix $\epsilon > 0$. According to De Ryck et al. [2021, Theorem 5.1], there exists $u_\theta \in \mathrm{NN}_2$ such that $\|u_\theta - \hat{f}\|_{W^{K,\infty}([0,1]^d)} \leqslant \epsilon$. Since $\tilde{f}$ is an extension of $\hat{f}$, $\tilde{f}|_{\tau(\Omega)} \in C^\infty(\bar{\Omega})$ and one also has $\|u_\theta - \hat{f}\|_{C^K(\tau(\Omega))} \leqslant \epsilon$.

Now, let $m \in \mathbb{N}$ and let α be a multi-index such that $\sum_{i=1}^{d_1} \alpha_i = m$. Then, clearly, $\partial^\alpha(\hat{f}(\tau)) = A_\tau^m \times \partial^\alpha \hat{f}(\tau)$. Therefore, $\|u_\theta(\tau) - \hat{f}(\tau)\|_{C^K(\Omega)} \leqslant \epsilon \times \max(1, A_\tau^K)$, that is

$$\|u_\theta(\tau) - f\|_{C^K(\Omega)} \leqslant \epsilon \times \max(1, A_\tau^K).$$

But, since τ is affine, $u_\theta(\tau)$ belongs to NN_2. This is the desires result.

Generalization to any number $H \geqslant 2$ of layers. We show in this paragraph that NN_H is dense in $(C^\infty(\bar{\Omega},\mathbb{R}),\|\cdot\|_{C^K(\Omega)})$ for all $H \geqslant 2$. The case $H = 2$ has been treated above and it is therefore assumed that $H \geqslant 3$.

Let $f \in C^\infty(\bar{\Omega},\mathbb{R})$. Introduce the function v defined by

$$v(x_1,\ldots,x_{d_1}) = (\tanh^{\circ(H-2)}(x_1),\ldots,\tanh^{\circ(H-2)}(x_{d_1})),$$

where $\tanh^{\circ(H-2)}$ stands for the $\tanh$ function composed $(H-2)$ times with itself. For all $u_\theta \in \mathrm{NN}_2$, $u_\theta(v) \in \mathrm{NN}_H$ is a neural network such that the first weights matrices $(W_\ell)_{1\leqslant\ell\leqslant H-2}$ are identity matrices and the first offsets $(b_\ell)_{1\leqslant\ell\leqslant H-2}$ are equal to zero. Since $\tanh$ is an increasing C^∞ function, v is a C^∞ diffeomorphism. Therefore, $v(\Omega)$ is a bounded Lipschitz domain and $f(v^{-1}) \in C^\infty(v(\Omega),\mathbb{R})$. Lemma C.2 shows that $f(v^{-1}) \in C^\infty(\bar{v}(\Omega),\mathbb{R})$, where $\bar{v}(\Omega)$ is the closure of $v(\Omega)$. According to the previous paragraph, there exists a sequence $(\theta_m)_{m\in\mathbb{N}}$ of parameters such that $u_{\theta_m} \in \mathrm{NN}_2$ and

$$\lim_{m\to\infty} \|u_{\theta_m} - f(v^{-1})\|_{C^K(v(\Omega))} = 0.$$

Thus, u_{θ_m} approximates $f(v^{-1})$, and we would like $u_{\theta_m}(v)$ to approximate f. From Lemma C.2,

$$\|u_{\theta_m}(v) - f\|_{C^K(\Omega)} \leqslant B_K \times \|u_{\theta_m} - f \circ v^{-1}\|_{C^K(\Omega)} \times (1 + \|\tanh^{\circ H-2}\|_{C^K(\mathbb{R})})^K,$$

while Corollary C.5 asserts that $\|\tanh^{\circ H-2}\|_{C^K(\mathbb{R})} < \infty$. Therefore, we deduce that $\lim_{m\to\infty} \|u_{\theta_m}(v) - f\|_{C^K(\Omega)} = 0$ with $u_{\theta_m}(v) \in \mathrm{NN}_H$, which proves the lemma for $H \geqslant 2$.

Generalization to all output dimension d_2. We have shown so far that for all $H \geqslant 2$, NN_H is dense in $(C^\infty(\bar\Omega,\mathbb{R}),\|\cdot\|_{C^K(\Omega)})$. It remains to establish that NN_H is dense in $(C^\infty(\bar\Omega,\mathbb{R}^{d_2}),\|\cdot\|_{C^K(\Omega)})$ for any output dimension d_2.

Let $f=(f_1,\dots,f_{d_2})\in C^\infty(\Omega,\mathbb{R}^{d_2})$. For all $1\leqslant i\leqslant d_2$, let $(\theta_m^{(i)})_{m\in\mathbb{N}}\in(\mathrm{NN}_H)^{\mathbb{N}}$ be a sequence of neural networks such that $\lim_{m\to\infty}\|u_{\theta_m^{(i)}}-f_i\|_{C^K(\Omega)}=0$. Denote by $u_{\theta_m}=(u_{\theta_m^{(1)}},\dots,u_{\theta_m^{(d_2)}})$ the stacking of these sequences. For all $m\in\mathbb{N}$, $u_{\theta_m}\in\mathrm{NN}_H$ and $\lim_{m\to\infty}\|u_{\theta_m}-f\|_{C^K(\Omega)}=0$. Therefore, NN_H is dense in $(C^\infty(\bar\Omega,\mathbb{R}),\|\cdot\|_{C^K(\Omega)})$.

APPENDIX E: PROOFS OF SECTION 3

E.1. Proof of Proposition 3.1. Consider $u_{\hat\theta(p,n_r,D)}\in\mathrm{NN}_H(D)$, the neural network defined by

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

where $\delta(n,n_r)$ is defined in (12) and where the observations have been reordered according to increasing values of the $\mathbf{X}_{(i)}$. According to Lemma C.6, one has, for all $1\leqslant i\leqslant n$, $\lim_{p\to\infty}u_{\hat\theta(p,n_r,D)}(\mathbf{X}_i)=Y_i$. Moreover, for all order $K\geqslant 1$ of differentiation and all $1\leqslant j\leqslant n_r$, $\lim_{p\to\infty}u^{(K)}_{\hat\theta(p,n_r,D)}(\mathbf{X}_j^{(r)})=0$. Recalling that $\mathscr{F}(u,\mathbf{x})=mu''(\mathbf{x})+\gamma u'(\mathbf{x})$, we have $\|\mathscr{F}(u,\mathbf{x})\|_2\leqslant m\|u''(\mathbf{x})\|_2+\gamma\|u'(\mathbf{x})\|_2$. We therefore conclude that $\lim_{p\to\infty}R_{n,n_r}(u_{\hat\theta(p,n_r,D)})=0$, which is the first statement of the proposition.

Next, using the Cauchy-Schwarz inequality, we have that, for any function $f\in C^2(\mathbb{R})$ and any $\varepsilon>0$,

$$2\varepsilon\int_{-\varepsilon}^{\varepsilon}(mf''+\gamma f')^2\geqslant\left(\int_{-\varepsilon}^{\varepsilon}mf''+\gamma f'\right)^2=\big[m(f'(\varepsilon)-f'(-\varepsilon))+\gamma(f(\varepsilon)-f(-\varepsilon))\big]^2.$$

Thus,

$$\begin{aligned}
&\mathscr{R}_n(u_{\hat\theta(p,n_r,D)})\\
&\quad\geqslant\frac{1}{T}\int_{[0,T]}\mathscr{F}(u_{\hat\theta(p,n_r,D)},\mathbf{x})^2d\mathbf{x}\\
&\quad\geqslant\frac{1}{T}\sum_{i=1}^{n}\int_{\mathbf{X}_{(i)}+\delta(n,n_r)/2-\varepsilon}^{\mathbf{X}_{(i)}+\delta(n,n_r)/2+\varepsilon}\mathscr{F}(u_{\hat\theta(p,n_r,D)},\mathbf{x})^2d\mathbf{x}\\
&\quad\geqslant\frac{1}{T}\sum_{i=1}^{n}\frac{1}{2\varepsilon}\big[m(u'_{\hat\theta(p,n_r,D)}(\mathbf{X}_{(i)}+\delta(n,n_r)/2+\varepsilon)-u'_{\hat\theta(p,n_r,D)}(\mathbf{X}_{(i)}+\delta(n,n_r)/2-\varepsilon))\\
&\qquad+\gamma(u_{\hat\theta(p,n_r,D)}(\mathbf{X}_{(i)}+\delta(n,n_r)/2+\varepsilon)-u_{\hat\theta(p,n_r,D)}(\mathbf{X}_{(i)}+\delta(n,n_r)/2-\varepsilon))\big]^2.
\end{aligned}$$

Observe that, as soon as $\delta(n,n_r)/4>\varepsilon$, one has, for all $1\leqslant i\leqslant n-1$,

$$\lim_{p\to\infty}u_{\hat\theta(p,n_r,D)}(\mathbf{X}_{(i)}+\delta(n,n_r)/2+\varepsilon)-u_{\hat\theta(p,n_r,D)}(\mathbf{X}_{(i)}+\delta(n,n_r)/2-\varepsilon)=Y_{(i+1)}-Y_{(i)},$$

and, for all $1\leqslant i\leqslant n-1$,

$$\lim_{p\to\infty}u'_{\hat\theta(p,n_r,D)}(\mathbf{X}_{(i)}+\delta(n,n_r)/2+\varepsilon)-u'_{\hat\theta(p,n_r,D)}(\mathbf{X}_{(i)}+\delta(n,n_r)/2-\varepsilon)=0.$$

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

$$\sum_{i=1}^{n} \frac{1}{2\varepsilon} \big[m(u'_{\hat{\theta}(p,n_r,D)}(\mathbf{X}_{(i)} + \delta(n,n_r)/2 - \varepsilon) - u'_{\hat{\theta}(p,n_r,D)}(\mathbf{X}_{(i)} + \delta(n,n_r)/2 - \varepsilon)) \\ + \gamma(u_{\hat{\theta}(p,n_r,D)}(\mathbf{X}_{(i)} + \delta(n,n_r)/2 - \varepsilon) - u_{\hat{\theta}(p,n_r,D)}(\mathbf{X}_{(i)} + \delta(n,n_r)/2 - \varepsilon)) \big]^2 \\ \xrightarrow[p\to\infty]{} \gamma \times \frac{\sum_{i=1}^{n-1} (Y_{(i+1)} - Y_{(i)})^2}{2\varepsilon}.$$

We have just proved that, for any $0 < \varepsilon < \delta(n, n_r)/4$, there exists $P \in \mathbb{N}$ such that, for all $p \geqslant P$,

$$\mathcal{R}_n(u_{\hat{\theta}(p,n_r,D)}) \geqslant \gamma \times \frac{\sum_{i=1}^{n-1} (Y_{(i+1)} - Y_{(i)})^2}{2\varepsilon T}.$$

Since we suppose that there exists two observations $Y_{(i)} \neq Y_{(j)}$, we conclude as desired that $\lim_{p\to\infty} \mathcal{R}_n(u_{\hat{\theta}(p,n_r,D)}) = \infty$

E.2. Proof of Proposition 3.2. Let $u_{\hat{\theta}(p,n_e,n_r,D)} \in \mathrm{NN}_H(4)$ be the neural network defined by

$$u_{\hat{\theta}(p,n_e,n_r,D)}(x,t) = \tanh^{\circ H}(x + 0.5 + pt) - \tanh^{\circ H}(x - 0.5 + pt) \\ + \tanh^{\circ H}(0.5 + pt) - \tanh^{\circ H}(1.5 + pt).$$

Clearly, for any $p \in \mathbb{N}$, $u_{\hat{\theta}(p,n_e,n_r,D)}$ satisfies the initial condition

$$u_{\hat{\theta}(p,n_e,n_r,D)}(x,0) = \tanh^{\circ H}(x + 0.5) - \tanh^{\circ H}(x - 0.5) + \tanh^{\circ H}(0.5) - \tanh^{\circ H}(1.5).$$

We are going to prove in the next paragraphs that the derivatives of $u_{\hat{\theta}(p,n_e,n_r,D)}$ vanishe as $p \to \infty$, starting with the temporal derivative and continuing with the spatial ones. According to Lemma C.4, for all $\varepsilon > 0$ and all $x \in [-1, 1]$, $\lim_{p\to\infty} \|u_{\hat{\theta}(p,n_e,n_r,D)}(x,\cdot)\|_{C^2([\varepsilon,T])} = 0$. Therefore, for any $\mathbf{X}_i^{(e)} \in \{-1, 1\} \times [0, T]$, $\lim_{p\to\infty} \|u_{\hat{\theta}(p,n_e,n_r,D)}(\mathbf{X}_i^{(e)})\|_2 = 0$ and, for any $\mathbf{X}_j^{(r)} \in \Omega$, $\lim_{p\to\infty} \|\partial_t u_{\hat{\theta}(p,n_e,n_r,D)}(\mathbf{X}_j^{(r)})\|_2 = 0$ (since $\mathbf{X}_j^{(r)} \notin \partial\Omega$).

Moreover, letting $v(x,t) = \tanh^{\circ H}(x + 0.5 + pt) - \tanh^{\circ H}(x - 0.5 + pt)$, it comes that $\partial^2_{x,x} u_{\hat{\theta}(p,n_e,n_r,D)} = p^{-2} \partial^2_{t,t} v$. Thus, invoking again Lemma C.4, for all $\varepsilon > 0$, and all $x \in [-1, 1]$,

$$\lim_{p\to\infty} p^{-2} \|\partial^2_{t,t} v(x,\cdot)\|_{\infty,[\varepsilon,T]} = \lim_{p\to\infty} \|\partial^2_{x,x} u_{\hat{\theta}(p,n_e,n_r,D)}(x,\cdot)\|_{\infty,[\varepsilon,T]} = 0.$$

Therefore, for any $\mathbf{X}_j^{(r)} \in \Omega$, one has $\lim_{p\to\infty} \|\partial^2_{x,x} u_{\hat{\theta}(p,n_e,n_r,D)}(\mathbf{X}_j^{(r)})\|_2 = 0$ and, in turn, one has $\lim_{p\to\infty} \|\mathscr{F}(u_{\hat{\theta}(p,n_e,n_r,D)}, \mathbf{X}_j^{(r)})\|_2 = 0$. We conclude that, for all $n_e, n_r \geqslant 0$, $\lim_{p\to\infty} R_{n_e,n_r}(u_{\hat{\theta}(p,n_e,n_r,D)}) = 0$.

Next, observe that $\mathscr{R}(u_{\hat{\theta}(p,n_e,n_r,D)}) \geqslant \int_{[-1,1]\times[0,T]} (\partial_t u_{\hat{\theta}(p,n_e,n_r,D)} - \partial^2_{x,x} u_{\hat{\theta}(p,n_e,n_r,D)})^2$. By the Cauchy-Schwarz inequality, for any $\delta > 0$,

$$\int_{[-1,1]\times[0,T]} (\partial_t u_{\hat{\theta}(p,n_e,n_r,D)} - \partial^2_{x,x} u_{\hat{\theta}(p,n_e,n_r,D)})^2 \\ \geqslant \delta^{-1} \int_{x=-1}^{1} \Big(\int_{t=0}^{\delta} \partial_t u_{\hat{\theta}(p,n_e,n_r,D)}(x,t) - \partial^2_{x,x} u_{\hat{\theta}(p,n_e,n_r,D)}(x,t) \Big)^2 dx$$

$$\geqslant \delta^{-1}\int_{x=-1}^{1}\Big(u_{\hat{\theta}(p,n_e,n_r,D)}(x,\delta)-u_{\hat{\theta}(p,n_e,n_r,D)}(x,0)-\int_{t=0}^{\delta}\partial^2_{x,x}u_{\hat{\theta}(p,n_e,n_r,D)}(x,t)dt\Big)^2dx.$$

Invoking again Lemma C.4, we know that $\lim_{p\to\infty}\|u_{\hat{\theta}(p,n_e,n_r,D)}(\cdot,\delta)\|_{[-1,1],\infty}=0$. Moreover, for all $t>0$ and all $-1\leqslant x\leqslant 1$, $\lim_{p\to\infty}\partial^2_{x,x}u_{\hat{\theta}(p,n_e,n_r,D)}(x,t)=0$. Besides, by Corollary C.5, $\|\partial^2_{x,x}u_{\hat{\theta}(p,n_e,n_r,D)}\|_{\infty,[0,1]\times[-1,1]}\leqslant 2\|\tanh^{\circ H}\|_{C^2(\mathbb{R})}<\infty$. Thus, by the dominated convergence theorem, for any $\delta>0$ and all p large enough,

$$\mathscr{R}(u_{\hat{\theta}(p,n_e,n_r,D)})\geqslant\frac{1}{2\delta}\int_{x=-1}^{1}\big(u_{\hat{\theta}(p,n_e,n_r,D)}(x,0)\big)^2dx.$$

Noticing that $u_{\hat{\theta}(p,n_e,n_r,D)}(x,0)$ corresponds to the initial condition, that does not depends on p, we conclude that $\lim_{p\to\infty}\mathscr{R}(u_{\hat{\theta}(p,n_e,n_r,D)})=\infty$.

APPENDIX F: PROOFS OF SECTION 4

F.1. Proof of Proposition 4.2. Recall that each neural network $u_\theta\in\mathrm{NN}_H(D)$ is written as $u_\theta=\mathcal{A}_{H+1}\circ(\tanh\circ\mathcal{A}_H)\circ\cdots\circ(\tanh\circ\mathcal{A}_1)$, where each $\mathcal{A}_k:\mathbb{R}^{L_{k-1}}\to\mathbb{R}^{L_k}$ is an affine function of the form $\mathcal{A}_k(x)=W_kx+b_k$, with W_k a $(L_{k-1}\times L_k)$-matrix, $b_k\in\mathbb{R}^{L_k}$ a vector, $L_0=d_1$, $L_1=\cdots=L_H=D$, $L_{H+1}=d_2$, and $\theta=(W_1,b_1,\ldots,W_{H+1},b_{H+1})\in\mathbb{R}^{\sum_{i=0}^{H}(L_i+1)\times L_i}$. For each $i\in\{1,\ldots,d_1\}$, we let π_i be the projection operator on the ith coordinate, defined by $\pi_i(x_1,\ldots,x_{d_1})=x_i$. Similarly, for a matrix $W=(W_{i,j})_{1\leqslant i\leqslant d_2,1\leqslant j\leqslant d_1}$, we let $\pi_{i,j}(W)=W_{i,j}$ and $\|W\|_\infty=\max_{1\leqslant i\leqslant d_2,1\leqslant j\leqslant d_1}|W_{i,j}|$. Note that $\|W_k\mathbf{x}\|_\infty\leqslant L_{k-1}\|W_k\|_\infty\|\mathbf{x}\|_\infty$. Clearly, $\max_{1\leqslant k\leqslant H+1}(\|W_k\|_\infty,\|b_k\|_\infty)\leqslant\|\theta\|_\infty\leqslant\|\theta\|_2$. Finally, we recursively define the constants $C_{K,H}$ for all $K\geqslant 0$ and all $H\geqslant 1$ by $C_{0,H}=1$, $C_{K,1}=2^{K-1}\times(K+2)!$, and

$$C_{K,H+1}=B_K2^{K-1}(K+2)!\max_{\substack{i_1,\ldots,i_K\in\mathbb{N}\\ i_1+2i_2+\cdots+Ki_K=K}}\prod_{1\leqslant\ell\leqslant K}C_{\ell,H},\tag{13}$$

where B_K is the Kth Bell number, defined in (7).

We prove the proposition by induction on H, starting with the case $H=1$. Clearly, for $H=1$, one has

$$\|u_\theta\|_\infty\leqslant\|W_2\times\tanh\circ\mathcal{A}_1\|_\infty+\|b_2\|_\infty\leqslant\|W_2\|_\infty D+\|b_2\|_\infty\leqslant(D+1)\|\theta\|_2.\tag{14}$$

Next, for any multi-index $\alpha=(\alpha_1,\ldots,\alpha_{d_1})$ such that $|\alpha|\geqslant 1$,

$$\partial^\alpha u_\theta(\mathbf{x})=W_2\begin{pmatrix}\pi_{1,1}(W_1)^{\alpha_1}\times\cdots\times\pi_{1,d_1}(W_1)^{\alpha_{d_1}}\times\tanh^{(|\alpha|)}(\pi_1(\mathcal{A}_1(\mathbf{x})))\\ \vdots\\ \pi_{1,d_1}(W_1)^{\alpha_1}\times\cdots\times\pi_{d_1,d_1}(W_1)^{\alpha_{d_1}}\times\tanh^{(|\alpha|)}(\pi_{d_1}(\mathcal{A}_1(\mathbf{x})))\end{pmatrix}.\tag{15}$$

Upon noting that $|\pi_{1,d_1}(W_1)|\leqslant\|\theta\|_\infty$, we see that

$$\|\partial^\alpha u_\theta\|_\infty\leqslant D\|W_2\|_\infty\|\theta\|_2^{|\alpha|}\|\tanh^{(|\alpha|)}\|_\infty\leqslant D\|\theta\|_2^{1+|\alpha|}\|\tanh^{(|\alpha|)}\|_\infty.\tag{16}$$

Therefore, combining (14) and (16), we deduce that, for any $K\geqslant 1$, $\|u_\theta\|_{C^K(\mathbb{R}^{d_1})}\leqslant(D+1)\max_{k\leq K}\|\tanh^{(k)}\|_\infty(1+\|\theta\|_2)^K\|\theta\|_2$. Applying Lemma C.3, we conclude that, for all $u\in\mathrm{NN}_1(D)$ and for all $K\geqslant 0$,

$$\|u_\theta\|_{C^K(\mathbb{R}^{d_1})}\leqslant C_{K,1}(D+1)(1+\|\theta\|_2)^K\|\theta\|_2.$$

Induction. Assume that for a given $H \geqslant 1$, one has, for any neural network $u_\theta \in \mathrm{NN}_H(D)$ and any $K \geqslant 0$,

$$\|u_\theta\|_{C^K(\mathbb{R}^{d_1})} \leqslant C_{K,H}(D+1)^{1+KH}(1+\|\theta\|_2)^{KH}\|\theta\|_2. \tag{17}$$

Our objective is to show that for any $u_\theta \in \mathrm{NN}_{H+1}(D)$ and any $K \geqslant 0$,

$$\|u_\theta\|_{C^K(\mathbb{R}^{d_1})} \leqslant C_{K,H+1}(D+1)^{1+K(H+1)}(1+\|\theta\|_2)^{K(H+1)}\|\theta\|_2.$$

For such a u_θ, we have, by definition, $u_\theta = \mathcal{A}_{H+2} \circ \tanh \circ v_\theta$, where $v_\theta \in \mathrm{NN}_H(D)$ (by a slight abuse of notation, the parameter of v_θ is in fact $\theta' = (W_1, b_1, \dots, W_{H+1}, b_{H+1})$ while $\theta = (W_1, b_1, \dots, W_{H+2}, b_{H+2})$, so $\|\theta'\|_2 \leqslant \|\theta\|_2$ and $\|\theta'\|_\infty \leqslant \|\theta\|_\infty$). Consequently,

$$\|u_\theta\|_\infty \leqslant \|W_{H+2}\|_\infty D + \|b_{H+2}\|_\infty \leqslant (D+1)\|\theta\|_2. \tag{18}$$

In addition, for any multi-index $\alpha = (\alpha_1, \dots, \alpha_{d_1})$ such that $|\alpha| \geqslant 1$,

$$\partial^\alpha u_\theta(\mathbf{x}) = W_{H+2} \begin{pmatrix} \partial^\alpha(\tanh \circ \pi_1 \circ v_\theta(\mathbf{x})) \\ \vdots \\ \partial^\alpha(\tanh \circ \pi_D \circ v_\theta(\mathbf{x})) \end{pmatrix}.$$

Thus, $\|\partial^\alpha u_\theta\|_\infty \leqslant D\|W_{H+2}\|_\infty \max_{j \leqslant D} \|\tanh \circ \pi_j \circ v_\theta\|_{C^K(\mathbb{R}^{d_1})}$. Invoking identity (8), one has

$$\|\tanh \circ \pi_j \circ v\|_{C^K(\mathbb{R}^{d_1})} \leqslant B_K \|\tanh\|_{C^K(\mathbb{R})} \max_{i_1+2i_2+\cdots+Ki_K=K} \prod_{1 \leqslant \ell \leqslant K} \|\pi_j \circ v_\theta\|^{i_\ell}_{C^\ell(\mathbb{R}^{d_1})}.$$

Observing that $\pi_j \circ v_\theta$ belongs to $\mathrm{NN}_H(D)$, Lemma C.3 and inequality (17) show that

$$\|\tanh \circ \pi_j \circ v_\theta\|_{C^\ell(\mathbb{R}^{d_1})} \leqslant C_{\ell,H+1}(D+1)^{1+\ell H}(1+\|\theta\|_2)^{1+\ell H}\|\theta\|_2.$$

Therefore, $\|\partial^\alpha u_\theta\|_\infty \leqslant C_{K,H+1}(D+1)^{1+KH}(1+\|\theta\|_2)^{K(H+1)}\|\theta\|_2$, which concludes the induction.

To complete the proof, it remains to show that the exponent of $\|\theta\|_2$ is optimal. To this aim, we let $d_1 = d_2 = 1$, $D = 1$. For each $H \geqslant 1$, we consider the sequence $(\theta_m^{(H)})_{m \in \mathbb{N}}$ defined by $\theta_m^{(H)} = (W_1^{(m)}, b_1^{(m)}, \dots, W_{H+1}^{(m)}, b_{H+1}^{(m)})$, with $W_i^m = m$ and $b_i^m = 0$. Then, for all $\theta = (W_1, b_1, \dots, W_{H+1}, b_{H+1}) \in \Theta_{H,1}$, the associated neural network's derivatives satisfy

$$\|u_\theta^{(k)}\|_\infty = \|(\tanh^{\circ H})^{(K)}\|_\infty |W_{H+1}| \prod_{i=1}^{H} |W_i|^K.$$

Next, since $\|\theta_m^{(H)}\|_2 = m\sqrt{H+1}$, we have

$$\|u_{\theta_m^{(H)}}\|_{C^K(\mathbb{R}^{d_1})} \geqslant \left\|u^{(K)}_{\theta_m^{(H)}}\right\|_\infty \geqslant \left\|(\tanh^{\circ H})^{(K)}\right\|_\infty m^{1+HK} \geqslant \bar{C}(H,K)\|\theta_m^{(H)}\|_2^{1+HK},$$

where $\bar{C}(H,K) = (H+1)^{-(1+HK)/2}\|(\tanh^{\circ H})^{(K)}\|_\infty$. Since $\lim_{m\to\infty} \|\theta_m^{(H)}\|_2 = \infty$, we conclude that the bound of inequality (17) is tight.

F.2. Lipschitz dependence of the Hölder norm in the NN parameters.

PROPOSITION F.1 (Lipschitz dependence of the Hölder norm in the NN parameters). *Consider the class* $\mathrm{NN}_H(D) = \{u_\theta, \theta \in \Theta_{H,D}\}$. *Let* $K \in \mathbb{N}$. *Then there exists a constant* $\tilde{C}_{K,H} > 0$, *depending only on* K *and* H, *such that, for all* $\theta, \theta' \in \Theta_{H,D}$,

$$\|u_\theta - u_{\theta'}\|_{C^K(\Omega)} \leqslant \tilde{C}_{K,H}(1+d_1 M(\Omega))(D+1)^{H+KH^2}(1+\|\theta\|_2)^{H+KH^2}\|\theta-\theta'\|_2,$$

where $M(\Omega) = \sup_{\mathbf{x}\in\Omega} \|\mathbf{x}\|_\infty$.

PROOF. We recursively define the constants $\tilde{C}_{K,H}$ for all $K \geqslant 0$ and all $H \geqslant 1$ by $\tilde{C}_{K,1} = (K+2)2^{2K-1}(K+2)!(K+3)!$, and

$$\tilde{C}_{K,H+1} = C_{K,H+1}[1 + (K+1)B_K 2^{2K-1}(K+3)!(K+2)!\tilde{C}_{K,H}].$$

Recall that π_i is the projection operator on the ith coordinate, defined by $\pi_i(x_1, \dots, x_{d_1}) = x_i$. Before embarking on the proof, observe that by identity (8), we have, for all $u_1, u_2 \in C^K(\Omega, \mathbb{R}^D)$, for all $1 \leqslant i \leqslant D$,

$$\begin{aligned}\partial^\alpha(\tanh \circ \pi_i \circ u_1 - \tanh \circ \pi_i \circ u_2) = &\sum_{P \in \Pi(K)} [\tanh^{(|P|)} \circ \pi_i \circ u_1] \prod_{S \in P} \partial^{\alpha(S)}(\pi_i \circ u_1) \\ &- [\tanh^{(|P|)} \circ \pi_i \circ u_2] \prod_{S \in P} \partial^{\alpha(S)}(\pi_i \circ u_2).\end{aligned}$$

In addition, for two sequences $(a_i)_{1 \leqslant i \leqslant n}$ and $(b_i)_{1 \leqslant i \leqslant n}$,

(19)
$$\prod_{i=1}^n a_i - \prod_{i=1}^n b_i = \sum_{i=1}^n (a_i - b_i)\Big(\prod_{j=i+1}^n a_j\Big)\Big(\prod_{j=1}^{i-1} b_j\Big) \leqslant n \max_{1 \leqslant i \leqslant n}\{|a_i - b_i|\} \prod_{i=1}^n \max(|a_i|, |b_i|).$$

Observe that for any $1 \leqslant i \leqslant d_2$ and $P \in \Pi(K)$, $[\tanh^{(|P|)} \circ \pi_i \circ u_1] \prod_{S \in P} \partial^{\alpha(S)}(\pi_i \circ u_1) - [\tanh^{(|P|)} \circ \pi_i \circ u_2] \prod_{S \in P} \partial^{\alpha(S)}(\pi_i \circ u_2)$ is the difference of two products of $|P|+1$ terms to which we can apply (19). So,

$$\begin{aligned}&\Big\|[\tanh^{(|\pi|)} \circ \pi_i \circ u_1] \prod_{S \in P} \partial^{\alpha(S)}(\pi_i \circ u_1) - [\tanh^{(|\pi|)} \circ \pi_i \circ u_2] \prod_{S \in \pi} \partial^{\alpha(S)}(\pi_i \circ u_2)\Big\|_{\infty,\Omega} \\ &\quad \leqslant (|P|+1)\big(\|\tanh^{(|P|)}\|_{\mathrm{Lip}} \|u_1 - u_2\|_{\infty,\Omega} + \|u_1 - u_2\|_{C^K(\Omega)}\big) \\ &\quad \times \|\tanh^{(|P|)}\|_\infty \prod_{S \in P} \max(\|\partial^{\alpha(S)} u_1\|_{\infty,\Omega}, \|\partial^{\alpha(S)} u_2\|_{\infty,\Omega}). \qquad (20)\end{aligned}$$

Notice finally that $\|\tanh^{(|P|)}\|_{\mathrm{Lip}} = \|\tanh^{(|P|+1)}\|_\infty$.

With the preliminary results out of the way, we are now equipped to prove the statement of the proposition, by induction on H. Assume first that $H = 1$. We start by examining the case $K = 0$ and then generalize to all $K \geqslant 1$. Let $u_\theta = \mathcal{A}_2 \circ \tanh \circ \mathcal{A}_1$ and $u_{\theta'} = \mathcal{A}_2' \circ \tanh \circ \mathcal{A}_1'$. Notice that

$$\|\mathcal{A}_1 - \mathcal{A}_1'\|_{\infty,\Omega} \leqslant \|b_1 - b_1'\|_\infty + d_1 M(\Omega)\|W_1 - W_1'\|_\infty \leqslant \|\theta - \theta'\|_2(1 + d_1 M(\Omega)),$$

where $M(\Omega) = \max_{\mathbf{x} \in \Omega} \|\mathbf{x}\|_\infty$. Since $\|\tanh\|_{\mathrm{Lip}} = 1$, we deduce that $\|\tanh \circ \mathcal{A}_1 - \tanh \circ \mathcal{A}_1'\|_\infty \leqslant \|\theta - \theta'\|_2(1 + d_1 M(\Omega))$. Similarly, $\|\mathcal{A}_2 - \mathcal{A}_2'\|_{\infty, B(1,\|\cdot\|_\infty)} \leqslant \|\theta - \theta'\|_2(1 + D)$. Next,

$$\begin{aligned}\|u_\theta - u_{\theta'}\|_{\infty,\Omega} &\leqslant \|(\mathcal{A}_2 - \mathcal{A}_2') \circ \tanh \circ \mathcal{A}_1\|_{\infty,\Omega} + \|\mathcal{A}_2' \circ \tanh \circ \mathcal{A}_1 - \mathcal{A}_2' \circ \tanh \circ \mathcal{A}_1'\|_{\infty,\Omega} \\ &\leqslant \|\mathcal{A}_2 - \mathcal{A}_2'\|_{\infty, B(1,\|\cdot\|_\infty)} + D\|W_2'\|_\infty \|\tanh \circ \mathcal{A}_1 - \tanh \circ \mathcal{A}_1'\|_{\infty,\Omega} \\ &\leqslant \|\theta - \theta'\|_2(1 + D + D\|\theta'\|_2(1 + d_1 M(\Omega))) \\ &\leqslant \tilde{C}_{0,1}(1 + d_1 M(\Omega))(D+1)(1 + \max(\|\theta\|_2, \|\theta'\|_2))\|\theta - \theta'\|_2.\end{aligned}$$

This shows the result for $H = 1$ and $K = 0$. Assume now that $K \geqslant 1$, and let α be a multi-index such that $|\alpha| = K$. Observe that

$$\begin{aligned}\|\partial^\alpha(u_\theta - u_{\theta'})\|_{\infty,\Omega} \leqslant &\|(W_2 - W_2')\partial^\alpha(\tanh \circ \mathcal{A}_1)\|_{\infty,\Omega} \\ &+ \|W_2' \partial^\alpha(\tanh \circ \mathcal{A}_1 - \tanh \circ \mathcal{A}_1')\|_{\infty,\Omega}. \qquad (21)\end{aligned}$$

By Lemma C.3 and an argument similar to the inequality (15), we have

$$\begin{aligned}\|(W_2 - W_2')\partial^\alpha(\tanh\circ\mathcal{A}_1)\|_{\infty,\Omega} &\leqslant (D+1)\|\theta-\theta'\|_2\|\theta\|_2^K\|\tanh\|_{C^K(\mathbb{R})}\\ &\leqslant 2^{K-1}(K+2)!(D+1)\|\theta-\theta'\|_2\|\theta\|_2^K. \end{aligned} \tag{22}$$

In order to bound the second term on the right-hand side of (21), we use inequality (20) with $u_1 = \mathcal{A}_1$ and $u_2 = \mathcal{A}_1'$. In this case, the only non-zero term on the right-hand side of (20) corresponds to the partition $\pi = \{\{1\},\{2\},\ldots,\{K\}\}$. Recall that $\|\mathcal{A}_1 - \mathcal{A}_1'\|_{\infty,\Omega} \leqslant \|\theta-\theta'\|_2(1+d_1M(\Omega))$, and note that whenever $|\alpha|=1$, $\|\partial^\alpha(\mathcal{A}_1-\mathcal{A}_1')\|_{\infty,\Omega}\leqslant\|\theta-\theta'\|_2$. Therefore, $\|\mathcal{A}_1-\mathcal{A}_1'\|_{C^K(\Omega)} = \|\mathcal{A}_1-\mathcal{A}_1'\|_{C^1(\Omega)} \leqslant \|\theta-\theta'\|_2(1+d_1M(\Omega))$. Observe, in addition, that $\prod_{B\in\{\{1\},\{2\},\ldots,\{K\}\}}\max(\|\partial^{\alpha(B)}\mathcal{A}_1\|_{\infty,\Omega}, \|\partial^{\alpha(B)}\mathcal{A}_1'\|_{\infty,\Omega}) \leqslant \max(\|\theta\|_2,\|\theta'\|_2)^K$. Thus, putting all the pieces together, we are led to

$$\begin{aligned}&\|\partial^\alpha(\tanh\circ\mathcal{A}_1 - \tanh\circ\mathcal{A}_1')\|_{\infty,\Omega}\\ &\quad\leqslant (K+1)\|\tanh^{(K+1)}\|_\infty\|\theta-\theta'\|_2(1+d_1M(\Omega))\|\tanh^{(K)}\|_\infty\max(\|\theta\|_2,\|\theta'\|_2)^K.\end{aligned}$$

Now, by Lemma C.3, $\|\tanh^{(K)}\|_\infty \leqslant 2^{K-1}(K+2)!$ So,

$$\begin{aligned}&\|\partial^\alpha(\tanh\circ\mathcal{A}_1 - \tanh\circ\mathcal{A}_1')\|_{\infty,\Omega}\\ &\quad\leqslant (K+1)2^{2K-1}(K+2)!(K+3)!\|\theta-\theta'\|_2(1+d_1M(\Omega))\max(\|\theta\|_2,\|\theta'\|_2)^K.\end{aligned} \tag{23}$$

Combining inequalities (21), (22), and (23), we conclude that

$$\|\partial^\alpha(u_\theta - u_{\theta'})\|_{\infty,\Omega} \leqslant \tilde{C}_{K,1}(1+d_1M(\Omega))(D+1)(1+\max(\|\theta\|_2,\|\theta'\|_2))^{K+1}\|\theta-\theta'\|_2,$$

so that $\|u_\theta - u_{\theta'}\|_{C^K(\Omega)} \leqslant \tilde{C}_{K,1}(1+d_1M(\Omega))(D+1)(1+\max(\|\theta\|_2,\|\theta'\|_2))^{K+1}\|\theta-\theta'\|_2$.

Induction. Fix $H \geqslant 1$, and assume that for all $u_\theta, u_{\theta'} \in \mathrm{NN}_H(D)$ and all $K \geqslant 0$,

$$\begin{aligned}&\|u_\theta - u_{\theta'}\|_{C^K(\Omega)}\\ &\quad\leqslant \tilde{C}_{K,H}(1+d_1M(\Omega))(D+1)^{H+KH^2}(1+\max(\|\theta\|_2,\|\theta'\|_2))^{H+KH^2}\|\theta-\theta'\|_2.\end{aligned} \tag{24}$$

Let $u_\theta, u_{\theta'} \in \mathrm{NN}_{H+1}(D)$. Observe that $u_\theta = \mathcal{A}_{H+2}\circ\tanh\circ v_\theta$ and $u_{\theta'} = \mathcal{A}_{H+2}'\circ\tanh\circ v_{\theta'}$, where $v_\theta, v_{\theta'} \in \mathrm{NN}_H(D)$. Moreover,

$$\begin{aligned}&\|\partial^\alpha(u_\theta - u_{\theta'})\|_{\infty,\Omega}\\ &\quad\leqslant \|(W_{H+2}-W_{H+2}')\partial^\alpha(\tanh\circ v_\theta)\|_{\infty,\Omega} + \|W_{H+2}'\partial^\alpha(\tanh\circ v_\theta - \tanh\circ v_{\theta'})\|_{\infty,\Omega}\\ &\quad\leqslant D(\|\theta-\theta'\|_2\times\|\partial^\alpha(\tanh\circ v_\theta)\|_{\infty,\Omega} + \|\theta'\|_2\times\|\partial^\alpha(\tanh\circ v_\theta - \tanh\circ v_{\theta'})\|_{\infty,\Omega}).\end{aligned} \tag{25}$$

Since $\tanh\circ v_\theta \in \mathrm{NN}_{H+1}(D)$, we have, by Proposition 4.2,

$$\|\partial^\alpha(\tanh\circ v_\theta)\|_{\infty,\Omega} \leqslant C_{K,H+1}(D+1)^{1+K(H+1)}(1+\|\theta\|_2)^{K(H+1)}\|\theta\|_2. \tag{26}$$

Moreover, using (20), Lemma C.3, and the definition of $C_{K,H+1}$ in (13), we have

$$\begin{aligned}&\|\partial^\alpha(\tanh\circ v_\theta - \tanh\circ v_{\theta'})\|_{\infty,\Omega}\\ &\quad\leqslant B_K(K+1)\|\tanh^{(K+1)}\|_\infty\|v_\theta - v_{\theta'}\|_{C^K(\Omega)}\|\tanh^{(K)}\|_\infty\\ &\qquad\times C_{K,H+1}(D+1)^{KH}(1+\max(\|\theta\|_2,\|\theta'\|_2))^{KH}\\ &\quad\leqslant 2^{2K-1}(K+3)!(K+2)!B_K(K+1)\|v_\theta - v_{\theta'}\|_{C^K(\Omega)}\\ &\qquad\times C_{K,H+1}(D+1)^{KH}(1+\max(\|\theta\|_2,\|\theta'\|_2))^{KH}.\end{aligned} \tag{27}$$

The term $\|v_\theta - v_{\theta'}\|_{C^K(\Omega)}$ in (27) can be upper bounded using the induction assumption (24). Thus, combining (25), (26), and (27), we conclude as desired that for all $u_\theta, u_{\theta'} \in \mathrm{NN}_{H+1}(D)$ and all $K \in \mathbb{N}$,

$$\|u_\theta - u_{\theta'}\|_{C^K(\Omega)} \leqslant \tilde{C}_{K,H+1}(1 + d_1 M(\Omega))(D+1)^{(H+1)+K(H+1)^2} \times (1 + \max(\|\theta\|_2, \|\theta'\|_2))^{(H+1)+K(H+1)^2} \|\theta - \theta'\|_2.$$

□

F.3. Uniform approximation of integrals. Throughout this section, the parameters $H, D \in \mathbb{N}^\star$ are held fixed, as well as the neural architecture $\mathrm{NN}_H(D)$ parameterized by $\Theta_{H,D}$. We let d be a metric in $\Theta_{H,D}$, and denote by $B(r,d)$ the closed ball in $\Theta_{H,D}$ centered at 0 and of radius r according to the metric d, that is, $B(r,d) = \{\theta \in \Theta_{H,D},\ d(0,\theta) \leqslant r\}$.

THEOREM F.2 (Uniform approximation of integrals). *Let $\Omega \subseteq \mathbb{R}^{d_1}$ be a bounded Lipschitz domain, let $\alpha_1 > 0$, and let $\mathbf{X}_1, \dots, \mathbf{X}_n$ be a sequence of i.i.d. random variables in $\bar{\Omega}$, with distribution μ_X. Let $f : C^\infty(\bar{\Omega}, \mathbb{R}^{d_2}) \times \bar{\Omega} \to \mathbb{R}^{d_2}$ be an operator, and assume that the following two requirements are satisfied:*

(i) there exist $C_1 > 0$ and $\beta_1 \in [0, 1/2[$ such that, for all $n \geqslant 1$ and all $\theta, \theta' \in B(n^{\alpha_1}, \|.\|_2)$,

$$\|f(u_\theta, \cdot) - f(u_{\theta'}, \cdot)\|_{\infty,\bar{\Omega}} \leqslant C_1 n^{\beta_1} \|\theta - \theta'\|_2; \tag{28}$$

(ii) there exist $C_2 > 0$ and $\beta_2 \in [0, 1/2[$ satisfying $\beta_2 > \alpha_1 + \beta_1$ such that, for all $n \geqslant 1$ and all $\theta \in B(n^{\alpha_1}, \|.\|_2)$,

$$\|f(u_\theta, \cdot)\|_{\infty,\bar{\Omega}} \leqslant C_2 n^{\beta_2}. \tag{29}$$

Then, almost surely, there exists $N \in \mathbb{N}^\star$ such that, for all $n \geqslant N$,

$$\sup_{\theta \in B(n^{\alpha_1}, \|.\|_2)} \left\| \frac{1}{n} \sum_{i=1}^n f(u_\theta, \mathbf{X}_i) - \int_{\bar{\Omega}} f(u_\theta, \cdot) d\mu_X \right\|_2 \leqslant \log^2(n) n^{\beta_2 - 1/2}.$$

(Notice that the rank N is random.)

PROOF. Let us start the proof by considering the case $d_2 = 1$. For a given $\theta \in B(n^{\alpha_1}, \| \cdot \|_2)$, we let

$$Z_{n,\theta} = \frac{1}{n} \sum_{i=1}^n f(u_\theta, \mathbf{X}_i) - \int_{\bar{\Omega}} f(u_\theta, \cdot) d\mu_X.$$

We are interested in bounding the random variable

$$Z_n = \sup_{\theta \in B(n^{\alpha_1}, \|\cdot\|_2)} |Z_{n,\theta}| = \sup_{\theta \in B(n^{\alpha_1}, \|\cdot\|_2)} Z_{n,\theta}.$$

Note that there is no need of absolute value in the rightmost term since, for any $\theta = (W_1, b_1, \dots, W_{H+1}, b_{H+1}) \in B(n^{\alpha_1}, \| \cdot \|_2)$, it is clear that $\theta' = (W_1, b_1, \dots, W_H, b_H, -W_{H+1}, -b_{H+1}) \in B(n^{\alpha_1}, \| \cdot \|_2)$ and $u_{\theta'} = -u_\theta$. Let $M(\Omega) = \max_{\mathbf{x} \in \bar{\Omega}} \|x\|_2$. Using inequality (28), we have, for any $\theta, \theta' \in B(n^{\alpha_1}, \| \cdot \|_2)$,

$$\left| \frac{1}{n} \Big(f(u_\theta, \mathbf{X}_i) - \int_{\bar{\Omega}} f(u_\theta, \cdot) d\mu_X \Big) - \frac{1}{n} \Big(f(u'_\theta, \mathbf{X}_i) - \int_{\bar{\Omega}} f(u'_\theta, \cdot) d\mu_X \Big) \right| \leqslant 2 C_1 n^{\beta_1 - 1} \|\theta - \theta'\|_2.$$

Thus, according to Hoeffding's theorem [van Handel, 2016, Lemma 3.6] the random variable $n^{-1}(f(u_\theta, \mathbf{X}_i) - \int_{\bar{\Omega}} f(u_\theta, \cdot) d\mu_X) - n^{-1}(f(u'_\theta, \mathbf{X}_i) - \int_{\bar{\Omega}} f(u'_\theta, \cdot) d\mu_X)$ is subgaussian with

parameter $4C_1^2 n^{2\beta_1-2}\|\theta-\theta'\|_2^2$. Invoking Azuma's theorem [van Handel, 2016, Lemma 3.7], we deduce that $Z_{n,\theta}-Z_{n,\theta'}$, is also subgaussian, with parameter $4C_1^2 n^{2\beta_1-1}\|\theta-\theta'\|_2^2$. Since $\mathbb{E}(Z_{n,\theta})=0$, we conclude that for all $n\geqslant 1$, $(Z_{n,\theta})_{\theta\in B(n^{\alpha_1},\|\cdot\|_2)}$ is a subgaussian process on $B(n^{\alpha_1},\|\cdot\|_2)$ for the metric $d(\theta,\theta')=2C_1 n^{\beta_1-1/2}\|\theta-\theta'\|_2$. Moreover, since $\theta\mapsto Z_{n,\theta}$ is continuous for the topology induced by the metric d, $(Z_{n,\theta})_{\theta\in B(n^{\alpha_1},\|\cdot\|_2)}$ is separable [van Handel, 2016, Remark 5.23]. Thus, by Dudley's theorem [van Handel, 2016, Corollary 5.25]

$$\mathbb{E}(Z_n)\leqslant 12\int_0^\infty [\log N(B(n^{\alpha_1},\|\cdot\|_2),d,r)]^{1/2}dr,$$

where $N(B(n^{\alpha_1},\|\cdot\|_2),d,r)$ is the minimum number of balls of radius r according to the metric d needed to cover the space $B(n^{\alpha_1},\|\cdot\|_2)$. Clearly, $N(B(n^{\alpha_1},\|\cdot\|_2),\ d,\ r)=N(B(n^{\alpha_1},\|\cdot\|_2),\ \|\cdot\|_2,\ n^{1/2-\beta_1}r/(2C_1))$. Thus,

$$\mathbb{E}(Z_n)\leqslant 24C_1 n^{\beta_1-1/2}\int_0^\infty [\log N(B(n^{\alpha_1},\|\cdot\|_2),\|\cdot\|_2,r)]^{1/2}dr$$

and, in turn,

$$\mathbb{E}(Z_n)\leqslant 24C_1 n^{\alpha_1+\beta_1-1/2}\int_0^\infty [\log N(B(1,\|\cdot\|_2),\ \|\cdot\|_2,r)]^{1/2}dr.$$

Upon noting that $N(B(1,\|\cdot\|_2),\ \|\cdot\|_2,r)=1$ for $r\geqslant 1$, we are led to

$$\mathbb{E}(Z_n)\leqslant 24C_1 n^{\alpha_1+\beta_1-1/2}\int_0^1 [\log N(B(1,\|\cdot\|_2),\ \|\cdot\|_2,r)]^{1/2}dr.$$

Since $\Theta_{H,D}=\mathbb{R}^{(d_1+1)D+(H-1)D(D+1)+(D+1)d_2}$, according to van Handel [2016, Lemma 5.13], one has

$$\log N(B(1,\|\cdot\|_2),\ \|\cdot\|_2,r)\leqslant[(d_1+1)D+(H-1)D(D+1)+(D+1)d_2]\log(3/r).$$

Notice that $\int_0^1\log(3/r)^{1/2}dr\leqslant 3/2$. Therefore,

$$\mathbb{E}(Z_n)\leqslant 36C_1[(d_1+1)D+(H-1)D(D+1)+(D+1)d_2]^{1/2}n^{\alpha_1+\beta_1-1/2}. \tag{30}$$

Next, observe that, by definition of $Z_n=Z_n(\mathbf{X}_1,\ldots,\mathbf{X}_n)$,

$$\begin{aligned}
&\sup_{\mathbf{x}_i\in\mathbb{R}^{d_1}} Z_n(\mathbf{X}_1,\ldots,\mathbf{X}_{i-1},\mathbf{x}_i,\mathbf{X}_{i+1},\ldots,\mathbf{X}_n)-\inf_{\mathbf{x}_i\in\mathbb{R}^{d_1}} Z_n(\mathbf{X}_1,\ldots,\mathbf{X}_{i-1},\mathbf{x}_i,\mathbf{X}_{i+1},\ldots,\mathbf{X}_n)\\
&\quad\leqslant 2n^{-1}\sup_{\theta\in B(n^{\alpha_1},\|\cdot\|_2)}\left\|f(u_\theta,\mathbf{X}_i)-\int_{\bar\Omega}f(u_\theta,\cdot)d\mu_X\right\|_2\\
&\quad\leqslant 4n^{-1}\sup_{\theta\in B(n^{\alpha_1},\|\cdot\|_2)}\|f(u_\theta,\cdot)\|_\infty.
\end{aligned}$$

Using inequality (29), McDiarmid's inequality [van Handel, 2016, Theorem 3.11] ensures that Z_n is subgaussian with parameter $4C_2^2 n^{2\beta_2-1}$. In particular, for all $t_n\geqslant 0$, $\mathbb{P}(|Z_n-\mathbb{E}(Z_n)|\geqslant t_n)\leqslant 2\exp(-n^{1-2\beta_2}t_n^2/(8C_2^2))$, which is summable with $t_n=C_3 n^{\beta_2-1/2}\log^2(n)$, where C_3 is any positive constant. Thus, recalling that $\beta_2>\alpha_1+\beta_1$, the Borel-Cantelli lemma and (30) ensure that, almost surely, for all n large enough, $0\leqslant Z_n\leqslant 2C_3 n^{\beta_2-1/2}\log^2(n)$. Taking $C_3=1/2$ yields the desired result.

The generalization to the case $d_2\geqslant 2$ is easy. Just note, letting $f=(f_1,\ldots,f_{d_2})$, that

$$\begin{aligned}
&\sup_{\theta\in B(n^{\alpha_1},\|\cdot\|_2)}\left\|\frac{1}{n}\sum_{i=1}^n f(u_\theta,\mathbf{X}_i)-\int_{\bar\Omega}f(u_\theta,\cdot)d\mu_X\right\|_2\\
&\qquad\leqslant\sqrt{d_2}\max_{1\leqslant j\leqslant d_2}\sup_{\theta\in B(n^{\alpha_1},\|\cdot\|_2)}\left\|\frac{1}{n}\sum_{i=1}^n f_j(u_\theta,\mathbf{X}_i)-\int_{\bar\Omega}f_j(u_\theta,\cdot)d\mu_X\right\|_2.
\end{aligned}$$

Taking $C_3 = d_2^{-1/2}/2$ as above leads to the result. □

PROPOSITION F.3 (Condition function). *Let Ω be a bounded Lipschitz domain, let E be a closed subset of $\partial\Omega$, and let $h \in \mathrm{Lip}(E, \mathbb{R}^{d_2})$. Then the operator $\mathscr{H}(u, \mathbf{x}) = \mathbf{1}_{\mathbf{x}\in E}\|u(\mathbf{x}) - h(\mathbf{x})\|^2$ satisfies inequalities (28) and (29) with $\alpha_1 < (3+H)^{-1}/2$, $\beta_1 = (1+H)\alpha_1$, and $1/2 > \beta_2 \geqslant (3+H)\alpha_1$.*

PROOF. First note, since $\mathrm{Lip}(E, \mathbb{R}^{d_2}) \subseteq C^0(E, \mathbb{R}^{d_2})$, that $\|h\|_\infty < \infty$. Observe also that for any $v, w \in \mathbb{R}^{d_2}$, $|\|v\|_2^2 - \|w\|_2^2| = |\langle v+w, v-w\rangle| \leqslant \|v+w\|_2\|v-w\|_2 \leqslant d_2\|v+w\|_\infty\|v-w\|_\infty$, where $\langle\cdot,\cdot\rangle$ denotes the canonical scalar product. Thus, we obtain, for all $\theta, \theta' \in B(n^{\alpha_1}, \|\cdot\|_2)$ and all $\mathbf{x} \in E$,

$$
\begin{aligned}
|\mathscr{H}(u_\theta, \mathbf{x}) - \mathscr{H}(u_{\theta'}, \mathbf{x})| &\leqslant (\|u_\theta(\mathbf{x})\|_2 + \|u_{\theta'}(\mathbf{x})\|_2 + 2\|h(\mathbf{x})\|_2)\|u_\theta(\mathbf{x}) - u_{\theta'}(\mathbf{x})\|_2 \\
&\leqslant d_2(\|u_\theta\|_{\infty,\bar\Omega} + \|u_{\theta'}\|_{\infty,\bar\Omega} + 2\|h\|_\infty)\|u_\theta - u_{\theta'}\|_{\infty,\bar\Omega} \\
&\leqslant d_2(2(D+1)n^{\alpha_1} + 2\|h\|_\infty)\|u_\theta - u_{\theta'}\|_{\infty,\bar\Omega} \text{ (by inequality (18))} \\
&\leqslant 2d_2((D+1)n^{\alpha_1} + \|h\|_\infty)\tilde C_{0,H}(1 + d_1 M(\Omega)) \\
&\qquad \times (D+1)^H(1+n^{\alpha_1})^H\|\theta - \theta'\|_2 \quad \text{(by Proposition F.1)} \\
&\leqslant C_1 n^{\beta_1}\|\theta - \theta'\|_2,
\end{aligned}
$$

where $\beta_1 = (1+H)\alpha_1$ and $C_1 = 2^{H+1}d_2(D+1+\|h\|_\infty)\tilde C_{0,H}(1+d_1M(\Omega))(D+1)^H$.

Next, using (18) once again, for all $\theta \in B(n^{\alpha_1}, \|.\|_2)$, $\|\mathscr{H}(u_\theta, \cdot)\|_{\infty,\bar\Omega} \leqslant d_2(\|u_\theta\|_{\infty,\bar\Omega} + \|h\|_\infty)^2 \leqslant d_2((D+1)n^{\alpha_1} + \|h\|_\infty)^2 \leqslant C_2 n^{2\alpha_1}$. Recall that for inequality (29), β_2 must satisfy $\alpha_1 + \beta_1 < \beta_2 < 1/2$. This is true for $\beta_2 = (3+H)\alpha_1$, which completes the proof. □

PROPOSITION F.4 (Polynomial operator). *Let Ω be a bounded Lipschitz domain, and let $\mathscr{F} \in \mathscr{P}_{\mathrm{op}}$. Then the operator $\mathbf{1}_{\mathbf{x}\in\Omega}\mathscr{F}(u_\theta, \mathbf{x})^2$ satisfies inequalities (28) and (29) with $\alpha_1 < [2 + H(1 + (2+H)\deg(\mathscr{F}))]^{-1}/2$, $\beta_1 = H(1+(2+H)\deg(\mathscr{F}))\alpha_1$, and $1/2 > \beta_2 \geqslant [2 + H(1+(2+H)\deg(\mathscr{F}))]\alpha_1$.*

PROOF. Let $\mathscr{F} \in \mathscr{P}_{\mathrm{op}}$. By definition, there exist a degree $s \geqslant 1$, a polynomial $P \in C^\infty(\mathbb{R}^{d_1}, \mathbb{R})[Z_{1,1}, \ldots, Z_{d_2,s}]$, and a sequence $(\alpha_{i,j})_{1\leqslant i\leqslant d_2, 1\leqslant j\leqslant s}$ of multi-indexes such that, for any $u \in C^\infty(\bar\Omega, \mathbb{R}^{d_2})$, $\mathscr{F}(u,\cdot) = P((\partial^{\alpha_{i,j}}u_i)_{1\leqslant i\leqslant d_2, 1\leqslant j\leqslant s})$. In other words, there exists $N(P) \in \mathbb{N}^\star$, exponents $I(i,j,k) \in \mathbb{N}$, and functions $\phi_1, \ldots, \phi_{N(P)} \in C^\infty(\bar\Omega, \mathbb{R})$, such that $P(Z_{1,1}, \ldots, Z_{d_2,s}) = \sum_{k=1}^{N(P)} \phi_k \times \prod_{i=1}^{d_2}\prod_{j=1}^{s} Z_{i,j}^{I(i,j,k)}$. Recall, by Definition 4.5, that $\deg(\mathscr{F}) = \max_k \sum_{i=1}^{d_2}\sum_{j=1}^{s}(1+|\alpha_{i,j}|)I(i,j,k)$.

Now, according to Proposition 4.2, there exists a positive constant $C_{\deg(\mathscr{F}),H}$ such that

$$
\begin{aligned}
&\|\mathscr{F}(u_\theta,\cdot)^2\|_{\infty,\bar\Omega} \\
&\quad\leqslant \left[\sum_{k=1}^{N(P)} \|\phi_k\|_{\infty,\bar\Omega}\prod_{i=1}^{d_2}\prod_{j=1}^{s}\|\partial^{\alpha_{i,j}}u_\theta\|_{\infty,\bar\Omega}^{I(i,j,k)}\right]^2 \\
&\quad\leqslant N^2(P)\big[\max_{1\leqslant k\leqslant N(P)}\|\phi_k\|_{\infty,\bar\Omega}\big]^2 C^2_{\deg(\mathscr{F}),H}(D+1)^{2H\deg(\mathscr{F})}(1+\|\theta\|_2)^{2H\deg(\mathscr{F})}.
\end{aligned}
$$

Thus, for any $\theta \in B(n^{\alpha_1}, \|\cdot\|_2)$, $\|\mathscr{F}(u_\theta,\cdot)^2\|_{\infty,\bar\Omega} \leqslant C_2 n^{\beta_2}$, where

$$
C_2 = 2^{2H\deg(\mathscr{F})}N^2(P)\big[\max_{1\leqslant k\leqslant N(P)}\|\phi_k\|_{\infty,\bar\Omega}\big]^2 C^2_{\deg(\mathscr{F}),H}(D+1)^{2H\deg(\mathscr{F})},
$$

and for any $\beta_2 \geqslant 2H\deg(\mathscr{F})\alpha_1$.

Next, observe that, any u and v, $||u|^2-|v|^2| = |(u+v)(u-v)| \leqslant |u+v||u-v|$. Therefore,

$$\begin{aligned}|\mathscr{F}(u_\theta,\mathbf{x})^2 - \mathscr{F}(u_{\theta'},\mathbf{x})^2| &\leqslant \big(|\mathscr{F}(u_\theta,\mathbf{x})| + |\mathscr{F}(u_{\theta'},\mathbf{x})|\big)|\mathscr{F}(u_\theta,\mathbf{x}) - \mathscr{F}(u_{\theta'},\mathbf{x})| \\ &\leqslant 2C_2^{1/2} n^{H\deg(\mathscr{F})\alpha_1}|\mathscr{F}(u_\theta,\mathbf{x}) - \mathscr{F}(u_{\theta'},\mathbf{x})|.\end{aligned}$$

Using inequality (19) (remark that the product $\prod_{i=1}^{d_2}\prod_{j=1}^{s} Z_{i,j}^{I(i,j,k)}$ has less than $\deg(\mathscr{F})$ terms different from 1), it is easy to see that

$$\begin{aligned}|\mathscr{F}(u_\theta,\mathbf{x}) - \mathscr{F}(u_{\theta'},\mathbf{x})| &\leqslant N(P)\big[\max_{1\leqslant k\leqslant N(P)}\|\phi_k\|_{\infty,\bar{\Omega}}\big]\deg(\mathscr{F})\|u_\theta - u_{\theta'}\|_{C^{\deg(\mathscr{F})}(\Omega)} \\ &\quad\times \max_{1\leqslant k\leqslant N(P)}\prod_{i,j}\max(\|u_\theta\|_{C^{|\alpha_{i,j}|}(\Omega)}, \|u_{\theta'}\|_{C^{|\alpha_{i,j}|}(\Omega)})^{I(i,j,k)}.\end{aligned}$$

From Proposition 4.2, we deduce that

$$\begin{aligned}&\max_{1\leqslant k\leqslant N(P)}\prod_{i,j}\max(\|u_\theta\|_{C^{|\alpha_{i,j}|}(\Omega)}, \|u_{\theta'}\|_{C^{|\alpha_{i,j}|}(\Omega)})^{I(i,j,k)} \\ &\quad\leqslant C_{\deg(\mathscr{F}),H}(D+1)^{H\deg(\mathscr{F})}(1+\max(\|\theta\|_2,\|\theta'\|_2))^{H\deg(\mathscr{F})}.\end{aligned}$$

Combining the last two inequalities with Proposition F.1 gives that

$$\begin{aligned}&|\mathscr{F}(u_\theta,\mathbf{x}) - \mathscr{F}(u_{\theta'},\mathbf{x})| \\ &\quad\leqslant N(P)\big[\max_{1\leqslant k\leqslant N(P)}\|\phi_k\|_{\infty,\bar{\Omega}}\big]\deg(\mathscr{F})\tilde{C}_{\deg(\mathscr{F}),H}(1+d_1M(\Omega))\|\theta-\theta'\|_2 \\ &\qquad\times C_{\deg(\mathscr{F}),H}(D+1)^{H(1+(1+H)\deg(\mathscr{F}))}(1+\max(\|\theta\|_2,\|\theta'\|_2))^{H(1+(1+H)\deg(\mathscr{F}))}.\end{aligned}$$

Hence, for all $\theta,\theta' \in B(n^{\alpha_1},\|\cdot\|_2)$, $|\mathscr{F}(u_\theta,\mathbf{x})^2 - \mathscr{F}(u_{\theta'},\mathbf{x})^2| \leqslant C_1 n^{\beta_1}\|\theta-\theta'\|_2$, where

$$\begin{aligned}C_1 &= 2C_2^{1/2}N(P)\big[\max_{1\leqslant k\leqslant N(P)}\|\phi_k\|_{\infty,\bar{\Omega}}\big]\deg(\mathscr{F})\tilde{C}_{\deg(\mathscr{F}),H}(1+d_1M(\Omega)) \\ &\quad\times C_{\deg(\mathscr{F}),H}(D+1)^{H(1+(1+H)\deg(\mathscr{F}))}2^{H(1+(1+H)\deg(\mathscr{F}))}\end{aligned}$$

and $\beta_1 = H(1+(2+H)\deg(\mathscr{F}))\alpha_1$.

Recall that for inequality (29), β_2 must satisfy $\alpha_1+\beta_1<\beta_2<1/2$. This is true for $\beta_2 = [2+H(1+(2+H)\deg(\mathscr{F}))]\alpha_1$ and $\alpha_1 < [2+H(1+(2+H)\deg(\mathscr{F}))]^{-1}/2$. □

F.4. Proof of Theorem 4.6. Let $u_0 = 0 \in \mathrm{NN}_H(D)$ be the neural network with parameter $\theta = (0,\dots,0)$. Obviously, $R^{(\mathrm{ridge})}_{n,n_e,n_r}(u_0) = R_{n,n_e,n_r}(u_0)$. Also,

$$R_{n,n_e,n_r}(u_0) \leqslant \frac{\lambda_d}{n}\sum_{i=1}^{n}\|Y_i\|_2^2 + \lambda_e\|h\|_\infty + \frac{1}{n_r}\sum_{k=1}^{M}\sum_{\ell=1}^{n_r}\|\mathscr{F}_k(0,\mathbf{X}_\ell^{(r)})\|_2^2.$$

Since each $\mathscr{F}_k$ is a polynomial operator (see Definition 4.4), it takes the form

$$\mathscr{F}_k(u,\mathbf{x}) = \sum_{\ell=1}^{N(P_k)}\phi_{\ell,k}\prod_{i=1}^{d_2}\prod_{j=1}^{s_k}(\partial^{\alpha_{i,j,k}}u_i(\mathbf{x}))^{I_k(i,j,\ell)}.$$

Therefore,

$$\begin{aligned}R_{n,n_e,n_r}(u_0) &\leqslant \frac{\lambda_d}{n}\sum_{i=1}^{n}\|Y_i\|_2^2 + \lambda_e\|h\|_\infty + \sum_{k=1}^{M}\sum_{\ell=1}^{N(P_k)}\|\phi_{\ell,k}\|_{\infty,\bar{\Omega}} \\ &:= I,\end{aligned} \tag{31}$$

where I does not depend on $\lambda_{(\mathrm{ridge})}$, n_e, and n_r.

Let $(\hat{\theta}^{(\mathrm{ridge})}(p,n_e,n_r,D))_{p\in\mathbb{N}}$ be any minimizing sequence of the empirical risk of the ridge PINN, i.e., $\lim_{p\to\infty} R^{(\mathrm{ridge})}_{n,n_e,n_r}(u_{\hat{\theta}^{(\mathrm{ridge})}(p,n_e,n_r,D)}) = \inf_{\theta\in\Theta_{H,D}} R^{(\mathrm{ridge})}_{n,n_e,n_r}(u_\theta)$. In the rest of the proof, we let $n_{r,e} = \min(n_r,n_e)$. We will make use of the following three sets: $\mathcal{E}_1(n_{r,e}) = \{\theta\in\Theta_{H,D},\ \|\theta\|_2 \geq n_{r,e}^{\kappa}\}$, $\mathcal{E}_2(n_{r,e}) = \{\theta\in\Theta_{H,D},\ n_{r,e}^{\kappa/4} \leq \|\theta\|_2 \leq n_{r,e}^{\kappa}\}$, and $\mathcal{E}_3(n_{r,e}) = \{\theta\in\Theta_{H,D},\ \|\theta\|_2 \leq n_{r,e}^{\kappa/4}\}$. Clearly, $\Theta_{H,D} = \mathcal{E}_1\cup\mathcal{E}_2\cup\mathcal{E}_3$. The proof relies on the argument that almost surely, given any n_r and n_e, for all p large enough, $\hat{\theta}^{(\mathrm{ridge})}(p,n_e,n_r,D)\in\mathcal{E}_2\cup\mathcal{E}_3$. Moreover, on $\mathcal{E}_2\cup\mathcal{E}_3$, the empirical risk function $R^{(\mathrm{ridge})}_{n,n_e,n_r}$ is close to the theoretical risk $\mathscr{R}_n$, when $n_{r,e}$ is large enough. For clarity, the proof is divided into four steps.

Step 1. We start by observing that, for any $\theta\in\mathcal{E}_1(n_{r,e})$, $R^{(\mathrm{ridge})}_{n,n_e,n_r}(\theta) \geqslant \lambda_{(\mathrm{ridge})}\|\theta\|_2^2 \geqslant n_{r,e}^{\kappa}$. Therefore, according to (31), once $n_{r,e}\geq (I+1)^{1/\kappa}$,

$$\inf_{\theta\in\mathcal{E}_3(n_{r,e})} R^{(\mathrm{ridge})}_{n,n_e,n_r}(u_\theta) + 1 \leqslant R^{(\mathrm{ridge})}_{n,n_e,n_r}(u_0) + 1 \leqslant \inf_{\theta\in\mathcal{E}_1(n_{r,e})} R^{(\mathrm{ridge})}_{n,n_e,n_r}(u_\theta).$$

This shows that, for all $n_{r,e}$ large enough and for all p large enough, $\hat{\theta}^{(\mathrm{ridge})}(p,n_e,n_r,D)\notin\mathcal{E}_1(n_{r,e})$.

Step 2. Applying Proposition F.3 and Proposition F.4 with $\alpha_1=\kappa$ and $\beta_2 = (2+H(1+(2+H)\max_k \deg(\mathscr{F}_k)))\alpha_1$, and then Theorem F.2, we know that, almost surely, there exists $N\in\mathbb{N}^\star$ such that, for all $n_{r,e}\geqslant N$,

$$\sup_{\theta\in\mathcal{E}_2(n_{r,e})\cup\mathcal{E}_3(n_{r,e})} \left|\frac{1}{n_e}\sum_{j=1}^{n_e}\|u_\theta(\mathbf{X}_j^{(e)}) - h(\mathbf{X}_j^{(e)})\|_2^2 - \mathbb{E}\|u_\theta(\mathbf{X}^{(e)}) - h(\mathbf{X}^{(e)})\|_2^2\right|$$

$$\leqslant \log^2(n_{r,e}) n_{r,e}^{\beta_2-1/2} \tag{32}$$

and, for each $1\leqslant k\leqslant M$,

$$\sup_{\theta\in\mathcal{E}_2(n_{r,e})\cup\mathcal{E}_3(n_{r,e})} \left|\frac{1}{n_r}\sum_{\ell=1}^{n_r}\mathscr{F}_k(u_\theta,\mathbf{X}_\ell^{(r)})^2 - \frac{1}{|\Omega|}\int_\Omega \mathscr{F}_k(u_\theta,\mathbf{x})^2 d\mathbf{x}\right| \leqslant \log^2(n_{r,e}) n_{r,e}^{\beta_2-1/2}. \tag{33}$$

Thus, almost surely, for all $n_{r,e}$ large enough and for all $\theta\in\mathcal{E}_2(n_{r,e})$,

$$R^{(\mathrm{ridge})}_{n,n_e,n_r}(u_\theta) \geqslant \mathscr{R}_n(u_\theta) + \lambda_{(\mathrm{ridge})}\|\theta\|_2^2 - (M+1)\log^2(n_{r,e}) n_{r,e}^{\beta_2-1/2}.$$

But, for all $\theta\in\mathcal{E}_2(n_{r,e})$, $\lambda_{(\mathrm{ridge})}\|\theta\|_2^2 \geqslant n_{e,r}^{-\kappa/2}$. Upon noting that $-\kappa/2 > \beta_2 - 1/2$, we conclude that, almost surely, for all $n_{r,e}$ large enough and for all $\theta\in\mathcal{E}_2(n_{r,e})$, $R^{(\mathrm{ridge})}_{n,n_e,n_r}(u_\theta) \geqslant \mathscr{R}_n(u_\theta)$.

Step 3. Clearly, for all $\theta\in\mathcal{E}_3(n_{r,e})$, $\lambda_{(\mathrm{ridge})}\|\theta\|_2^2 \leqslant n_{e,r}^{-\kappa/2}$. Using inequalities (32) and (33), we deduce that, almost surely, for all $n_{r,e}$ large enough and for all $\theta\in\mathcal{E}_3(n_{r,e})$, $|R^{(\mathrm{ridge})}_{n,n_e,n_r}(u_\theta) - \mathscr{R}_n(u_\theta)| \leqslant (M+2)\log^2(n_{r,e}) n_{r,e}^{-\kappa/2}$.

Step 4. Fix $\varepsilon>0$. Let $(\theta_p)_{p\in\mathbb{N}}$ be any minimizing sequence of the theoretical risk function $\mathscr{R}_n$, that is, $\lim_{p\to\infty}\mathscr{R}_n(u_{\theta_p}) = \inf_{\theta\in\Theta_{H,D}}\mathscr{R}_n(u_\theta)$. Thus, by definition, there exists some $P_\varepsilon\in\mathbb{N}$ such that $|\mathscr{R}_n(u_{\theta_{P_\varepsilon}}) - \inf_{\theta\in\Theta_{H,D}}\mathscr{R}_n(u_\theta)| \leqslant \varepsilon$.

For fixed $n_{r,e}$, according to Step 1, we have, for all p large enough, $\hat{\theta}^{(\mathrm{ridge})}(p,n_e,n_r,D)\in\mathcal{E}_2(n_{r,e})\cup\mathcal{E}_3(n_{r,e})$. So, according to Step 2 and Step 3,

$$\mathscr{R}_n(u_{\hat{\theta}^{(\mathrm{ridge})}(p,n_e,n_r,D)}) \leqslant R^{(\mathrm{ridge})}_{n,n_e,n_r}(u_{\hat{\theta}^{(\mathrm{ridge})}(p,n_e,n_r,D)}) + (M+2)\log^2(n_{r,e}) n_{r,e}^{-\kappa/2}.$$

Now, by definition of the minimizing sequence $(\hat{\theta}^{(\mathrm{ridge})}(p,n_e,n_r,D))_{p\in\mathbb{N}}$, for all p large enough, $R^{(\mathrm{ridge})}_{n,n_e,n_r}(u_{\hat{\theta}^{(\mathrm{ridge})}(p,n_e,n_r,D)}) \leqslant \inf_{\theta\in\Theta_{H,D}} R^{(\mathrm{ridge})}_{n,n_e,n_r}(u_\theta)+\varepsilon$. Also, according to Step 3,

$$\begin{aligned}\inf_{\theta\in\mathcal{E}_2(n_{r,e})\cup\mathcal{E}_3(n_{r,e})} R^{(\mathrm{ridge})}_{n,n_e,n_r}(u_\theta) &\leqslant \inf_{\theta\in\mathcal{E}_3(n_{r,e})} R^{(\mathrm{ridge})}_{n,n_e,n_r}(u_\theta)\\ &\leqslant \inf_{\theta\in\mathcal{E}_3(n_{r,e})} \mathscr{R}_n(u_\theta)+(M+2)\log^2(n_{r,e})n_{r,e}^{-\kappa/2}.\end{aligned}$$

Observe that, for all $n_{r,e}$ large enough, $\theta_{P_\varepsilon}\in\mathcal{E}_3(n_{r,e})$. Therefore, $\inf_{\theta\in\mathcal{E}_3(n_{r,e})}\mathscr{R}_n(u_\theta)\leqslant \mathscr{R}_n(u_{\theta_{P_\varepsilon}})$. Combining the previous inequalities, we conclude that, almost surely, for all $n_{r,e}$ large enough and for all p large enough,

$$\mathscr{R}_n(u_{\hat{\theta}^{(\mathrm{ridge})}(p,n_e,n_r,D)}) \leqslant \inf_{\theta\in\Theta_{H,D}}\mathscr{R}_n(u_\theta)+3\varepsilon.$$

Since ε is arbitrary, then, almost surely, $\lim_{n_e,n_r\to\infty}\lim_{p\to\infty}\mathscr{R}_n(u_{\hat{\theta}^{(\mathrm{ridge})}(p,n_e,n_r,D)}) = \inf_{\theta\in\Theta_{H,D}}\mathscr{R}_n(u_\theta)$.

F.5. Proof of Theorem 4.7. The result is a direct consequence of Theorem 4.6, Proposition 2.3 and of the continuity of $\mathscr{R}_n$ with respect to the $C^K(\Omega)$ norm.

APPENDIX G: PROOFS OF SECTION 5

G.1. Proof of Proposition 5.5. Since the functions in $H^{m+1}(\Omega,\mathbb{R}^{d_2})$ are only defined almost everywhere, we first have to give a meaning to the pointwise evaluations $u(\mathbf{X}_i)$ when $u\in H^{m+1}(\Omega,\mathbb{R}^{d_2})$. Since Ω is a bounded Lipschitz domain and $(m+1)>d_1/2$, we can use the Sobolev embedding of Theorem B.1. Clearly, $\tilde{\Pi}$ is linear and $\|\tilde{\Pi}(u)\|_\infty\leqslant C_\Omega\|u\|_{H^{m+1}(\Omega)}$. The natural choice to evaluate $u\in H^{m+1}(\Omega,\mathbb{R}^{d_2})$ at the point $\mathbf{X}_i$ is therefore to evaluate its unique continuous modification $\tilde{\Pi}(u)$ at $\mathbf{X}_i$.

By assumption, $\mathscr{F}_k(u,\cdot)=\mathscr{F}_k^{(\mathrm{lin})}(u,\cdot)+B_k$, where $\mathscr{F}_k^{(\mathrm{lin})}(u,\cdot)=\sum_{|\alpha|\leqslant K}\langle A_{k,\alpha},\partial^\alpha u\rangle$ and $A_{k,\alpha}\in C^\infty(\bar{\Omega},\mathbb{R}^{d_1})$. Next, consider the symmetric bilinear form, defined for all $u,v\in H^{m+1}(\Omega,\mathbb{R}^{d_2})$ by

$$\begin{aligned}\mathcal{A}_n(u,v) &= \frac{\lambda_d}{n}\sum_{i=1}^n\langle\tilde{\Pi}(u)(\mathbf{X}_i),\tilde{\Pi}(v)(\mathbf{X}_i)\rangle+\lambda_e\mathbb{E}\langle\tilde{\Pi}(u)(\mathbf{X}^{(e)}),\tilde{\Pi}(v)(\mathbf{X}^{(e)})\rangle\\ &\quad+\frac{1}{|\Omega|}\sum_{k=1}^M\int_\Omega\mathscr{F}_k^{(\mathrm{lin})}(u,\mathbf{x})\mathscr{F}_k^{(\mathrm{lin})}(v,\mathbf{x})d\mathbf{x}+\frac{\lambda_t}{|\Omega|}\sum_{|\alpha|\leqslant m+1}\int_\Omega\langle\partial^\alpha u(\mathbf{x}),\partial^\alpha v(\mathbf{x})\rangle d\mathbf{x},\end{aligned}$$

along with the linear form defined for all $u\in H^{m+1}(\Omega,\mathbb{R}^{d_2})$ by

$$\begin{aligned}\mathcal{B}_n(u) &= \frac{\lambda_d}{n}\sum_{i=1}^n\langle Y_i,\tilde{\Pi}(u)(\mathbf{X}_i)\rangle+\lambda_e\mathbb{E}\langle\tilde{\Pi}(u)(\mathbf{X}^{(e)}),h(\mathbf{X}^{(e)})\rangle\\ &\quad-\frac{1}{|\Omega|}\sum_{k=1}^M\int_\Omega B_k(\mathbf{x})\mathscr{F}_k^{(\mathrm{lin})}(v,\mathbf{x})d\mathbf{x}.\end{aligned}$$

Observe that

$$\mathcal{A}_n(u,u)-2\mathcal{B}_n(u)=\mathscr{R}_n^{(\mathrm{reg})}(u)-\frac{\lambda_d}{n}\sum_{i=1}^n\|Y_i\|_2^2-\lambda_e\mathbb{E}\|h(\mathbf{X}^{(e)})\|_2^2-\frac{1}{|\Omega|}\sum_{k=1}^M\int_\Omega B_k(\mathbf{x})^2d\mathbf{x}.$$

In addition, $\mathcal{A}_n(u,u) \geqslant \lambda_t \|u\|^2_{H^{m+1}(\Omega)}$, where $\lambda_t > 0$, so that $\mathcal{A}_n$ is coercive on the normed space $(H^{m+1}(\Omega), \|\cdot\|_{H^{m+1}(\Omega)})$. Since $(m+1) > \max(d_1/2, K)$, one has that

$$|\mathcal{A}_n(u,v)| \leqslant ((\lambda_d + \lambda_e)C_\Omega^2 + \sum_{1\leqslant k\leqslant M} (\sum_{|\alpha|\leqslant K} \|A_{k,\alpha}\|_{\infty,\Omega})^2 + \lambda_t)\|u\|_{H^{m+1}(\Omega)}\|v\|_{H^{m+1}(\Omega)},$$

and

$$|\mathcal{B}_n(u)| \leqslant C_\Omega\Big(\frac{\lambda_d}{n}\sum_{i=1}^n \|Y_i\|_2 + \lambda_e\|h\|_\infty + \sum_{k=1}^M (\|B_k\|_{\infty,\Omega} \sum_{|\alpha|\leqslant K} \|A_{k,\alpha}\|_{\infty,\Omega})\Big)\|u\|_{H^{m+1}(\Omega)}.$$

This shows that the operators $\mathcal{A}_n$ and $\mathcal{B}_n$ are continuous. Therefore, by the Lax-Milgram theorem [e.g., Brezis, 2010, Corollary 5.8], there exists a unique $\hat{u} \in H^{m+1}(\Omega, \mathbb{R}^{d_2})$ such that $\mathcal{A}_n(\hat{u},\hat{u}) - 2\mathcal{B}_n(\hat{u}) = \min_{u\in H^{m+1}(\Omega,\mathbb{R}^{d_2})} \mathcal{A}_n(u,u) - 2\mathcal{B}_n(u)$. This directly implies that $\hat{u}$ is the unique minimizer of $\mathscr{R}_n^{(\mathrm{reg})}$ over $H^{m+1}(\Omega, \mathbb{R}^{d_2})$. Furthermore, the Lax-Milgram theorem also states that $\hat{u}$ is the unique element of $H^{m+1}(\Omega, \mathbb{R}^{d_2})$ such that, for all $v \in H^{m+1}(\Omega, \mathbb{R}^{d_2})$, $\mathcal{A}_n(\hat{u}, v) = \mathcal{B}_n(v)$. This concludes the proof of the proposition.

G.2. Proof of Proposition 5.6. Let $\hat{u}_n$ be the unique minimizer of the regularized theoretical risk $\mathscr{R}_n^{(\mathrm{reg})}$ over $H^{m+1}(\Omega, \mathbb{R}^{d_2})$ given by Proposition 5.5. Notice that

$$\inf_{u\in C^\infty(\bar{\Omega},\mathbb{R}^{d_2})} \mathscr{R}_n^{(\mathrm{reg})}(u) = \inf_{u\in H^{m+1}(\Omega,\mathbb{R}^{d_2})} \mathscr{R}_n^{(\mathrm{reg})}(u) = \mathscr{R}_n(\hat{u}_n).$$

The first equality is a consequence of the density of $C^\infty(\bar{\Omega}, \mathbb{R}^{d_2})$ in $H^{m+1}(\Omega, \mathbb{R}^{d_2})$, together with the continuity of the function $\mathscr{R}_n^{(\mathrm{reg})} : H^{m+1}(\Omega, \mathbb{R}^{d_2}) \to \mathbb{R}$ with respect to the $H^{m+1}(\Omega)$ norm (see the proof of Proposition 5.5). The density argument follows from the extension theorem of Stein [1970, Chapter VI.3.3, Theorem 5] and from Evans [2010, Chapter 5.3, Theorem 3].

Our goal is to show that the regularized theoretical risk satisfies some form of continuity, so that we can connect $\mathscr{R}^{(\mathrm{reg})}(u_p)$ and $\mathscr{R}^{(\mathrm{reg})}(\hat{u}_n)$. Recall that, by assumption, $\mathscr{F}_k(u,\cdot) = \mathscr{F}_k^{(\mathrm{lin})}(u,\cdot) + B_k$, where $\mathscr{F}_k^{(\mathrm{lin})}(u,\cdot) = \sum_{|\alpha|\leqslant K}\langle A_{k,\alpha}(\cdot), \partial^\alpha u(\cdot)\rangle$ and $A_{k,\alpha} \in C^\infty(\bar{\Omega}, \mathbb{R}^{d_1})$. Observe that

$$\mathscr{R}_n^{(\mathrm{reg})}(u) = F(u) + \frac{1}{|\Omega|} I(u), \tag{34}$$

where

$$F(u) = \frac{\lambda_d}{n}\sum_{i=1}^n \|\tilde{\Pi}(u)(\mathbf{X}_i) - Y_i\|_2^2 + \lambda_e \mathbb{E}\|\tilde{\Pi}(u)(\mathbf{X}^{(e)}) - h(\mathbf{X}^{(e)})\|_2^2,$$

$$I(u) = \int_\Omega L((\partial^{m+1}_{i_1,\ldots,i_{m+1}} u(\mathbf{x}))_{1\leqslant i_1,\ldots,i_{m+1}\leqslant d_1}, \ldots, u(\mathbf{x}), \mathbf{x})d\mathbf{x},$$

and where the function L satisfies

$$L(x^{(m+1)}, \ldots, x^{(0)}, z) = \sum_{k=1}^M \Big(B_k(z) + \sum_{|\alpha|\leqslant K} \langle A_{k,\alpha}(z), x_\alpha^{(|\alpha|)}\rangle\Big)^2 + \lambda_t \sum_{j=0}^{m+1} \|x^{(j)}\|_2^2.$$

(The term $x^{(j)} \in \mathbb{R}^{\binom{d_1+j-1}{j-1}d_2}$ corresponds to the to the concatenation of all the partial derivatives of order j, i.e., to the term $(\partial^j_{i_1,\ldots,i_j} u(\mathbf{x}))_{1\leqslant i_1,\ldots,i_j\leqslant d_1}$.) Clearly, $L \geqslant 0$ and, since

$(m+1) > K$, the Lagrangian L is convex in $x^{(m+1)}$. Therefore, according to Lemma C.11, the function I is weakly lower-semi continuous on $H^{m+1}(\Omega, \mathbb{R}^{d_2})$.

Now, let us proceed by contradiction and assume that there is a sequence $(u_p)_{p\in\mathbb{N}}$ of functions such that (i) $u_p \in C^\infty(\bar{\Omega}, \mathbb{R}^{d_2})$, (ii) $\lim_{p\to\infty} \mathscr{R}_n^{(\mathrm{reg})}(u_p) = \mathscr{R}_n^{(\mathrm{reg})}(\hat{u}_n)$, and (iii) $(u_p)_{p\in\mathbb{N}}$ does not converge to $\hat{u}_n$ with respect to the $H^m(\Omega)$ norm. Therefore, upon passing to a subsequence, there exists $\varepsilon > 0$ such that, for all $p \geqslant 0$, $\|u_p - \hat{u}_n\|_{H^m(\Omega)} \geqslant \varepsilon$.

Since $\mathscr{R}_n^{(\mathrm{reg})}(u_p) \geqslant \lambda_t \|u_p\|_{H^{m+1}(\Omega)}$, $\lambda_t > 0$, and $(u_p)_{p\in\mathbb{N}}$ is a minimizing sequence, $(u_p)_{p\in\mathbb{N}}$ is bounded in $H^{m+1}(\Omega, \mathbb{R}^{d_2})$. Therefore, Theorem B.4 states that passing to a subsequence, $(u_p)_{p\in\mathbb{N}}$ converges to a limit, say u_∞, both weakly in $H^{m+1}(\Omega, \mathbb{R}^{d_2})$ and with respect to the $H^m(\Omega)$ norm. Then, since I is weakly lower-semi continuous on $H^{m+1}(\Omega, \mathbb{R}^{d_2})$, we deduce that

$$\lim_{p\to\infty} I(u_p) \geqslant I(u_\infty). \tag{35}$$

Recalling the definition of $\tilde{\Pi}$ in Theorem B.1, we know that there exists a constant $C_\Omega > 0$ such that $\|u_p - \tilde{\Pi}(u_\infty)\|_{\infty,\Omega} = \|\tilde{\Pi}(u_p - u_\infty)\|_{\infty,\Omega} \leqslant C_\Omega \|u_p - u_\infty\|_{H^m(\Omega)}$. We deduce that $\lim_{p\to\infty} F(u_p) = F(u_\infty)$. Therefore, combining this result with (34) and (35), we deduce that $\lim_{p\to\infty} \mathscr{R}_n^{(\mathrm{reg})}(u_p) \geqslant \mathscr{R}_n^{(\mathrm{reg})}(u_\infty)$. However, recalling that $\lim_{p\to\infty} \mathscr{R}_n^{(\mathrm{reg})}(u_p) = \mathscr{R}_n^{(\mathrm{reg})}(\hat{u}_n)$ and that $\hat{u}_n$ is the unique minimizer of $\mathscr{R}_n^{(\mathrm{reg})}$ over $H^{m+1}(\Omega, \mathbb{R}^{d_2})$, we conclude that $u_\infty = \hat{u}_n$.

We just proved that there exists a subsequence of $(u_p)_{p\in\mathbb{N}}$ which converges to $\hat{u}_n$ with respect to the $H^m(\Omega)$ norm. This contradicts the assumption $\|u_p - \hat{u}_n\|_{H^m(\Omega)} \geqslant \varepsilon$ for all $p \geqslant 0$.

G.3. Proof of Theorem 5.7. The result is an immediate consequence of Theorem 4.7, Propositions 5.5, and Proposition 5.6.

G.4. Proof of Theorem 5.8. Throughout the proof, since no data are involved, we denote the regularized theoretical risk by $\mathscr{R}^{(\mathrm{reg})}$ instead of $\mathscr{R}_n^{(\mathrm{reg})}$. Also, to make the dependence in the hyperparameter λ_t transparent, we denote by $u(\lambda_t)$ the unique minimizer of $\mathscr{R}^{(\mathrm{reg})}$ instead of $\hat{u}_n$.

We proceed by contradiction and assume that $\lim_{\lambda_t\to 0} \|u(\lambda_t) - u^\star\|_{H^m(\Omega)} \neq 0$. If this is true, then, upon passing to a subsequence $(\lambda_{t,p})_{p\in\mathbb{N}}$ such that $\lim_{p\to\infty} \lambda_{t,p} = 0$, there exists $\varepsilon > 0$ such that, for all $p \geqslant 0$, $\|u(\lambda_{t,p}) - u^\star\|_{H^m(\Omega)} \geqslant \varepsilon$.

Notice that $\|u(\lambda_{t,p})\|_{H^{m+1}(\Omega)} \leqslant \mathscr{R}^{(\mathrm{reg})}(u^\star)/\lambda_{t,p} = \|u^\star\|_{H^{m+1}(\Omega)}$. Theorem B.4 proves that upon passing to a subsequence, $(u(\lambda_{t,p}))_{p\in\mathbb{N}}$ converges with respect to the $H^m(\Omega)$ norm to a function $u_\infty \in H^{m+1}(\Omega, \mathbb{R}^{d_2})$. Since $m \geqslant K$, the theoretical risk $\mathscr{R}$ is continuous with respect to the $H^m(\Omega)$ norm and we have that $\mathscr{R}(u_\infty) = \lim_{p\to\infty} \mathscr{R}(u(\lambda_{t,p}))$. Moreover, by definition of $u(\lambda_{t,p})$ and since $\mathscr{R}(u^\star) = 0$, we have that $\mathscr{R}(u(\lambda_{t,p})) + \lambda_{t,p}\|u(\lambda_{t,p})\|_{H^{m+1}(\Omega)} \leqslant \lambda_{t,p}\|u^\star\|_{H^{m+1}(\Omega)}$. Therefore, $\mathscr{R}(u_\infty) = 0$ and $u_\infty = u^\star$. This contradicts the assumption that for all $p \geqslant 0$, $\|u(\lambda_{t,p}) - u^\star\|_{H^m(\Omega)} \geqslant \varepsilon$.

G.5. Proof of Proposition 5.11. We prove the proposition in several steps. In the sequel, given a measure μ on Ω and a function $u \in H^{m+1}(\Omega, \mathbb{R}^{d_2})$, we let $\|u\|^2_{L^2(\mu)} = \int_\Omega \|\tilde{\Pi}(u)(\mathbf{x})\|_2^2 d\mu(\mathbf{x})$, where, as usual, $\tilde{\Pi}(u)$ is the unique continuous function such that $\tilde{\Pi}(u) = u$ almost everywhere.

Step 1: Decomposing the problem into two simpler ones. Following the framework of Arnone et al. [2022], the core idea is to decompose the problem into two simpler ones thanks to the linearity in $\hat{u}_n$ and in Y_i of the identity

$$\forall v \in H^{m+1}(\Omega, \mathbb{R}^{d_2}), \quad \mathcal{A}_n(\hat{u}_n, v) = \mathcal{B}_n(v)$$

of Proposition 5.5. Thus, recalling that $Y_i = u^\star(\mathbf{X}_i) + \varepsilon_i$, we let

$$\mathcal{B}_n^\star(v) = \frac{\lambda_d}{n}\sum_{i=1}^n \langle u^\star(\mathbf{X}_i), \tilde{\Pi}(v)(\mathbf{X}_i)\rangle + \lambda_e \mathbb{E}\langle \tilde{\Pi}(v)(\mathbf{X}^{(e)}), h(\mathbf{X}^{(e)})\rangle$$
$$- \frac{1}{|\Omega|}\sum_{k=1}^M \int_\Omega B_k(\mathbf{x}) \mathscr{F}_k^{(\mathrm{lin})}(v, \mathbf{x}) d\mathbf{x}$$

and

$$\mathcal{B}_n^{(\mathrm{noise})}(v) = \frac{\lambda_d}{n}\sum_{i=1}^n \langle \varepsilon_i, \tilde{\Pi}(v)(\mathbf{X}_i)\rangle.$$

Clearly, $\mathcal{B}_n = \mathcal{B}_n^\star + \mathcal{B}_n^{(\mathrm{noise})}$. Using Proposition 5.5 with Y_i instead of ε_i, and setting $\lambda_e = 0$, we see that there exists a unique $\hat{u}_n^{(\mathrm{noise})} \in H^{m+1}(\Omega, \mathbb{R}^{d_2})$ such that, for all $v \in H^{m+1}(\Omega, \mathbb{R}^{d_2})$, $\mathcal{A}_n(\hat{u}_n^{(\mathrm{noise})}, v) = \mathcal{B}_n^{(\mathrm{noise})}(v)$. Furthermore, $\hat{u}_n^{(\mathrm{noise})}$ is the unique minimizer over $H^{m+1}(\Omega, \mathbb{R}^{d_2})$ of

$$\mathscr{R}_n^{(\mathrm{noise})}(u) = \frac{\lambda_d}{n}\sum_{i=1}^n \|\tilde{\Pi}(u)(\mathbf{X}_i) - \varepsilon_i\|_2^2 + \lambda_e \mathbb{E}\|u(\mathbf{X}^{(e)})\|_2^2 + \frac{1}{|\Omega|}\sum_{k=1}^M \int_\Omega \mathscr{F}_k^{(\mathrm{lin})}(u, \mathbf{x})^2 d\mathbf{x}$$
$$+ \lambda_t \|u\|_{H^{m+1}(\Omega)}^2.$$

Similarly, Proposition 5.5 shows that there exists a unique $\hat{u}_n^\star \in H^{m+1}(\Omega, \mathbb{R}^{d_2})$ such that, for all $v \in H^{m+1}(\Omega, \mathbb{R}^{d_2})$, $\mathcal{A}_n(\hat{u}_n^\star, v) = \mathcal{B}_n^\star(v)$, and $\hat{u}_n^\star$ is the unique minimizer over $H^{m+1}(\Omega, \mathbb{R}^{d_2})$ of

$$\mathscr{R}_n^\star(u) = \frac{\lambda_d}{n}\sum_{i=1}^n \|\tilde{\Pi}(u - u^\star)(\mathbf{X}_i)\|_2^2 + \lambda_e \mathbb{E}\|\tilde{\Pi}(u)(\mathbf{X}^{(e)}) - h(\mathbf{X}^{(e)})\|_2^2$$
$$+ \frac{1}{|\Omega|}\sum_{k=1}^M \int_\Omega \mathscr{F}_k(u, \mathbf{x})^2 d\mathbf{x} + \lambda_t \|u\|_{H^{m+1}(\Omega)}^2.$$

By the bilinearity of $\mathcal{A}_n$, one has, for all $v \in H^{m+1}(\Omega, \mathbb{R}^{d_2})$, $\mathcal{A}_n(\hat{u}_n^\star + \hat{u}_n^{(\mathrm{noise})}, v) = \mathcal{B}_n(v)$. However, according to Proposition 5.5, $\hat{u}_n$ is the unique element of $H^{m+1}(\Omega, \mathbb{R}^{d_2})$ satisfying this property. Therefore, $\hat{u}_n = \hat{u}_n^\star + \hat{u}_n^{(\mathrm{noise})}$.

Step 2: Some properties of the minimizers. According to Lemma C.12, $\hat{u}_n$, $\hat{u}_n^\star$, and $\hat{u}_n^{(\mathrm{noise})}$ are random variables. Our goal in this paragraph is to prove that $\mathbb{E}\|\hat{u}_n\|_{H^{m+1}(\Omega)}^2$, $\mathbb{E}\|\hat{u}_n^\star\|_{H^{m+1}(\Omega)}^2$, and $\mathbb{E}\|\hat{u}_n^{(\mathrm{noise})}\|_{H^{m+1}(\Omega)}^2$ are finite, so that we can safely use conditional expectations on $\hat{u}_n$, $\hat{u}_n^\star$, and $\hat{u}_n^{(\mathrm{noise})}$. Recall that, since $\lambda_t \|\hat{u}_n\|_{H^{m+1}(\Omega)}^2 \leqslant \mathscr{R}_n^{(\mathrm{reg})}(\hat{u}_n) \leqslant \mathscr{R}_n^{(\mathrm{reg})}(0)$, and since $\mathscr{F}_k^{(\mathrm{lin})}(0, \cdot) = 0$,

$$\lambda_t \|\hat{u}_n\|_{H^{m+1}(\Omega)}^2 \leqslant \frac{\lambda_d}{n}\sum_{i=1}^n \|Y_i\|_2^2 + \lambda_e \mathbb{E}\|h(\mathbf{X}^{(e)})\|_2^2 + \frac{1}{|\Omega|}\sum_{k=1}^M \int_\Omega B_k(\mathbf{x})^2 d\mathbf{x}.$$

Hence,

$$\mathbb{E}\|\hat{u}_n\|^2_{H^{m+1}(\Omega)} \leqslant \lambda_t^{-1}\Big(\lambda_d \mathbb{E}\|u^\star(\mathbf{X}) + \varepsilon\|_2^2 + \lambda_e \mathbb{E}\|h(\mathbf{X}^{(e)})\|_2^2 + \frac{1}{|\Omega|}\sum_{k=1}^M \int_\Omega B_k(\mathbf{x})^2 d\mathbf{x}\Big).$$

Similarly,

$$\mathbb{E}\|\hat{u}_n^\star\|^2_{H^{m+1}(\Omega)} \leqslant \lambda_t^{-1}\Big(\lambda_d \mathbb{E}\|u^\star(\mathbf{X})\|_2^2 + \lambda_e \mathbb{E}\|h(\mathbf{X}^{(e)})\|_2^2 + \frac{1}{|\Omega|}\sum_{k=1}^M \int_\Omega B_k(\mathbf{x})^2 d\mathbf{x}\Big),$$

and $\mathbb{E}\|\hat{u}_n^{(\mathrm{noise})}\|^2_{H^{m+1}(\Omega)} \leqslant \lambda_t^{-1}\lambda_d \mathbb{E}\|\varepsilon\|_2^2$.

Step 3: Bias-variance decomposition. In this paragraph, we use the notation $\mathcal{A}_{(\mathbf{x},e)}(u,u)$ instead of $\mathcal{A}_n(u,u)$, to make the dependence of $\mathcal{A}_n$ in the random variables $\mathbf{x} = (\mathbf{X}_1,\dots,\mathbf{X}_n)$ and $e = (\varepsilon_1,\dots,\varepsilon_n)$ more explicit. We do the same with $\mathcal{B}_n$ and $\hat{u}_n^{(\mathrm{noise})}$. Observe that, for any $(\mathbf{x},e) \in \Omega^n \times \mathbb{R}^{nd_2}$ and for any $u \in H^{m+1}(\Omega,\mathbb{R}^{d_2})$, one has

$$\mathcal{A}_{(\mathbf{x},-e)}(u,u) - 2\mathcal{B}^{(\mathrm{noise})}_{(\mathbf{x},e)}(u) = \mathcal{A}_{(\mathbf{x},e)}(-u,-u) - 2\mathcal{B}^{(\mathrm{noise})}_{(\mathbf{x},-e)}(-u).$$

Therefore, $\hat{u}^{(\mathrm{noise})}_{(\mathbf{x},e)} = -\hat{u}^{(\mathrm{noise})}_{(\mathbf{x},-e)}$.

Since, by assumption, ε has the same law as $-\varepsilon$, this implies $\mathbb{E}(\hat{u}_n^{(\mathrm{noise})} \mid \mathbf{X}_1,\dots,\mathbf{X}_n) = 0$, and so $\mathbb{E}(\hat{u}_n^{(\mathrm{noise})}) = 0$. Moreover, since $\hat{u}_n^\star$ is a measurable function of $\mathbf{X}_1,\dots,\mathbf{X}_n$, we have $\mathbb{E}(\hat{u}_n^\star \mid \mathbf{X}_1,\dots,\mathbf{X}_n) = \hat{u}_n^\star$. Recalling (Step 1) that $\hat{u}_n = \hat{u}_n^\star + \hat{u}_n^{(\mathrm{noise})}$, we deduce the following bias-variance decomposition:

$$\mathbb{E}\|\hat{u}_n - u^\star\|^2_{L^2(\mu_{\mathbf{X}})} = \mathbb{E}\|\hat{u}_n^\star - u^\star\|^2_{L^2(\mu_{\mathbf{X}})} + \mathbb{E}\|\hat{u}_n^{(\mathrm{noise})}\|^2_{L^2(\mu_{\mathbf{X}})}. \tag{36}$$

Step 4: Bounding the bias. Recall that $\hat{u}_n^\star$ minimizes $\mathscr{R}_n^\star$ over $H^{m+1}(\Omega,\mathbb{R}^{d_2})$, so that $\mathscr{R}_n^\star(u^\star) \geqslant \mathscr{R}_n^\star(\hat{u}_n^\star)$. Therefore, $\mathrm{PI}(u^\star) + \lambda_t\|u^\star\|^2_{H^{m+1}(\Omega)} \geqslant \frac{\lambda_d}{n}\sum_{i=1}^n \|\tilde{\Pi}(\hat{u}_n^\star - u^\star)(\mathbf{X}_i)\|_2^2$. We deduce that

$$\begin{aligned}
&\frac{1}{\lambda_d}\big(\mathrm{PI}(u^\star) + \lambda_t\|u^\star\|^2_{H^{m+1}(\Omega)}\big) \\
&\geqslant \frac{\|\hat{u}_n^\star - u^\star\|^2_{H^{m+1}(\Omega)}}{n}\sum_{i=1}^n \Big\|\tilde{\Pi}\Big(\frac{\hat{u}_n^\star - u^\star}{\|\hat{u}_n^\star - u^\star\|_{H^{m+1}(\Omega)}}\Big)(\mathbf{X}_i)\Big\|_2^2 \\
&\geqslant \|\hat{u}_n^\star - u^\star\|^2_{L^2(\mu_{\mathbf{X}})} \\
&\quad - \|\hat{u}_n^\star - u^\star\|^2_{H^{m+1}(\Omega)} \sup_{\|u\|_{H^{m+1}(\Omega)}\leqslant 1}\Big(\mathbb{E}\|\tilde{\Pi}(u)(\mathbf{X})\|_2^2 - \frac{1}{n}\sum_{i=1}^n \|\tilde{\Pi}(u)(\mathbf{X}_i)\|_2^2\Big) \\
&\geqslant \|\hat{u}_n^\star - u^\star\|^2_{L^2(\mu_{\mathbf{X}})} \\
&\quad - 2\big(\|\hat{u}_n^\star\|^2_{H^{m+1}(\Omega)} + \|u^\star\|^2_{H^{m+1}(\Omega)}\big) \sup_{\|u\|_{H^{m+1}(\Omega)}\leqslant 1}\Big(\mathbb{E}\|\tilde{\Pi}(u)(\mathbf{X})\|_2^2 - \frac{1}{n}\sum_{i=1}^n \|\tilde{\Pi}(u)(\mathbf{X}_i)\|_2^2\Big).
\end{aligned}$$

Moreover, $\mathrm{PI}(u^\star) + \lambda_t\|u^\star\|^2_{H^{m+1}(\Omega)} \geqslant \lambda_t\|\hat{u}_n^\star\|^2_{H^{m+1}(\Omega)}$. Taking expectations, we conclude by Lemma C.14 that there exists a constant C'_Ω, depending only on Ω, such that

$$\mathbb{E}\|\hat{u}_n^\star - u^\star\|^2_{L^2(\mu_{\mathbf{X}})} \leqslant \frac{1}{\lambda_d}\big(\mathrm{PI}(u^\star) + \lambda_t\|u^\star\|^2_{H^{m+1}(\Omega)}\big) + \frac{C'_\Omega d_2^{1/2}}{n^{1/2}}\Big(2\|u^\star\|^2_{H^{m+1}(\Omega)} + \frac{\mathrm{PI}(u^\star)}{\lambda_t}\Big).$$

Step 5: Bounding the variance. Since $\hat{u}_n^{(\text{noise})}$ minimizes $\mathscr{R}_n^{(\text{noise})}$ over $H^{m+1}(\Omega, \mathbb{R}^{d_2})$, we have $\mathscr{R}_n^{(\text{noise})}(0) \geqslant \mathscr{R}_n^{(\text{noise})}(\hat{u}_n^{(\text{noise})})$. So,

$$\frac{\lambda_d}{n}\sum_{i=1}^n \|\varepsilon_i\|_2^2 \geqslant \frac{\lambda_d}{n}\sum_{i=1}^n \|\tilde{\Pi}(\hat{u}_n^{(\text{noise})})(\mathbf{X}_i) - \varepsilon_i\|_2^2.$$

Observing that $\|\tilde{\Pi}(\hat{u}_n^{(\text{noise})})(\mathbf{X}_i) - \varepsilon_i\|_2^2 = \|\tilde{\Pi}(\hat{u}_n^{(\text{noise})})(\mathbf{X}_i)\|_2^2 - 2\langle\tilde{\Pi}(\hat{u}_n^{(\text{noise})})(\mathbf{X}_i), \varepsilon_i\rangle + \|\varepsilon_i\|_2^2$, we deduce that

$$\frac{2}{n}\sum_{i=1}^n \langle\tilde{\Pi}(\hat{u}_n^{(\text{noise})})(\mathbf{X}_i), \varepsilon_i\rangle \geqslant \frac{1}{n}\sum_{i=1}^n \|\tilde{\Pi}(\hat{u}_n^{(\text{noise})})(\mathbf{X}_i)\|_2^2,$$

and

$$\begin{aligned}&\Big\langle \int_\Omega \tilde{\Pi}(\hat{u}_n^{(\text{noise})}) d\mu_{\mathbf{X}}, \frac{2}{n}\sum_{i=1}^n \varepsilon_i\Big\rangle + \frac{2}{n}\sum_{i=1}^n \Big\langle \tilde{\Pi}(\hat{u}_n^{(\text{noise})})(\mathbf{X}_i) - \int_\Omega \tilde{\Pi}(\hat{u}_n^{(\text{noise})}) d\mu_{\mathbf{X}}, \varepsilon_i\Big\rangle \\ &\quad\geqslant \frac{1}{n}\sum_{i=1}^n \|\tilde{\Pi}(\hat{u}_n^{(\text{noise})})(\mathbf{X}_i)\|_2^2.\end{aligned}$$

Therefore,

$$\begin{aligned}\|\hat{u}_n^{(\text{noise})}\|_{L^2(\mu_{\mathbf{X}})}^2 &\leqslant \Big\langle \int_\Omega \tilde{\Pi}(\hat{u}_n^{(\text{noise})}) d\mu_{\mathbf{X}}, \frac{2}{n}\sum_{i=1}^n \varepsilon_i\Big\rangle \\ &\quad + \|\hat{u}_n^{(\text{noise})}\|_{H^{m+1}(\Omega)} \sup_{\|u\|_{H^{m+1}(\Omega)}\leqslant 1} \frac{1}{n}\sum_{j=1}^n \langle\tilde{\Pi}(u)(\mathbf{X}_j) - \mathbb{E}(\tilde{\Pi}(u)(\mathbf{X})), \varepsilon_j\rangle \\ &\quad + \|\hat{u}_n^{(\text{noise})}\|_{H^{m+1}(\Omega)}^2 \sup_{\|u\|_{H^{m+1}(\Omega)}\leqslant 1} \Big(\mathbb{E}\|\tilde{\Pi}(u)(\mathbf{X}_i)\|_2^2 - \frac{1}{n}\sum_{i=1}^n \|\tilde{\Pi}(u)(\mathbf{X}_i)\|_2^2\Big) \\ &:= A + B + C.\end{aligned}$$

According to the Cauchy-Schwarz inequality,

$$\mathbb{E}(A) \leqslant \Big(\mathbb{E}\Big\| \int_\Omega \tilde{\Pi}(\hat{u}_n^{(\text{noise})}) d\mu_{\mathbf{X}}\Big\|_2^2\Big)^{1/2} \times \frac{2(\mathbb{E}\|\varepsilon\|_2^2)^{1/2}}{n^{1/2}},$$

and so, by Jensen's inequality,

$$\mathbb{E}(A) \leqslant \big(\mathbb{E}\|\hat{u}_n^{(\text{noise})}\|_{L^2(\mu_{\mathbf{X}})}^2\big)^{1/2} \times \frac{2(\mathbb{E}\|\varepsilon\|_2^2)^{1/2}}{n^{1/2}}.$$

The inequality $\mathscr{R}_n^{(\text{noise})}(0) \geqslant \mathscr{R}_n^{(\text{noise})}(\hat{u}_n^{(\text{noise})})$ also implies that

$$\frac{\lambda_d}{n}\sum_{i=1}^n \|\varepsilon_i\|_2^2 \geqslant \frac{\lambda_d}{n}\sum_{i=1}^n \|\tilde{\Pi}(\hat{u}_n^{(\text{noise})})(\mathbf{X}_i) - \varepsilon_i\|_2^2 + \lambda_t \|\hat{u}_n^{(\text{noise})}\|_{H^{m+1}(\Omega)}^2.$$

Therefore,

$$\frac{\lambda_d}{n\lambda_t}\sum_{i=1}^n 2\langle\tilde{\Pi}(\hat{u}_n^{(\text{noise})})(\mathbf{X}_i), \varepsilon_i\rangle \geqslant \|\hat{u}_n^{(\text{noise})}\|_{H^{m+1}(\Omega)}^2,$$

and

$$\frac{\lambda_d}{\lambda_t}\sup_{\|u\|_{H^{m+1}(\Omega)}\leqslant 1}\frac{1}{n}\sum_{j=1}^{n}\langle\tilde{\Pi}(u)(\mathbf{X}_j),\varepsilon_j\rangle \geqslant \|\hat{u}_n^{(\mathrm{noise})}\|_{H^{m+1}(\Omega)}.$$

By Theorem B.1, if $\|u\|_{H^{m+1}(\Omega)}\leqslant 1$, then $\langle\mathbb{E}(\tilde{\Pi}(u)(\mathbf{X})),\frac{1}{n}\sum_{j=1}^{n}\varepsilon_j\rangle \leqslant \frac{C_\Omega d_2^{1/2}}{n}\|\sum_{i=1}^{n}\varepsilon_i\|_2$. Thus,

$$\begin{aligned}&\|\hat{u}_n^{(\mathrm{noise})}\|_{H^{m+1}(\Omega)}\\ &\leqslant \frac{\lambda_d}{\lambda_t}\Big(\frac{C_\Omega d_2^{1/2}}{n}\|\sum_{i=1}^{n}\varepsilon_i\|_2+\sup_{\|u\|_{H^{m+1}(\Omega)}\leqslant 1}\frac{1}{n}\sum_{j=1}^{n}\langle\tilde{\Pi}(u)(\mathbf{X}_j)-\mathbb{E}(\tilde{\Pi}(u)(\mathbf{X})),\varepsilon_j\rangle\Big).\end{aligned}$$

Using Lemma C.15 together with the fact that, for all $\mathbf{x},\mathbf{y}\in\mathbb{R}$, $(\mathbf{x}+\mathbf{y})^2\leqslant 2(\mathbf{x}^2+\mathbf{y}^2)$,

$$\mathbb{E}\|\hat{u}_n^{(\mathrm{noise})}\|_{H^{m+1}(\Omega)}^2\leqslant\frac{4\lambda_d^2}{n\lambda_t^2}C_\Omega^2 d_2\mathbb{E}\|\varepsilon\|_2^2.$$

Similarly, observing that for all random variables $X,Y\in\mathbb{R}$, $\mathbb{E}(XY)^2\leqslant\mathbb{E}(X^2)\mathbb{E}(Y^2)$,

$$\mathbb{E}(B)\leqslant\frac{4\lambda_d}{n\lambda_t}C_\Omega^2 d_2\mathbb{E}\|\varepsilon\|_2^2.$$

Moreover, by Lemma C.14 and the inequality $\mathbb{E}(XYZ)^2\leqslant\mathbb{E}(X^2)\mathbb{E}(Y^2)\mathbb{E}(Z^2)$,

$$\mathbb{E}(C)\leqslant\frac{\lambda_d^2}{n^{3/2}\lambda_t^2}C_\Omega^2 d_2^{3/2}\mathbb{E}\|\varepsilon\|_2^2.$$

Therefore, we conclude that there exists a constant $C_\Omega>0$, depending only on Ω, such that

$$\begin{aligned}\mathbb{E}\|\hat{u}_n^{(\mathrm{noise})}\|_{L^2(\mu_\mathbf{X})}^2&\leqslant\big(\mathbb{E}\|\hat{u}_n^{(\mathrm{noise})}\|_{L^2(\mu_\mathbf{X})}^2\big)^{1/2}\frac{2(\mathbb{E}\|\varepsilon\|_2^2)^{1/2}}{n^{1/2}}\\&\quad+\frac{4\lambda_d}{n\lambda_t}C_\Omega^2 d_2\mathbb{E}\|\varepsilon\|_2^2+\frac{\lambda_d^2}{n^{3/2}\lambda_t^2}C_\Omega^2 d_2^{3/2}\mathbb{E}\|\varepsilon\|_2^2.\end{aligned}$$

Hence, using elementary algebra,

$$\big(\mathbb{E}\|\hat{u}_n^{(\mathrm{noise})}\|_{L^2(\mu_\mathbf{X})}^2\big)^{1/2}\leqslant\frac{(\mathbb{E}\|\varepsilon\|_2^2)^{1/2}}{n^{1/2}}\Big(2+2C_\Omega d_2^{3/4}\Big(\frac{\lambda_d^{1/2}}{\lambda_t^{1/2}}+\frac{\lambda_d}{\lambda_t n^{1/4}}\Big)\Big)$$

and

$$\mathbb{E}\|\hat{u}_n^{(\mathrm{noise})}\|_{L^2(\mu_\mathbf{X})}^2\leqslant\frac{8\mathbb{E}\|\varepsilon\|_2^2}{n}\Big(1+C_\Omega d_2^{3/2}\Big(\frac{\lambda_d}{\lambda_t}+\frac{\lambda_d^2}{\lambda_t^2 n^{1/2}}\Big)\Big).$$

Step 6: Putting everything together. Combining Steps 3, 4, and 5, we conclude that

$$\begin{aligned}\mathbb{E}\|\hat{u}_n-u^\star\|_{L^2(\mu_\mathbf{X})}^2&\leqslant\frac{1}{\lambda_d}\big(\mathrm{PI}(u^\star)+\lambda_t\|u^\star\|_{H^{m+1}(\Omega)}^2\big)+\frac{C'_\Omega d_2^{1/2}}{n^{1/2}}\Big(2\|u^\star\|_{H^{m+1}(\Omega)}^2+\frac{\mathrm{PI}(u^\star)}{\lambda_t}\Big)\\&\quad+\frac{8\mathbb{E}\|\varepsilon\|_2^2}{n}\Big(1+C_\Omega d_2^{3/2}\Big(\frac{\lambda_d}{\lambda_t}+\frac{\lambda_d^2}{\lambda_t^2 n^{1/2}}\Big)\Big).\end{aligned}$$

G.6. Proof of Proposition 5.12. By definition, $\hat{u}_n$ minimizes $\mathscr{R}_n^{(\mathrm{reg})}$ over $H^{m+1}(\Omega, \mathbb{R}^{d_2})$. So, $\mathscr{R}_n^{(\mathrm{reg})}(u^\star) \geqslant \mathscr{R}_n^{(\mathrm{reg})}(\hat{u}_n)$. Moreover, since

$$\|\tilde{\Pi}(\hat{u}_n)(\mathbf{X}_i) - Y_i\|_2^2 = \|\tilde{\Pi}(\hat{u}_n - u^\star)(\mathbf{X}_i)\|_2^2 - 2\langle \tilde{\Pi}(\hat{u}_n - u^\star)(\mathbf{X}_i), \varepsilon_i\rangle + \|\varepsilon_i\|_2^2,$$

one has

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \|\tilde{\Pi}(\hat{u}_n)(\mathbf{X}_i) - Y_i\|_2^2 \\
&\quad \geqslant -2\|\hat{u}_n - u^\star\|_{H^{m+1}(\Omega)} \times \sup_{\|u\|_{H^{m+1}(\Omega)} \leqslant 1} \frac{1}{n}\sum_{j=1}^n \langle \tilde{\Pi}(u)(\mathbf{X}_j) - \mathbb{E}(\tilde{\Pi}(u)(\mathbf{X})), \varepsilon_j\rangle \\
&\qquad - 2\Big\langle \int_\Omega \tilde{\Pi}(\hat{u}_n - u^\star) d\mu_{\mathbf{X}}, \frac{1}{n}\sum_{i=1}^n \varepsilon_i \Big\rangle + \frac{1}{n}\sum_{i=1}^n \|\varepsilon_i\|_2^2.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \|\tilde{\Pi}(\hat{u}_n)(\mathbf{X}_i) - Y_i\|_2^2 \\
&\quad \geqslant -2(\|\hat{u}_n\|_{H^{m+1}(\Omega)} + \|u^\star\|_{H^{m+1}(\Omega)}) \sup_{\|u\|_{H^{m+1}(\Omega)} \leqslant 1} \frac{1}{n}\sum_{j=1}^n \langle \tilde{\Pi}(u)(\mathbf{X}_j) - \mathbb{E}(\tilde{\Pi}(u)(\mathbf{X})), \varepsilon_j\rangle \\
&\qquad - 2\Big\langle \int_\Omega \tilde{\Pi}(\hat{u}_n - u^\star) d\mu_{\mathbf{X}}, \frac{1}{n}\sum_{i=1}^n \varepsilon_i \Big\rangle + \frac{1}{n}\sum_{i=1}^n \|\varepsilon_i\|_2^2.
\end{aligned} \tag{37}$$

Recall from Steps 4 and 5 of the proof of Theorem 5.11 that

$$\begin{aligned}
\mathbb{E}\|\hat{u}_n\|_{H^{m+1}(\Omega)}^2 &\leqslant 2\mathbb{E}\|\hat{u}_n^\star\|_{H^{m+1}(\Omega)}^2 + 2\mathbb{E}\|\hat{u}_n^{(\mathrm{noise})}\|_{H^{m+1}(\Omega)}^2 \\
&\leqslant 2\Big(\frac{\mathrm{PI}(u^\star)}{\lambda_t} + \|u^\star\|_{H^{m+1}(\Omega)}^2\Big) + \frac{8\lambda_d^2}{n\lambda_t^2} C_\Omega^2 d_2 \mathbb{E}\|\varepsilon\|_2^2
\end{aligned}$$

Therefore, Lemma C.15 and the inequality $\mathbb{E}(XY)^2 \leqslant \mathbb{E}(X)^2\mathbb{E}(Y)^2$ show that

$$\mathbb{E}\Big(\|\hat{u}_n\|_{H^{m+1}(\Omega)} \sup_{\|u\|_{H^{m+1}(\Omega)} \leqslant 1} \frac{1}{n}\sum_{j=1}^n \langle \tilde{\Pi}(u)(\mathbf{X}_j) - \mathbb{E}(\tilde{\Pi}(u)(\mathbf{X})), \varepsilon_j\rangle\Big) = \mathcal{O}_{n\to\infty}\Big(\frac{\lambda_d}{n\lambda_t}\Big).$$

By Theorem 5.11,

$$\mathbb{E}\Big|\Big\langle \int_\Omega \tilde{\Pi}(\hat{u}_n - u^\star) d\mu_{\mathbf{X}}, \frac{1}{n}\sum_{i=1}^n \varepsilon_i \Big\rangle\Big| \leqslant \big(\mathbb{E}\|u^\star - \hat{u}_n\|_{L^2(\mu_{\mathbf{X}})}^2\big)^{1/2} \frac{\mathbb{E}\|\varepsilon\|_2^2}{n^{1/2}} = \mathcal{O}_{n\to\infty}\Big(\frac{\lambda_d}{n^2\lambda_t}\Big)^{1/2}.$$

Combining these three results with (37), we conclude that

$$\mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^n \|\tilde{\Pi}(\hat{u}_n)(\mathbf{X}_i) - Y_i\|_2^2\Big) \geqslant \mathbb{E}\|\varepsilon\|_2^2 + \mathcal{O}_{n\to\infty}\Big(\frac{\lambda_d}{n\lambda_t}\Big).$$

Therefore, since $\lim_{n\to\infty} \frac{\lambda_d^2}{n\lambda_t} = 0$ and since $\mathscr{R}_n^{(\mathrm{reg})}(\hat{u}_n) = \frac{\lambda_d}{n}\sum_{i=1}^n \|\tilde{\Pi}(\hat{u}_n)(\mathbf{X}_i) - Y_i\|_2^2 + \mathrm{PI}(\hat{u}_n) + \lambda_t\|\hat{u}_n\|_{H^{m+1}(\Omega)}^2$,

$$\mathbb{E}\big(\mathscr{R}_n^{(\mathrm{reg})}(\hat{u}_n)\big) \geqslant \lambda_d\mathbb{E}\|\varepsilon\|_2^2 + \mathbb{E}(\mathrm{PI}(\hat{u}_n)) + o_{n\to\infty}(1).$$

Similarly, almost everywhere,

$$\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\Pi}(\hat{u}^\star)(\mathbf{X}_i)-Y_i\|_2^2=\frac{1}{n}\sum_{i=1}^{n}\|\varepsilon_i\|_2^2.$$

Hence,

$$\mathbb{E}\big(\mathscr{R}_n^{(\mathrm{reg})}(u^\star)\big)=\lambda_d\mathbb{E}\|\varepsilon\|_2^2+\mathrm{PI}(u^\star)+\lambda_t\|u^\star\|_{H^{m+1}(\Omega)}^2.$$

Since $\mathbb{E}(\mathscr{R}_n^{(\mathrm{reg})}(\hat{u}_n))\leqslant\mathbb{E}(\mathscr{R}_n^{(\mathrm{reg})}(u^\star))$ and since $\lambda_t\to 0$, we are led to

$$\mathbb{E}(\mathrm{PI}(\hat{u}_n))\leqslant\mathrm{PI}(u^\star)+o_{n\to\infty}(1),$$

which is the desired result.