Some elliptic second order problems and neural network solutions: Existence and error estimates

Jerome Pousin

Journal of Computational and Applied Mathematics 436 (2024) 115398

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Some elliptic second order problems and neural network solutions: Existence and error estimates

Jerome Pousin

Université de Lyon, Institut C. Jordan, INSA de Lyon CNRS (UMR5208), 20 Av. A. Einstein, 69100 Villeurbanne Cedex, France

ARTICLE INFO

Article history:
Received 6 April 2023
Received in revised form 2 June 2023

MSC:
00-01
99-00 65L09
65D05
65G99

Keywords:
Learning neural network
Elliptic PDE
Gradient flow strategy
Error estimates

ABSTRACT

Recently some neural networks have been proposed for computing approximate solutions to partial differential equations. For some second order elliptic or parabolic PDEs, error estimates are proved between the solution and the computed one with neural networks, assuming this one minimizes a L^p norm or a dual norm of the residual, or an abstract loss function. In this article, for some second order elliptic PDEs, thanks to a gradient flow strategy, we prove the existence of a neural network solution minimizing the loss function with respect to the neural network parameters and we give an error estimate between the solution and the computed one with neural networks. For some nonsymmetric elliptic PDEs, the problem is expressed in form of a MinMax problem which is approximating with a double NN. Thanks to a diagonal extraction process a result of convergence is established with respect to the parameters of NNs and errors estimates are also given.

1. Introduction

Feedforward neural networks (FNNs), or machine learning have become very popular in computer science in recent years, particularly for images processing and signal processing. These methods aim to approximate functions with layers neurons (or units), connected by linear operations between units and nonlinear activations within units, by minimizing a loss function J over a learning set, see [1–3] and references therein. Motivated by the performance of deep learning-based solutions in classical machine learning tasks, and since FNNs do not require meshes allowing to consider problems in large space dimension, FNNs are, now some time, used for computing approximate solutions of some PDEs. In [4], it is proved that a ReLU deep neural networks has the ability to represent the basis functions of simplicial finite element of order one, allowing so to approximate elliptic second order PDE problems which can be formulated as minimizing an energy functional as in [5]. For some elliptic second order PDEs, generally, the loss functions are based on a least squares formulation of the residual of the PDE in L^p norm or in a dual norm of the residual (see [6,7] for example), eventually augmented with a penalized term for the boundary conditions and a regularization term. In the first case, error estimates have been proposed between a NN-solution of the PDE and the exact solution in [8,9]. In the second case, if the elliptic operator is not symmetric, the problem can be formulated in form of a Min-Max problem, that optimization problem is used for calculating NN-solutions (also called adversial solutions). Approximating a Min-Max problem requires to deal with two NN. Error estimates have also been proposed in [10,11] and [12] for example. Here we have to notice that, in both cases, the NN-solutions considered minimize an abstract formulation of the loss function, that is to say that the optimization process is with respect to NN-functions as elements Sobolev spaces and not with respect to parameters of the NN. Unfortunately, it is well known that the realization spaces of NN-functions have very few topological properties [13].

E-mail address: jerome.pousin@insa-lyon.fr.

https://doi.org/10.1016/j.cam.2023.115398
0377-0427/

In [14], by using a temporal discretization, nonlinear parabolic PDEs are approximatively solved with a FNN. The loss function consists of a L^2-least squares formulation of the equation augmented with the boundary and initial conditions. The neural network is then trained to minimize, with a stochastic gradient descent, the loss functional by a discretization of integrals and randomly sampling spatial points. A similar way is adopted in [15].

For elliptic PDEs, The loss function minimizing problems have a gradient flow structure (see Section 3), which validates that a gradient descent (respectively ascent) strategy converges towards a minimum (respectively a maximum). In this paper, for some elliptic problems, it is proved that a gradient technic applied to the NN parameters set converges towards a minimum of the loss function and errors estimates are given. For the nonsymmetric elliptic PDEs, the MinMax problem is approximating with a double NN, and thanks to a diagonal extraction process a result of convergence is established and Errors estimates are also given.

The paper is organized as follows. Second section is dedicated to some properties of NNs. Third section consider self adjoint elliptic operator. A collocation method with a Gauss quadrature is introduced for evaluating the residual, and the NN is trained with a gradient descent technic with respect to parameters. A convergence result is given and error estimates are presented. Section 3 investigates the case of nonsymmetric elliptic operator and the Min-Max problem associated. By using a diagonal extraction principle an existence results is given and errors estimates are proved. The section is ended with some remarks concerning the way in which boundary conditions are accounted.

2. Neural networks

Consider a NN with a scalar-valued output $N(L,\theta)$ defined as composition of many layers of functions:

$$N(L,\theta) = \mathcal{N}(L,\omega_L,b_L)\circ\mathcal{N}(L-1,\omega_{L-1},b_{L-1})\cdots\mathcal{N}(2,\omega_2,b_2)\circ\mathcal{N}(1,\omega_1,b_1), \tag{1}$$

where the symbol $\circ$ denotes the composition of functions, L is the depth of the network and $\mathcal{N}(l,\omega_l,b_l)$ is called the lth hidden layer of the network for $1 \le l \le L-1$. A layer $\mathcal{N}(l,\omega_l,b_l) : \mathbb{R}^{d_{l-1}} \to \mathbb{R}^{d_l}$ is defined as composition of an affine transformation $\mathbb{R}^{d_{l-1}} \underset{x_{l-1}:\to\omega_l x_{l-1}+b_l}{\longrightarrow} \mathbb{R}^{d_l}$ where ω_l is a $d\times(d-1)$ weights matrix, b_l a $\mathbb{R}^d$ bias vector, with an activation function

$$\begin{array}{rccc}\sigma: & \mathbb{R} & \longrightarrow & \mathbb{R}\\ & t & \mapsto & t^+ = max(0,t)\end{array}$$

or the hyperbolic tangent or the sigmoid function if more regularity is required for the functions that the NN yield. We have

$$\mathcal{N}(l,\omega_l,b_l) = \sigma(\omega_l x_{l-1} + b_l) \tag{2}$$

here the function σ acts component-wise for a vector. The last layer is a linear transformation. Each component of the vector valued function $\mathcal{N}(l,\omega_l,b_l)$ is seen as a neuron and its dimension defines the width or the number of neurons of the lth layer. The set of $\sum_{l=1}^{L} d_l\times(d_{l-1}+1)$ parameters of the NN is denoted by θ, and its cardinal is denoted by $|\theta|$.

Let us note that Z, the set of all NN defined by (1), is a finite dimension vector space the dimension of which is $|\theta|$, which is isomorphic to $\mathbb{R}^{|\theta|}$. This vector space is equipped with the following norm:

$$\|N(L,\theta)\|_{max} = \max\left(\|N(L,\theta)\|_{\infty,sc}, \max_{1\le l\le L}\|b_l\|_\infty\right)$$

with

$$\|N(L,\theta)\|_{\infty,sc} = \max_{1\le l\le L}\|\omega_l\|_\infty.$$

For d be given, let $\Omega \subset \mathbb{R}^d$ be a convex bounded open subset with a $C^{0,1}$ boundary $\partial\Omega$ with $\overline{\Omega}$ compact, and denote by $C^1(\Omega;\mathbb{R})$ the set of differentiable functions with a continuous derivative. Introduce the realization map $R_\sigma : Z \underset{N(L,\theta)\to R_\sigma N(L,\theta)}{\longrightarrow} C^1(\Omega;\mathbb{R})$ when using the sigmoid function or hyperbolic tangent for function σ which will be assumed in this paper.

Lemma 1. *Let $\Omega \subset \mathbb{R}^d$ be compact, and let σ globally Lipschitz continuous, then there exists a constant $0 < C(\sigma,Z)$, such that*

$$Lipschitz(R_\sigma N(L,\theta)) \le C(\sigma,Z)\|N(L,\theta)\|_{\infty,sc};$$

Moreover, if $\theta \in B_{\mathbb{R}^{|\theta|}}(0,B)$, then $\|R_\sigma N(L,\theta)\|_{W^{1,\infty}(\Omega)}$, $\|\frac{\partial R_\sigma N(L,\theta)}{\partial\tilde{\theta}}\|_{L^\infty(\Omega)}$, and $\|\frac{\partial^2 R_\sigma N(L,\theta)}{\partial\tilde{\theta}^2}\|_{L^\infty(\Omega)}$ are bounded for any parameter $\tilde{\theta}\in\theta$.

Proof. See [13] prop. 4.1 for the bound $C(\sigma, Z)$ of the Lipschitz constant of map R_σ, depending on σ and on the structure of the NN, which implies the bound in $W^{1,\infty}$ norm. The other bounds are consequence of composition of affine functions with functions σ.

2.1. Feedforward neural networks

The parameters (weight matrices and bias vectors) are to be determined with a training set by minimizing a convex functional J usually called loss function.

For given integers $\{d_l\}_{l=1}^{L}$, define $\mathcal{C}(L, \theta)$ the set of realizations of FNN by:

$$\mathcal{C}(L, \theta) = \{x \mapsto R_\sigma N(L, \theta)(x) = \mathcal{N}(L, \omega_L, b_L) \circ \cdots \circ \mathcal{N}(1, \omega_1, b_1)(x)\} \tag{3}$$

where $N(L, \theta)$ is defined by (1), and $\mathcal{N}(l, \omega_l, b_l)$ by (2).

The space $\mathcal{C}(L, \theta)$ is not a vectorial subspace (see [7] Section 2 for a trivial example when σ is a max function, but a star-shaped with respect to 0 subspace of Lipschitz functions). It is not convex, neither closed for the L^p topology (see Theorem 2.1 and Theorem 3.1 in [13]).

2.2. Ability to approximate functions with a NN

Approximating a function with a neural network has been considered for a long time by many authors and is known as the universal approximation property. The smaller the precision, the more neurons in the hidden layers one should take to reach the required precision [16]. For $1 \leq p \leq \infty$, we denote by $L^p(\Omega)$ the standard Lebesgue's space of functions defined on Ω. For $1 \leq n$, the Sobolev's space $W^{n,p}(\Omega)$ is defined as the set of functions in $L^p(\Omega)$, the distributional derivatives of order up to n are in $L^p(\Omega)$:

$$W^{n,p}(\Omega) = \{f \in L^p(\Omega);\ D^\alpha f \in L^p(\Omega),\ 0 < |\alpha| \leq n\}.$$

$W^{n,p}(\Omega)$ is a Banach space when is endowed with the following norm [17,18]

$$\|f\|_{W^{n,p}(\Omega)} = \left(\sum_{0 \leq |\alpha| \leq n} \|D^\alpha f\|^p_{L^p(\Omega)} \right)^{\frac{1}{p}}$$

Now we give the universal approximation property proved in [16]. For a fixed $0 < B$, denote by $\mathcal{B} = \{f \in W^{n,p}(\Omega);\ \|f\|_{W^{n,p}(\Omega)} \leq B\}$ the B-radius ball of $W^{n,p}(\Omega)$.

Theorem 2. *Let $0 < d$; $1 \leq p \leq \infty$; $2 \leq n$; $0 < B$ and $0 \leq s \leq 1$ be given. Then for any $\epsilon \in (0, \frac{1}{2})$ and for any $f \in \mathcal{B}$ then there exists a FNN $N(L, \theta)$, the deep of which is L with a parameters set θ and a function $R_\sigma N(L, \theta_f) \in \mathcal{C}(L, \theta)$ verifying*

$$\|R_\sigma N(L, \theta_f) - f(\cdot)\|_{W^{s,p}(\Omega)} \leq \epsilon. \tag{4}$$

For the case $\Omega = (0, 1)^d$, in [19] Theorem 4.1, the following bounds are provided: there exists $0 < c(d, n, p, B, s)$ be such that

$$L \leq c(d, n, p, B, s) \log_2 (\epsilon^{\frac{-n}{n-s}})$$

and $|\theta|$, the cardinal of parameters set, is bounded by:

$$|\theta| \leq c(d, n, p, B, s) \epsilon^{\frac{-d}{n-s}} \log_2 (\epsilon^{\frac{-n}{n-s}}).$$

Assume $\Omega = (0, 1)^d$, let us end this section with a classical result concerning numerical integration.

Lemma 3. *Let $0 < \mu$ be a small parameter, the interval $(0, 1)$ is divided into subintervals the length of which is μ. Denote by $\{x_i, W_i\}_{i=1}^{M}$ the tensorized Gauss points and weights for Ω, and denote by $\{y_j w_j\}_{j=1}^{m}$ the tensorized Gauss points and weights for $\partial\Omega$,. Then there exist $0 < C_1, C_2$ not depending on μ such that for every $f \in W^{1,\infty}(\Omega) \cap W^{1,\infty}(\partial\Omega)$ we have:*

$$\left| \int_\Omega f(x)\, dx - \sum_{i=1}^{M} W_i f(x_i) \right| \leq C_1(\Omega) \frac{\mu\sqrt{d}}{2} \|f\|_{W^{1,\infty}},$$

$$\left| \int_{\partial\Omega} f(s)\, ds - \sum_{j=1}^{m} w_j f(y_j) \right| \leq C_2(\partial\Omega) \frac{\mu\sqrt{d-1}}{2} \|f\|_{W^{1,\infty}}.$$

Remark that for domains Ω which can be obtained by deforming hypercubes, similar results concerning numerical integration are valid.

3. The case of some second order self-adjoint elliptic PDE in d-dimension

Consider an operator A in a divergence form, $Av = \sum_{i,j=1}^{d} -D_i(a_{ij}D_j v) + a_0 v$, with $a_{ij} = a_{ji} \in C^{1,1}(\overline{\Omega})$; $0 < \underline{a} \leq a_0 \in C^{1,1}(\overline{\Omega})$ and where there exists $0 < \alpha$ such that

$$\alpha \|\xi\|^2_{\mathbb{R}^d} \leq \sum_{i,j=1}^{d} a_{ij}(x)\xi_i\xi_j$$

for all $x \in \overline{\Omega}$ and $\xi \in \mathbb{R}^d$. Problem PE reads: for $g \in C^1(\Omega)$ be given, find $u \in W^{1,2}(\Omega)$ verifying:

$$\begin{array}{l} Au = g \ \text{ in } \Omega; \\ u = 0 \text{ on } \partial\Omega \end{array} \tag{5}$$

Problem PE has one solution in $W^{2,2}(\Omega) \cap \left(W^{1,2}(\Omega) \cap Ker(\gamma)\right)$, where $\gamma : W^{1,2}(\Omega) \to W^{\frac{1}{2},2}(\partial\Omega) \subset L^2(\partial\Omega)$ is the trace operator (see [17] Chapter 2 or [18]).

It is not possible to include the boundary conditions in the space $\mathcal{C}(L,\theta)$, thus a penalization strategy is proposed for the Dirichlet boundary condition $\gamma(u) = 0$ in the same way as in [7]. For $0 < \eta$, define V the Hilbert space $W^{1,2}(\Omega)$ endowed with the following scalar product:

$$(v, w)_\eta = \int_\Omega \sum_{i,j=1}^{d} a_{ij}D_i v D_j w + a_0 v w \, dx + \frac{1}{\eta}\int_{\partial\Omega} v w \, ds, \ \forall v, w.$$

The associated norm is denoted by

$$|||v|||^2_\eta = \int_\Omega \sum_{i,j=1}^{d} a_{ij}D_i v D_j v + a_0 v^2 \, dx + \frac{1}{\eta}\int_{\partial\Omega} v^2 \, ds \tag{6}$$

Lemma 4. *The norm $|||\cdot|||_\eta$ defined by (6) is equivalent to the norm of $W^{1,2}$.*

For a proof see Lemma 3 in [7] where $\int_\Omega |\nabla v|^2 \, dx$ is replaced by $\int_\Omega \sum_{i,j=1}^{d} a_{ij}D_i v D_j v + a_0 v^2 \, dx$.

We have $u \in W^{2,2}(\Omega) \cap \left(W^{1,2}(\Omega) \cap Ker(\gamma)\right)$ the solution to Problem PE verifies the following minimization problem:

$$u = \underset{v \in W^{1,2}(\Omega)\cap Ker(\gamma)}{\text{Argmin}} J_\eta(v) = \frac{1}{2}\int_\Omega \sum_{i,j=1}^{d} a_{ij}D_i v D_j v + a_0 v^2 - 2gv \, dx + \frac{1}{2\eta}\int_{\partial\Omega} v^2 \, ds \tag{7}$$

$$= \underset{v \in W^{1,2}(\Omega)\cap Ker(\gamma)}{\text{Argmin}} J_\eta(v) = \frac{1}{2}\left(|||v-u|||^2_\eta - |||u|||^2_\eta\right) \tag{8}$$

since a variational formulation of Problem PE reads:

$$\int_\Omega \sum_{i,j=1}^{d} a_{ij}D_i u D_j v + a_0 uv - gv \, dx = 0 \ \forall v \in W^{1,2}(\Omega) \cap Ker(\gamma).$$

When approximating Problem PE with a NN, it will be needed to keep parameters θ bounded, therefore a penalized term is added to J_η. Let $0 < B$ be given, define

$$J_\eta(R_\sigma N(L,\theta)) = \frac{1}{2}\left(|||R_\sigma N(L,\theta) - u|||^2_\eta - |||u|||^2_\eta\right) + \frac{1}{4\eta}\sum_{q=1}^{|\theta|}\left(\theta_q^2 - B^2\right)^{+^2} \tag{9}$$

$$\begin{array}{l} \underline{u} = \underset{R_\sigma N(L,\theta)\in C(L,\theta)}{\text{Argmin}} J^C_\eta(R_\sigma N(L,\theta)) = \\ \frac{1}{2}\sum_{i=1}^{M} W_i \left(\sum_{k,l=1}^{d} a_{ij}(x_i) D_l R_\sigma N(L,\theta) D_k R_\sigma N(L,\theta)(x_i) + \right. \\ \left. a_0(x_i)\left(R_\sigma N(L,\theta)\right)^2(x_i) - 2g R_\sigma N(L,\theta)(x_i) \right) + \\ \frac{1}{2\eta}\sum_{j=1}^{m} w_j \left(R_\sigma N(L,\theta)\right)^2(y_j) + \frac{1}{4\eta}\sum_{q=1}^{|\theta|}\left(\theta_q^2 - B^2\right)^{+^2}. \end{array} \tag{10}$$

Now we give an existence result for Problem (10).

Lemma 5. *Assume $\Omega = (0,1)^d$ and σ to be hyperbolic tangent or sigmoid function, then there exists $\underline{u} \in \mathcal{C}(L,\theta)$ solution to Problem (10).*

Proof. In what follows the notation $R_\sigma N(L,\theta)$ will be replaced by $N(\theta)$ with θ a real parameter for keeping the presentation as simple as possible, the case of the NN is a straightforward generalization. First, the case of function J_η is considered, and we show the gradient flow structure of the minimization problem. After simple calculations we check that J_η is differentiable with respect to θ:

$$D_\theta J_\eta(N(\theta)) = \left(N(\theta) - u, \frac{\partial N(\theta)}{\partial \theta}\right)_\eta + \frac{1}{\eta}(\theta^2 - B^2)^+\theta$$

We have:

$$\frac{d}{dt}J_\eta(N(\theta(t))) = D_\theta J_\eta(N(\theta(t)))\frac{d}{dt}\theta(t) =$$
$$\left(\left(N(\theta) - u, \frac{\partial N(\theta(t))}{\partial \theta}\right)_\eta + \frac{1}{\eta}(\theta^2 - B^2)^+\theta\right)\frac{d}{dt}\theta(t)$$

Choose for a positive parameter h

$$\frac{d}{dt}\theta(t) = -hD_\theta J_\eta(N(\theta)),$$

and we get that $t \mapsto J_\eta(N(\theta(t)))$ is a decreasing function, allowing to use a descent gradient strategy for minimizing the function.

Now skip to Problem (10) with the function J_η^G. Introduce the positive semi-definite bilinear form $(\cdot,\cdot)_\eta^G$ defined on $\mathcal{C}(L,\theta)\times\mathcal{C}(L,\theta)$ and the linear form $(g,\cdot)^G$ defined on $\mathcal{C}(L,\theta)$ be such that

$$J_\eta^G(N(\theta)) = \frac{1}{2}\left[(N(\theta),N(\theta))_\eta^G - 2\,(g,N(\theta))^G\right] + \frac{1}{4\eta}\left(\theta^2 - B^2\right)^{+^2}.$$

The derivatives of function J_η^G are given by:

$$D_\theta J_\eta^G(N(\theta)) = \left(N(\theta), \frac{\partial N(\theta)}{\partial \theta}\right)_\eta^G - \left(g, \frac{\partial N(\theta)}{\partial \theta}\right)^G + \tag{11}$$

$$\frac{1}{\eta}(\theta^2 - B^2)^+\theta \tag{12}$$

$$D^2_{\theta^2} J_\eta^G(N(\theta)) = \left(\frac{\partial N(\theta)}{\partial \theta}, \frac{\partial N(\theta)}{\partial \theta}\right)_\eta^G + \left(N(\theta), \frac{\partial^2 N(\theta)}{\partial \theta^2}\right)_\eta^G - \tag{13}$$

$$\left(g, \frac{\partial^2 N(\theta)}{\partial \theta^2}\right)^G + \frac{1}{\eta}\left((\theta^2 - B^2)^+ + 2\theta^2 sgn^+(\theta^2 - B^2)\right) \tag{14}$$

For minimizing J_η^G a gradient descent algorithm is applied, which reads:

- Choose θ_0, $n = 0$,
- * $\theta_{n+1} = \theta_n - hD_\theta J_\eta^G(N(\theta_n))$
- While $|\theta_{n+1} - \theta_n|$ is not small enough do
  - compute $D_\theta J_\eta^G(N(\theta_{n+1}))$
  - n=n+1 and go to *

The function J_η^G is coercitive at infinity ($\lim_{|\theta|\to+\infty} J_\eta^G(\theta) = +\infty$). Thus we can assume there exists $0 < B_1$ such that for minimizing sequences $\theta_n^2 \le B_1^2$, otherwise, if θ_n were not bounded, it would not minimize J_η^G. Lemma 1 claims that $N(\theta_n)$; $\frac{\partial N(\theta_n)}{\partial \theta}$; $\frac{\partial^2 N(\theta_n)}{\partial \theta^2}$ are bounded in L^∞ norm. The second derivative of J_N^G is constituted of scalar products computed with numerical integration, involving $N(\theta_n)$; $\frac{\partial N(\theta_n)}{\partial \theta}$ and $\frac{\partial^2 N(\theta_n)}{\partial \theta^2}$. Thus $|D^2_{\theta^2} J_\eta^G(N(\theta_n) + \xi(N(\theta_{n+1}) - N(\theta_n)))|$ is bounded. The parameter h is chosen such that for all $\theta_n^2, \theta_{n+1}^2 \le B_1^2$ we have:

$$h|D^2_{\theta^2} J_\eta^G(N(\theta_n) + \xi(N(\theta_{n+1}) - N(\theta_n)))| < 1;\ \forall \xi \in (0,1).$$

We have:

$$J_\eta^G(N(\theta_{n+1})) - J_\eta^G(N(\theta_n)) = -h\left(D_\theta J_\eta^G(N(\theta_n))\right)^2 + \frac{h^2}{2}$$
$$D^2_{\theta^2} J_\eta^G(N(\theta_n) + \xi(N(\theta_{n+1}) - N(\theta_n)))\left(D_\theta J_\eta(N(\theta_n))\right)^2$$
$$\le -\frac{h}{2}\left(D_\theta J_\eta^G(N(\theta_n))\right)^2.$$

The sequence $J_\eta^G(N(\theta_n))$ is decreasing, and since $J_\eta^G(N(\theta_n)) \leq J_\eta^G(N(\theta_0)) = 0$ we deduce the following bound for θ_n:

$$\theta_n^2 \leq B^2 + 2\sqrt{\eta}\frac{2}{\underline{a}}\sum_{i=1}^{M} W_i g^2(x_i).$$

We have:

$$-\frac{2}{\underline{a}}\sum_{i=1}^{M} W_i g^2(x_i) \leq J_\eta^G(N(\theta_n)) \leq J_\eta^G(N(\theta_0)).$$

The function J_η^G is bounded from below there exists a subsequence $J_\eta^G(N(\theta_{n_p}))$ and a real a with $J_\eta^G(N(\theta_{n_p})) \to a$. The sequence $J_\eta^G(N(\theta_{n_p}))$ converges.

$$\sum_{p=1}^{\infty} -h\left(D_\theta J_\eta(N(\theta_n))\right)^2 + \frac{h^2}{2}$$
$$D^2_{\theta^2} J_\eta^G(N(\theta_n) + \xi(N(\theta_{n+1}) - N(\theta_n)))\left(D_\theta J_\eta(N(\theta_n))\right)^2;$$

is bounded from above by

$$\sum_{p=1}^{\infty} -\frac{h}{2}\left(D_\theta J_\eta(N(\theta_n))\right)^2 \text{ and thus since } J_\eta^G(N(\theta_{n_p})) \text{ converges}$$

$$\sum_{p=1}^{\infty} \frac{h}{2}\left(D_\theta J_\eta(N(\theta_n))\right)^2 \text{ converges .}$$

There exists a constant $C(a_0)$ such that

$$0 \leq J_\eta(N(\theta_{n_p})) + C(a_0)(g,g)^G.$$

Then we have the following estimates:

$$\left[J_\eta^G(N(\theta_{n_p})) + C(a_0)(g,g)^G\right]^{\frac{1}{2}} - \left[J_\eta^G(N(\theta_{n_p+1})) + C(a_0)(g,g)^G\right]^{\frac{1}{2}} \leq$$
$$\left|\left[J_\eta^G(N(\theta_{n_p})) + C(a_0)(g,g)^G\right]^{\frac{1}{2}} - \left[J_\eta^G(N(\theta_{n_p+1})) + C(a_0)(g,g)^G\right]^{\frac{1}{2}}\right| \leq$$
$$\left[J_\eta^G(N(\theta_{n_p})) - J_\eta^G(N(\theta_{n_p+1}))\right]^{\frac{1}{2}} \leq \left[\frac{h}{2} D_\theta J_\eta(N(\theta_{n_p}))^2\right]^{\frac{1}{2}}.$$

The function $\sqrt{(\cdot)}$ is continuous, thus the sequence

$$\{\left[J_\eta(N(\theta_{n_p})) + C(a_0)(g,g)\right]^{\frac{1}{2}}\}_{p=1}^{\infty}$$

converges and the serie $\sqrt{\frac{h}{2}}\sum_{p=1}^{\infty} D_\theta J_\eta(N(\theta_{n_p}))$ absolutely converges. We get the existence of a subsequence $\{\theta_{n_p}\}_{p\in\mathbb{N}}$ converging towards $\underline{\theta}$. Since the realization map $R_\sigma \in C^0(Z; C^1(\overline{\Omega}; \mathbb{R}))$ we have

$$R_\sigma N(\theta_{n_p}) \to R_\sigma N(\underline{\theta}),$$

and $J_\eta^G(R_\sigma N(\theta_{n_p})) \to J_\eta^G(R_\sigma N(\underline{\theta}))$ which is a solution to Problem (10).

In what follows, we deal with the expression (7) for function J_η since function u is not known. We have the following existence and error estimate for Problem (10):

Theorem 6. *Assume $\Omega = (0,1)^d$ and σ to be the hyperbolic tangent or the sigmoid function, let $0 < \eta$ and $0 < \epsilon < \frac{1}{2}$ be given, denote by $u \in W^{2,2}(\Omega)$ the solution to Problem (5) and by $R_\sigma N(L,\theta_u) \in \mathcal{C}(L,\theta)$ the approximation of u into $\mathcal{C}(L,\theta)$ given by Theorem 2. Then there exists $\underline{u} \in \mathcal{C}(L,\theta)$ solution to Problem (10). Moreover, assuming $\theta_u \in B(0,B)$ and $R_\sigma N(L,\theta_u)$ is not a minima of J_η^G, then there exists a constant $0 < C$ independent of η and ϵ be such that*

$$\frac{1}{4}\|\underline{u} - u\|_\eta^2 \leq \frac{3}{4}\|R_\sigma N(L,\theta_u) - u\|_\eta^2 + 8\eta C_\gamma \|u\|_{W^{2,2}(\Omega)} + \tag{15}$$
$$\mu\Big(C(\underline{u}, g) + C(R_\sigma N(L,\theta_u), g)\Big); \tag{16}$$

and thus thanks to Theorem 2

$$\frac{1}{4}\|\underline{u} - u\|_\eta^2 \leq C\epsilon^2 + 8\eta C_\gamma \|u\|_{W^{2,2}(\Omega)} + \mu\Big(C(\underline{u}, g) + C(R_\sigma N(L,\theta_u), g)\Big) \tag{17}$$

Proof. Lemma 5 yields the existence of $\underline{u} = R_\sigma N(L, \underline{\theta})$ solution to Problem (10). Since $R_\sigma N(L, \theta_u)$ is not a minima of J_η^G we have:

$$J_\eta^G(\underline{u}) \leq J_\eta^G(R_\sigma N(L, \theta_u)).$$

Lemma 3 yields estimates between J_η and J_η^G which combined with previous inequality give:

$$J_\eta(\underline{u}) - \mu C(\underline{u}, g) \leq J_\eta(R_\sigma N(L, \theta_u)) + \mu C(R_\sigma N(L, \theta_u), g). \tag{18}$$

Denote by $(\cdot, \cdot)_A$ the inner product induced by the differential operator A. We have $u \in W^{2,2}(\Omega)$ with $\gamma(u) = 0$, thus integrating by parts leads to:

$$\int_\Omega g\underline{u}\, dx = \int_\Omega Au\underline{u}\, dx = (u, \underline{u})_A + \int_{\partial\Omega} \sum_{i,j=1}^d a_{ij} D_j u\, n_i \underline{u}\, ds$$

where n is the outward normal to $\partial\Omega$. By using the continuity of the trace operator γ for the normal derivative: $\partial_{n_A} = \sum_{i,j=1}^d a_{ij} D_j \cdot n_i$ (see [17] for example) we get $\|\gamma(\partial_{n_A} u)\|_{L^2(\partial\Omega)} \leq C_{\gamma,A} \|u\|_{W^{2,2}(\Omega)}$. The young inequality, provides:

$$-4\eta C_{\gamma,A} \|u\|_{W^{2,2}(\Omega)} - \frac{1}{4\eta} \int_{\partial\Omega} (u - \underline{u})^2\, dx \leq \int_{\partial\Omega} |\partial_{n_A} u\underline{u}|\, dx \tag{19}$$

We deduce the following bound from below for J_η:

$$\frac{1}{4} \|\underline{u} - u\|_\eta^2 - \frac{1}{2}(u, u)_A - 4\eta C_{\gamma,A} \|u\|_{W^{2,2}(\Omega)} \leq J_\eta(\underline{u}) \tag{20}$$

Arguing in the same way and since $\theta_u \in B(0, R)$, we get the following bound from above for $J_\eta(R_\sigma N(L, \theta_u))$

$$J_\eta(R_\sigma N(L, \theta_u)) \leq \frac{3}{4} \|R_\sigma N(L, \theta_u) - u\|_\eta^2 - \frac{1}{2}(u, u)_A + 4\eta C_{\gamma,A} \|u\|_{W^{2,2}(\Omega)}$$

and we get the announced error estimate.

Remark 7. There does not exist a unique minimum for function J_η^G. While local minima are numerous, they are relatively easy to find, and they are all more or less equivalent. This peculiar property is analysed for the loss function of a typical multilayer net with ReLU activation function in [20] with the use of random matrix theory applied to the analysis of critical points in high degree polynomials on the sphere.

In practice R can be chosen sufficiently large for satisfying $\theta_u \in B(0, R)$.

Remark 8. The penalization function is positive, thus does not play any role in the bound from below in the error estimate. Since R is sufficiently large for θu be in the ball $B(0, R)$, the penalization function does not appear in the bound from above in the error estimate.

4. The case of some second order elliptic PDEs in d-dimension

Consider an operator A in divergence form,

$$Av = \sum_{i,j=1}^d -D_i(a_{ij} D_j v) + \sum_{j=1}^d b_j D_j v + a_0 v,$$

with $a_{ij} \in C^{1,1}(\overline{\Omega})$; $b \in C^1(\overline{\Omega})$; $\mathbb{R}^d$; $0 < \underline{a} \leq a_0 \in C^0(\overline{\Omega})$ with $0 < a_0 - \frac{1}{2} div(b)$ (to keep technical difficulties as simple as possible) and where there exists $0 < \alpha$ such that

$$\alpha \leq \sum_{i,j=1}^d a_{ij}(x) \xi_i \xi_j$$

for all $x \in \overline{\Omega}$ and $\xi \in \mathbb{R}^d$. Problem (21) reads: for $g \in C^1(\overline{\Omega})$ be given, find $u \in W^{1,2}(\Omega)$ verifying:

$$\begin{aligned} &Au = g \ \text{ in } \Omega; \\ &u = 0 \text{ on } \partial\Omega. \end{aligned} \tag{21}$$

Let us denote by $W_0^{1,2}(\Omega) = W^{1,2}(\Omega) \cap Ker(\gamma)$, Problem PE has one solution in $W^{2,2}(\Omega) \cap W_0^{1,2}(\Omega)$ (see [17] Chapter 2). A variational formulation for problem PE reads:

$$\int_\Omega \sum_{i,j=1}^d a_{ij} D_i u D_j v + v \sum_{j=1}^d b_j D_j u + a_0 uv - g v\, dx = 0 \, \forall v \in W_0^{1,2}(\Omega) \tag{22}$$

The variational formulation (22) can be expressed in form of the following saddle point problem with $J : W_0^{1,2}(\Omega) \times W_0^{1,2}(\Omega) \to \mathbb{R}$ defined by:

$$J(w, v) = \int_\Omega \sum_{i,j=1}^{d} a_{ij} D_i w D_j v + v \sum_{j=1}^{d} b_j D_j w + a_0 w v - g v \, dx, \tag{23}$$

$$\inf_{w} \sup_{\|v\|_{W^{1,2}}=1} J(w, v) \tag{24}$$

The function J is bilinear, and twice continuously differentiable. The optimality conditions for the saddle point $(\underline{u}, \overline{v})$ verifying:

$$\forall \varphi, w \in W_0^{1,2}(\Omega), J(\underline{u}, w) \le J(\underline{u}, \overline{v}) \le J(\varphi, \overline{v})$$

read: there exists $\lambda \in \mathbb{R}$ be such that $\forall w \in W_0^{1,2}(\Omega)$:

$$\begin{array}{ll} D_1 J(\underline{u}, \overline{v}) w & = \int_\Omega \sum_{i,j=1}^{d} a_{ij} D_i w D_j \overline{v} + \overline{v} \sum_{j=1}^{d} b_j D_j w + a_0 w \overline{v} \, dx \qquad = 0; \\ D_2 J(\underline{u}, \overline{v}) w & = \int_\Omega \sum_{i,j=1}^{d} a_{ij} D_i \underline{u} D_j w + w \sum_{j=1}^{d} b_j D_j \underline{u} + a_0 \underline{u} w - g w \, dx \\ & = \lambda (\overline{v}, w)_{W^{1,2}(\Omega)} \end{array} \tag{25}$$

The expression $(\cdot, \cdot)_{W^{1,2}(\Omega)}$ denotes the inner product of $W^{1,2}$, and the right hand side of the second equation is due to the constraint $(v, v)^{\frac{1}{2}}_{W^{1,2}} = 1$. From the first equation of (25) associated to the $W^{1,2}$-coercivity of the bilinear form defined by $\overline{v}, w \mapsto D_1 J(\overline{u}, \overline{v}) w$ we deduce that $\overline{v} = 0$ and thus the second equation of (25) reduces to the variational formulation (22).

In the same way that have been done in Section 3, for defining the approximated problem we need to consider two NNs: $N(L, \theta)$ for functions u and $N(\tilde{L}, \tilde{\theta})$ for functions v. In what follows, for simplifying the notations $R_\sigma N(L, \theta)$ and $R_\sigma N(\tilde{L}, \tilde{\theta})$ will be denoted by $N(\theta)$, and by $N(\tilde{\theta})$. Consider the bilinear form $(\cdot, \cdot)^G_\eta \, \mathcal{C}(L, \theta) \times \mathcal{C}(\tilde{L}, \tilde{\theta}) \to \mathbb{R}$ and the linear forms $(g, \cdot)^G_\eta : \mathcal{C}(L, \theta) \to \mathbb{R}$ defined by:

$$\begin{array}{l} (\varphi, v)^G_\eta = \sum_{i=1}^{M} W_i \Big(\sum_{k,l=1}^{d} a_{kl} D_l \varphi D_k v(x_i) + v \sum_{j=1}^{d} b_j D_j \varphi(x_i) + a_0 \varphi v(x_i) \Big) + \\ \frac{1}{\eta} \sum_{j=1}^{m} w_j \varphi v(y_j); \\ (g, v)^G = \sum_{i=1}^{M} W_i \Big(\sum_{k,l=1}^{d} g v(x_i) \Big). \end{array} \tag{26}$$

Introduce the two following bilinear forms:

$$\begin{array}{l} J_\eta(\varphi, v) = \int_\Omega \sum_{i,j=1}^{d} a_{ij} D_i \varphi D_j v + v \sum_{j=1}^{d} b_j D_j \varphi + a_0 \varphi v - g v \, dx + \\ \frac{1}{\eta} \int_{\partial\Omega} \varphi v \, ds + \frac{1}{4\eta} \left(\theta^2 - B^2 \right)^{+2} \\ J^G_\eta(\varphi, v) = (\varphi, v)^G_\eta - (g, v)^G + \frac{1}{4\eta} \left(\theta^2 - B^2 \right)^{+2}. \end{array} \tag{27}$$

Remark that whatever $0 < M$ is, we have (since J_η is linear with respect to v):

$$\begin{array}{l} \underset{N(\theta) \in \mathcal{C}(L,\theta) N(\tilde{\theta}) \neq 0 \, \|N(\tilde{\theta})\|_{W^{1,2}} \le M}{\operatorname{Argmin}} \quad \underset{}{\operatorname{Argmax}} \quad \frac{J_\eta(N(\theta), N(\tilde{\theta}))}{\|N(\tilde{\theta})\|_{W^{1,2}}} = \\ \underset{N(\theta) \in \mathcal{C}(L,\theta) N(\tilde{\theta}) \neq 0 \, \|N(\tilde{\theta})\|_{W^{1,2}} = 1}{\operatorname{Argmin}} \quad \underset{}{\operatorname{Argmax}} \quad J_\eta(N(\theta), N(\tilde{\theta})) \end{array} \tag{28}$$

and thus thanks to Lemma 1

$$\begin{array}{l} \underset{N(\theta) \in \mathcal{C}(L,\theta) N(\tilde{\theta}) \neq 0 \, \|N(\tilde{\theta})\|_{W^{1,2}} = 1}{\operatorname{Argmin}} \quad \underset{}{\operatorname{Argmax}} \quad J_\eta(N(\theta), N(\tilde{\theta})) \\ \underset{N(\theta) \in \mathcal{C}(L,\theta) \tilde{\theta} \in B(0,\tilde{B}), \|N(\tilde{\theta})\|_{W^{1,2}} = 1}{\operatorname{Argmin}} \quad \underset{}{\operatorname{Argmax}} \quad J_\eta(N(\theta), N(\tilde{\theta})) \end{array} \tag{29}$$

The NN approximated problem is defined by:

$$\left\{ \underline{u}, \overline{v} \right\} = \underset{N(\theta) \in \mathcal{C}(L,\theta) \tilde{\theta} \in B(0,\tilde{B}), \|N(\tilde{\theta})\|_{W^{1,2}} = 1}{\operatorname{Argmin}} \quad \underset{}{\operatorname{Argmax}} \quad J^G_\eta(N(\theta), N(\tilde{\theta})). \tag{30}$$

The function J_η^G is twice continuously differentiable since it is a bilinear function added to a C^2 functions. Simple calculations provide for all φ, v:

$$\begin{aligned}
&D_\theta J_\eta^G(N(\theta), v) = (\tfrac{\partial N(\theta)}{\partial \theta}, v)_\eta^G + \tfrac{1}{\eta}(\theta^2 - B^2)^+\theta;\\
&D_{\tilde\theta} J_\eta^G(\varphi, N(\tilde\theta)) = (\varphi, \tfrac{\partial N(\tilde\theta)}{\partial \tilde\theta})_\eta^G - (g, \tfrac{\partial N(\tilde\theta)}{\partial \tilde\theta})^G;\\
&D^2_{\theta 2} J_\eta^G(N(\theta), v) = (\tfrac{\partial^2 N(\theta)}{\partial \theta^2}, v)_\eta^G + \tfrac{1}{\eta}\left((\theta^2 - B^2)^+ + 2\theta^2 sgn^+(\theta^2 - B^2)\right);\\
&D^2_{\tilde\theta 2} J_\eta^G(\varphi, N(\tilde\theta)) = (\varphi, \tfrac{\partial^2 N(\tilde\theta)}{\partial \tilde\theta^2})_\eta^G - (g, \tfrac{\partial^2 N(\tilde\theta)}{\partial \tilde\theta^2})^G;\\
&D^2_{\theta\tilde\theta} J_\eta^G(N(\theta), N(\tilde\theta)) = (\tfrac{\partial N(\theta)}{\partial \theta}, \tfrac{\partial N(\tilde\theta)}{\partial \tilde\theta})_\eta^G.
\end{aligned} \tag{31}$$

Consider the two following intermediate optimization problems. For $u, \in \mathcal{C}(L, \theta)$, $v \in \mathcal{C}(\tilde L, \tilde\theta)$ given find $\overline{v}(u), \underline{u}(v) \in \mathcal{C}(\tilde L, \tilde\theta) \times \mathcal{C}(L, \theta)$ verifying:

$$\overline{v}(u) = \underset{\tilde\theta \in B(0,\tilde B) \| N(\tilde\theta)\|_{W^{1,2}} = 1}{\text{Argmax}} \quad J_\eta^G(u, N(\tilde\theta)) \tag{32}$$

$$\underline{u}(v) = \underset{N(\theta)\in\mathcal{C}(L,\theta)}{\text{Argmin}}\ J_\eta^G(N(\theta), v) \tag{33}$$

Lemma 9. *Assume $\Omega = (0,1)^d$ and σ to be hyperbolic tangent or sigmoid function, and let $u \in C(L, \theta)$ be given, then there exists $\overline{v}(u) \in \mathcal{C}(\tilde L, \tilde\theta)$ solution to Problem (32), calculated with a gradient ascent algorithm.*

Proof. See Annex A

Lemma 10. *Assume $\Omega = (0,1)^d$ and σ to be hyperbolic tangent or sigmoid function, and let $v \in \mathcal{C}(\tilde L, \tilde\theta)$ be given, then there exists $\underline{u}(v) \in \mathcal{C}(L, \theta)$ solution to Problem (33) calculated with a gradient descent algorithm.*

Proof. See Annex B

Now, we give the algorithm for solving Problem (30). With Lemma 9 and Lemma 10 we are able to compute a double index sequence $\{N(\theta_{n_p}), N(\tilde\theta_{n_q})\}_{p,q\in\mathbb{N}}$ which for p fixed converges when q goes to infinity towards a solution to Problem (32), and which for q fixed converges when p goes to infinity to a solution to Problem (33). By using the diagonal extraction principle we have a converging sequence $\{N(\theta_{n_p}), N(\tilde\theta_{n_p})\}_{p\in\mathbb{N}}$ towards a solution to Problem (30). The proposed ascent–descent algorithm reads:

- $0 < h$ and $0 < \tilde h$, be given, define $n = 0$,
  - □ $n = n + 1$
  - – Choose $\tilde\theta_0 \in B(0, \tilde B)$, $N(\tilde\theta_0) = \frac{N(\tilde\theta_0)}{\|N(\tilde\theta_0)\|_{W^{1,2}}}$
  - – For $p = 0$, to n
    - ∗ $\tilde\theta_{p+1} = \tilde\theta_p + hD_{\tilde\theta} J_\eta^G(N(\theta_n), N(\tilde\theta_p))$
    - ∗ $\tilde\theta_{p+1} =$ its projection into the ball $B(0, \tilde B)$
    - ∗ $N(\tilde\theta_{p+1}) = \frac{N(\tilde\theta_{p+1})}{\|N(\tilde\theta_{p+1})\|_{W^{1,2}}}$
    - ∗ compute $D_{\tilde\theta} J_\eta^G(N(\tilde\theta_n), N(\tilde\theta_{p+1}))$
  - – end of loop for p
  - – choose $\theta_0 = 0$
  - – For $q = 0$, to n
  - – $\theta_{q+1} = \theta_q - hD_\theta J_\eta^G(N(\theta_q), N(\tilde\theta_n))$
  - – compute $D_\theta J_\eta^G(N(\theta_{q+1}), N(\tilde\theta_n))$
  - – end of loop for q
  - – If $|\tilde\theta_{n+1} - \tilde\theta_n|$ and $|\theta_{n+1} - \theta_n|$ are not small enough go to □

Define the norm $|\|\cdot\||_\eta^2 = \|\cdot\|_{W^{1,2}}^2 + \frac{1}{\eta}\|\cdot\|_{L^2(\partial\Omega)}^2$ which is equivalent to the $W^{1,2}$ norm (see [7]). Finally we have the following existence and error estimate results.

Theorem 11. *Assume $\Omega = (0,1)^d$ and σ to be the hyperbolic tangent or the sigmoid function, let $0 < \eta$ and $0 < \epsilon < \frac{1}{2}$ be given, denote by $u \in W^{2,2}(\Omega)$ the solution to Problem (22) and by $R_\sigma N(L, \theta_u) \in \mathcal{C}(L, \theta)$ the approximation of u into $\mathcal{C}(L, \theta)$ given by Theorem 2. Then there exists $(\underline{u}, \overline{v}) \in \mathcal{C}(L, \theta) \times \mathcal{C}(\tilde L, \tilde\theta)$ solution to Problem (30). Moreover, assuming $\theta_u \in B(0, B)$ and*

$R_\sigma N(L,\theta_u)$ is not a solution to Problem (33) with $v=\overline{v}$, then the following estimate is valid

$$\|\|\underline{u}-u\|\|_\eta \le C(A,\gamma)\|\|R_\sigma N(L,\theta_u)-u\|\|_\eta + \tag{34}$$

$$C(A,\gamma,\alpha,\eta)\|\|R_\sigma(N(\tilde{L},\tilde{\theta}_{\underline{u}-u}))-(\underline{u}-u)\|\|_\eta + \tag{35}$$

$$\mu\Big(C(\underline{u},g)+C(R_\sigma N(L,\theta_u),g)\Big); \tag{36}$$

and thus thanks to Theorem 2

$$\|\|\underline{u}-u\|\|_\eta \le C(A,\gamma)\epsilon + C(A,\gamma,\alpha,\eta)\epsilon_1 + \mu\Big(C(\underline{u},\overline{v},g)+C(R_\sigma N(L,\theta_u),g)\Big) \tag{37}$$

Proof. Lemmas 9, 10 yield the existence of $(\underline{u},\overline{v})$ solution to Problem (30). Since $R_\sigma N(L,\theta_u)$ is not a minima of $J^G_\eta(\cdot,\overline{v})$ we have:

$$J^G_\eta(\underline{u},\overline{v}) \le J^G_\eta(R_\sigma N(L,\theta_u),\overline{v}).$$

Lemma 3 yields estimates between J_η and J^G_η which combined with previous inequality and since $\theta_u \in B(0,B)$ give:

$$J_\eta(\underline{u},\overline{v})-\mu C(\underline{u},\overline{v},g) \le J_\eta(R_\sigma N(L,\theta_u),\overline{v})+\mu C(R_\sigma N(L,\theta_u),\overline{v},g). \tag{38}$$

we have:

$$J_\eta(R_\sigma N(L,\theta_u),\overline{v}) = \int_\Omega \sum_{i,j=1}^d a_{ij}D_i(R_\sigma N(L,\theta_u)-u)D_j\overline{v} +$$

$$\overline{v}\sum_{j=1}^d b_jD_j(R_\sigma N(L,\theta_u)-u)\,dx +$$

$$\int_\Omega a_0(R_\sigma N(L,\theta_u)-u)\overline{v}\,dx + \frac{1}{\eta}\int_{\partial\Omega}(R_\sigma N(L,\theta_u)-u)\overline{v}\,ds$$

which is bounded:

$$|J_\eta(R_\sigma N(L,\theta_u),\overline{v})| \le C(A,\gamma,\eta)\|\|R_\sigma N(L,\theta_u)-u|\|_\eta.$$

Whatever $w \in \mathcal{C}(\tilde{L},\tilde{\theta})$ is with a $W^{1,2}$ norm one, for getting a bounded from below, start from the inequality:

$$\int_\Omega \sum_{i,j=1}^d a_{ij}D_i(\underline{u}-u)D_jw + w\sum_{j=1}^d b_jD_j(\underline{u}-u)\,dx +$$

$$\int_\Omega a_0(\underline{u}-u)w\,dx + \frac{1}{\eta}\int_{\partial\Omega}(\underline{u}-u)w\,ds \le J_\eta(\underline{u},\overline{v}).$$

Then choose

$$w = \frac{R_\sigma(N(\tilde{L},\tilde{\theta}_{\underline{u}-u}))}{\|R_\sigma(N(\tilde{L},\tilde{\theta}_{\underline{u}-u}))\|_{W^{1,2}}} =$$

$$\frac{1}{\|R_\sigma(N(\tilde{L},\tilde{\theta}_{\underline{u}-u}))\|_{W^{1,2}}}\Big(\underline{u}-u+R_\sigma(N(\tilde{L},\tilde{\theta}_{\underline{u}-u}))-(\underline{u}-u)\Big),$$

with thanks to Theorem 2

$$\|R_\sigma(N(\tilde{L},\tilde{\theta}_{\underline{u}-u}))-(\underline{u}-u)\|_{W^{1,2}} \le \epsilon_1 \text{ and } \|R_\sigma(N(\tilde{L},\tilde{\theta}_{\underline{u}-u}))\|_{W^{1,2}} \le \|(\underline{u}-u)\|_{W^{1,2}}+\epsilon_1$$

By using the coercivity of the bilinear form, we have:

$$\frac{1}{\|R_\sigma(N(\tilde{L},\tilde{\theta}_{\underline{u}-u}))\|_{W1,2}}$$

$$\Big[\min(\alpha,1)\|\|\underline{u}-u\|\|^2_\eta - C(A,\gamma)\|\|\underline{u}-u\|\|_\eta\|\|R_\sigma(N(\tilde{L},\tilde{\theta}_{\underline{u}-u}))-(\underline{u}-u)\|\|_\eta\Big]$$

and we get the following bound from below:

$$\frac{1}{2\|\underline{u}-u\|_{W^{1,2}}}$$

$$\Big[\min(\alpha,1)\|\|\underline{u}-u\|\|^2_\eta - C(A,\gamma)\|\|\underline{u}-u\|\|_\eta\|\|R_\sigma(N(\tilde{L},\tilde{\theta}_{\underline{u}-u}))-(\underline{u}-u)\|\|_\eta\Big].$$

Since the norms $\|\cdot\|_{W^{1,2}}$ and $|||\cdot|||_\eta$ are equivalent, there exists a constant $C(\eta)$ such that for all v, $0 < C(\eta) \leq \frac{|||v|||_\eta}{\|v\|_{W^{1,2}}}$ and we get the following bound from below for $J_\eta(\underline{u}, \overline{v})$:

$$\frac{C(\eta)}{2}\left[\min(\alpha, 1)|||\underline{u} - u|||_\eta - C(A, \gamma)|||R_\sigma(N(\tilde{L}, \tilde{\theta}_{\underline{u}-u})) - (\underline{u} - u)|||_\eta\right] \leq J_\eta(\underline{u}, \overline{v}).$$

Gathering this bound from below with the bound from above gives the announced estimate.

Let us end this section with some comments about some existing ways for enforcing the Dirichlet boundary conditions [21]. One can use an approximate distance function to exactly impose the boundary conditions by modifying the last layer of the NN, or by adding this distance function to the variational formulation. The Nitsche method [22] where the boundary conditions are variationally imposed, and finally the penalization method presented in this study. The gradient strategy and the existence and error estimates results presented in this paper can be extended to the Nitsche method.

Data availability

No data was used for the research described in the article.

Annex A

Proof of Lemma 9. A gradient ascent algorithm is applied, which reads:

- Choose $\tilde{\theta}_0 \in B(0, \tilde{B})$, $n = 0$, $N(\tilde{\theta}_0) = \frac{N(\tilde{\theta}_0)}{\|N(\tilde{\theta}_0)\|_{W^{1,2}}}$
- * $\tilde{\theta}_{n+1} = \tilde{\theta}_n + hD_{\tilde{\theta}}J_\eta^G(u, N(\tilde{\theta}_n))$
- $\tilde{\theta}_{n+1} =$ its projection into the ball $B(0, \tilde{B})$
- $N(\tilde{\theta}_{n+1}) = \frac{N(\tilde{\theta}_{n+1})}{\|N(\tilde{\theta}_{n+1})\|_{W^{1,2}}}$
- While $|\tilde{\theta}_{n+1} - \tilde{\theta}_n|$ is not small enough do
  - compute $D_{\tilde{\theta}}J_\eta^G(u, N(\tilde{\theta}_{n+1}))$
  - n=n+1 and go to *

Since $\tilde{\theta}_n \in B(0, \tilde{B})$ for all n, Lemma 1 implies $N(\tilde{\theta}_n)$ is bounded in $W^{1,\infty}$ norm. We deduce that $J_\eta^G(u, N(\tilde{\theta}_n))$ is bounded.

The sequence $J_\eta^G(u, N(\tilde{\theta}_n))$ is increasing. We have:

$$J_\eta^G(u, N(\tilde{\theta}_{n+1})) - J_\eta^G(u, N(\tilde{\theta}_n)) = hD_{\tilde{\theta}}J_\eta^G(u, N(\tilde{\theta}_n))^2 \tag{39}$$

$$+\frac{h^2}{2}D_{\tilde{\theta}^2}^2 J_\eta^G(u, N(\tilde{\theta}_{n+1}) + \xi(N(\tilde{\theta}_{n+1}) - N(\tilde{\theta}_n)))D_{\tilde{\theta}}J_\eta^G(u, N(\tilde{\theta}_n))^2 \tag{40}$$

Lemma 1 implies $D_{\tilde{\theta}^2}^2 J_\eta^G(u, N(\tilde{\theta}_{n+1}) + \xi(N(\tilde{\theta}_{n+1}) - N(\tilde{\theta}_n)))$ is bounded, thus we can choose h sufficiently small such that $h|D_{\tilde{\theta}^2}^2 J_\eta^G(u, N(\tilde{\theta}_{n+1}) + \xi(N(\tilde{\theta}_{n+1}) - N(\tilde{\theta}_n)))| < 1$ and we get the following inequality:

$$\frac{h}{2}D_{\tilde{\theta}}J_\eta^G(u, N(\tilde{\theta}_n))^2 \leq J_\eta^G(u, N(\tilde{\theta}_{n+1})) - J_\eta^G(u, N(\tilde{\theta}_n)). \tag{41}$$

There exists a subsequence n_p be such as $\{J_\eta^G(u, N(\tilde{\theta}_{n_p}))\}_{p\in\mathbb{N}}$ converges towards a. We get the convergence of the following series:

$$\sum_{p\in\mathbb{N}} D_{\tilde{\theta}}J_\eta^G(u, N(\tilde{\theta}_{n_p}))^2.$$

$J_\eta^G(u, N(\tilde{\theta}_{n_p}))$ is bounded there exists a constant C such that $0 \leq J_\eta^G(u, N(\tilde{\theta}_{n_p})) + C$. Arguing in the same way as for the proof of Lemma 5 and by using the $\sqrt{\cdot}$ function we deduce the convergence of the series

$$\sum_{p\in\mathbb{N}} D_{\tilde{\theta}}J_\eta^G(u, N(\tilde{\theta}_{n_p})),$$

which implies the convergence of the sequence $\{\tilde{\theta}_{n_p}\}_{p\in\mathbb{N}}$ and since the realization map R_σ is continuous, the convergence of $\{R_\sigma(N(\tilde{L}, \tilde{\theta}_{n_p}))\}_{p\in\mathbb{N}}$ towards a maximum of function J_η^G

Annex B

Proof of Lemma 10. Thanks to Lemma 1, we have $\frac{\|N(\theta)\|_{W^{1,\infty}}}{|\theta|} \leq C(\sigma, Z)$, we deduce that the function J_η^G is coercitive at infinity ($\lim_{|\theta|\to+\infty} \frac{J_\eta^G(\theta)}{|\theta|} = +\infty$).

There exists $0 < B_2$ such that minimizing sequences verify $\theta_n \in B(0, B_2)$. Otherwise, if θ_n were not bounded, it would not minimize J_η^G.

A gradient descent algorithm is applied, which reads:

- Choose θ_0, $n = 0$,
- * $\theta_{n+1} = \theta_n - hD_\theta J_\eta^G(N(\theta_n), v)$
- While $|\theta_{n+1} - \theta_n|$ is not small enough do
  - compute $D_\theta J_\eta^G(N(\theta_{n+1}), v)$
  - n = n+1 and go to *

Let us prove the convergence of the algorithm for a subsequence. The parameter h is chosen such that for all $\theta_n, \theta_{n+1} \in B(0, B_2)$ the following inequality is verified:

$$h|D^2_{\theta^2} J_\eta^G(N(\theta_n) + \xi(N(\theta_{n+1}) - N(\theta_n)))| < 1.$$

which is possible thanks to Lemma 1.

We have:

$$\begin{aligned} &J_\eta^G(N(\theta_{n+1})) - J_\eta^G(N(\theta_n)) = -h\left(D_\theta J_\eta^G(N(\theta_n))\right)^2 + \frac{h^2}{2} \\ &D^2_{\theta^2} J_\eta^G(N(\theta_n) + \xi(N(\theta_{n+1}) - N(\theta_n)))\left(D_\theta J_\eta(N(\theta_n))\right)^2 \\ &\leq -\frac{h}{2}\left(D_\theta J_\eta^G(N(\theta_n))\right)^2. \end{aligned}$$

The sequence $J_\eta^G(N(\theta_n), v)$ is decreasing, and since $J_\eta^G(N(\theta_n), v) \leq J_\eta^G(N(\theta_0), v) = 0$. We deduce the following bound for θ_n:

$$|\theta_n| \leq B + 4\eta C(A, C(\sigma, Z), g)\|v\|_{W^{1,\infty}}.$$

As a consequence of Lemma 1, the sequence $\{J_\eta^G(N(\theta_n), v)\}_{n\in\mathbb{N}}$ is bounded, thus there exists a subsequence and a real a such that: $\lim_{p\to+\infty} J_\eta^G(N(\theta_{n_p}), v) = a$. We get the convergence of the following series:

$$\sum_{p\in\mathbb{N}} D_\theta J_\eta^G(N(\theta_n), v)^2.$$

$J_\eta^G(N(\theta_{n_p}), v)$ is bounded there exists a constant C such that $0 \leq J_\eta^G(N(\theta_{n_p}), v) + C$. Arguing in the same way as for the proof of Lemma 5 and by using the $\sqrt{\cdot}$ function we deduce the convergence of the series

$$\sum_{p\in\mathbb{N}} D_\theta J_\eta^G(N(\theta_{n_p}), v),$$

which implies the convergence of the sequence $\{\theta_{n_p}\}_{p\in\mathbb{N}}$ and since the realization map R_σ is continuous, the convergence of $\{R_\sigma(N(L, \theta_{n_p}))\}_{p\in\mathbb{N}}$ towards a minimum of function J_η^G

References

[1] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016, available from http://www.deeplearningbook.org.
[2] Namig J. Guliyev, Vugar E. Ismailov, On the approximation by single hidden layer feedforward neural networks with fixed weights, Neural Netw. 98 (2018).
[3] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.
[4] J. He, L. Li, J. Xu, C. Zheng, Relu deep neural networks and linear finite elements, 2018, arXiv preprint arXiv:1807.03973.
[5] E. Weinan, Yu. Bing, The deep ritz method: A deep learning-based numerical algorithm for solving variational problems, 2017, arXiv:1710.00211v1 [Cs.LG].
[6] R. Glowinski, Lectures on Numerical Methods for Non-Linear Variational Problems, in: Scientific computation series, Springer-Verlag, 1984.
[7] J. Pousin, Least squares formulations for some elliptic second order problems, feedforward neural network solutions and convergence results., J. Comput. Math. Data Sci. 2 (2022).
[8] A. Biswas, J. Tian, S. Ulusoy, Error estimates for deep learning methods in fluid dynamics, Numer. Math. 151 (2022).
[9] S. Mishra, R. Molinaro, Estimates on the generalization error of physics-informed neural networks for approximating PDEs, IMA J. Numer. Anal. 43 (2023).
[10] Yaohua Zang, Gang Bao, Xiaojing Ye, Weak adversarial networks for high-dimensional partial differential equations, J. Comput. Phys. 411 (2020).
[11] Gang Bao, Xiaojing Ye, Yaohua Zang, Numerical solution of inverse problems by weak adversarial networks, Inverse Problems 36 (2020).
[12] Carlos Uriarte, David Pardo, Ignacio Muga, Judit Muñoz-Matute, A deep double ritz method (DRM) for solving partial differential equations using neural networks, Comput. Methods Appl. Mech. Engrg. 405 (2023) 15.

[13] P. Pertersen, M. Raslan, F. Voigtaender, Topological properties of the set of functions generated by neural networks of fixed size, Found. Comput. Math. 21 (2021) 375–444.
[14] J. Han, A. Jentzen, E. Weiman, Solving high-dimensional partial differential equations using deep learning, Proc. Natl. Acad. Sci. 115 (34) (2018) 8505–8510.
[15] J. Sirignano, K. Spiliopoulos, DGM: A deep learning algorithm for solving partial differential equations, J. Comput. Phys. 375 (2018) 1339–1364.
[16] K. Hornik, Approximation capabilities of multilayer feedforward networks, Neural Netw. 4 (2) (1991) 251–257.
[17] P. Grisvard, Elliptic Problems in Nonsmooth Domains, SIAM, 2011.
[18] L. awrence, C. Evans, Partial Differential Equations, in: Graduate Studies in Mathematics, vol 19, AMS, 1998.
[19] Ingo Guhring, Gitta Kutyniok, Philipp Petersen, Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms, Anal. Appl. 18 (5) (2020).
[20] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gerard Ben Arous, Yann LeCun, The loss surfaces of multilayer networks, in: Proceedings of the 18th International Conference on Artificial Intelligence and Statistics, Vol. 38, AISTATS, W & CP, San Diego, CA, USA. JMLR, 2015.
[21] S. Berrone, C. Canuto, M. Pintore, N. Sukumar, Enforcing Dirichlet boundary conditions in physics-informed neural networks and variational physics-informed neural networks, 2022, arXiv:2210.14795v.
[22] J.A. Nitsche, Uber ein Variationsprinzip zur Losung Dirichlet-Problemen bei Verwendung von Teilraumen, die keinen Randbedingungen unteworfen sind, Abh. Math. Semin. Univ. Hambg. 36 (1971).