Weierstraß-Institut
für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

A neural network approach to learning solutions of a class of elliptic variational inequalities

Amal Alphonse[1], Michael Hintermüller[1,2], Alexander Kister[3], Chin Hang Lun[4],

Clemens Sirotenko[1]

submitted: December 16, 2024

[1] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
E-Mail: amal.alphonse@wias-berlin.de
michael.hintermueller@wias-berlin.de
clemens.sirotenko@wias-berlin.de

[2] Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin
Germany
E-Mail: hint@math.hu-berlin.de

[3] Federal Institute for Materials Research and Testing (BAM)
Unter den Eichen 87
12205 Berlin
Germany
E-Mail: alexander.kister@bam.de

[4] Tripadvisor, Inc
7 Soho Square
London W1D 3QB
United Kingdom
E-Mail: clun@tripadvisor.com

No. 3152
Berlin 2024

2020 *Mathematics Subject Classification.* 49J40, 90C47, 68T07, 68Q32, 49M20, 49M29, 65K15.

Key words and phrases. Variational inequalities, neural networks, weak adversarial networks, infsup problems, nonsmooth optimisation.

Code is available at https://github.com/amal-alphonse/NNVI.

We thank Guozhi Dong (Central South University, Changsha) and Pavel Dvurechensky (WIAS, Berlin) for helpful discussions. MH was partially supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the DFG SPP 1962 Priority Programme *Non-smooth and Complementarity-based Distributed Parameter Systems: Simulation and Hierarchical Optimization* within project 10, and partially by the DFG under Germany's Excellence Strategy – The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689).

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

A neural network approach to learning solutions of a class of elliptic variational inequalities

Amal Alphonse, Michael Hintermüller, Alexander Kister, Chin Hang Lun, Clemens Sirotenko

Abstract

We develop a weak adversarial approach to solving obstacle problems using neural networks. By employing (generalised) regularised gap functions and their properties we rewrite the obstacle problem (which is an elliptic variational inequality) as a minmax problem, providing a natural formulation amenable to learning. Our approach, in contrast to much of the literature, does not require the elliptic operator to be symmetric. We provide an error analysis for suitable discretisations of the continuous problem, estimating in particular the approximation and statistical errors. Parametrising the solution and test function as neural networks, we apply a modified gradient descent ascent algorithm to treat the problem and conclude the paper with various examples and experiments. Our solution algorithm is in particular able to easily handle obstacle problems that feature biactivity (or lack of strict complementarity), a situation that poses difficulty for traditional numerical methods.

1 Introduction

In this paper, we use neural networks to find solutions of variational inequalities (VIs) of the following type:

$$\text{find } u \in K : \langle Au - f, u - v\rangle_{V^*,V} \leq 0 \quad \text{for all } v \in K, \tag{1}$$

where $V := H^1(\Omega)$ is the usual Sobolev space on a bounded Lipschitz domain $\Omega \subset \mathbb{R}^n$, $\langle \cdot, \cdot \rangle_{V^*,V}$ denotes the duality pairing between V and its topological dual V^*, the constraint set K is defined as

$$K := \left\{ u \in H^1(\Omega) \mid u \geq \psi \text{ in } \Omega, u = h \text{ on } \partial\Omega \right\}, \tag{2}$$

$h \in H^{1/2}(\partial\Omega)$ is given boundary data, $\psi \in H^1(\Omega)$ is a given obstacle that satisfies $\psi \leq h$ on $\partial\Omega$, and $f \in L^2(\Omega)$ is a given source term. Concerning the involved Sobolev and Lebesgue spaces we refer, e.g., to [1]. The operator $A\colon K \subset V \to V^*$ appearing in (1) is assumed to be Lipschitz and coercive, i.e., there exist constants $C_a, C_b > 0$ such that

$$\|Au - Av\|_{V^*} \leq C_b \|u - v\|_V \quad \forall u, v \in K, \tag{3a}$$

$$\langle Au - Av, u - v\rangle_{V^*,V} \geq C_a \|u - v\|_V^2 \quad \forall u, v \in K, \tag{3b}$$

and for simplicity, we focus our attention on linear differential operators of the form

$$\langle Au, u - v\rangle_{V^*,V} = \int_\Omega \nabla u \cdot \nabla(u - v) + \sum_{i=1}^n b_i \partial_{x_i} u(u - v) + ku(u - v), \tag{4}$$

where $k, b_i \geq 0$, $i = 1, \ldots, n$, are constants (which of course have to be such that (3) is satisfied), $\partial_{x_i} u$ is the weak partial derivative of u with respect to the i^{th} coordinate and ∇u is the weak gradient

DOI 10.20347/WIAS.PREPRINT.3152 Berlin 2024

of u. In operator form we have $A = -\Delta + \sum_{i=1}^{n} b_i \partial_{x_i} + k\mathrm{Id}$ with Δ representing the weak Laplacian and Id the identity map. The setting $b_i = 0$ for $i = 1, \dots, n$, $k = 0$, $h \equiv 0$ is the prototypical example of an elliptic variational inequality and is commonly referred to as the obstacle problem [39].

Variational inequalities of the type (1) have numerous applications in diverse scientific areas; we mention in particular contact mechanics, processes in biological cells, ecology, fluid flow, and finance, see for example [52, 39, 55]. They are also fundamental objects of study in applied analysis due to their interesting structure. Indeed, VIs are examples of free boundary problems. Obstacle problems sometimes are stated in the form

$$0 \leq (Au - f) \perp (u - \psi) \geq 0 \quad \text{a.e. on } \Omega, \tag{5a}$$

$$u = h \quad \text{a.e on } \partial\Omega, \tag{5b}$$

where $a \perp b$ stands for $ab = 0$. This formulation is equivalent to (1) under sufficient regularity (see Proposition 2.2). Classical methods for solving obstacle problems or variational inequalities include projection methods, multilevel and multigrid methods [40, 30], primal dual active set strategies and path following schemes [29], semismooth Newton schemes [25, 33], shape and topological sensitivity based methods [26, 27], level set methods and discontinuous Galerkin schemes [58]; see also [20, 9, 37, 19] and references therein.

The aim of this work is to formulate, analyse and implement a deep neural network approach to compute solutions of obstacle problems like (1) or (5). More specifically, we rephrase the VI as a minmax optimisation problem involving minimisation over the feasible set and maximisation over all feasible test functions, both of which are parametrised by neural networks, and we use a modified gradient descent ascent scheme to numerically solve for the solution. Our motivation stems from the fact that neural networks can efficiently represent nonlinear, nonsmooth functions and have the added advantage of not being intrinsically reliant on a mesh: they provide a naturally global and meshless representation of the solution, offering an advantage over other methods such as finite elements. Furthermore, they are able to handle complicated geometries and high-dimensional problems without great cost. Our work can also be considered as a first step in studying more complicated problems involving for example operator learning.

Related papers in the literature addressing solving elliptic variational inequalities or hemivariational inequalities via neural networks include [13, 64, 53, 31, 7]. A typical path taken by many works entails rewriting (1) as a minimisation problem. Indeed, if the operator A generates a bilinear form which is symmetric, then, as explained, e.g., in [52, §4:3, Remark 3.5, p. 97], the VI (1) is equivalent to the minimisation problem

$$\min_{u \in K} \frac{1}{2}\langle Au, u\rangle - \langle f, u\rangle.$$

This formulation gives rise to a natural loss function that can be tackled via neural networks, as done in [13, 31, 64]. If A is non-symmetric, this equivalence is not available and one cannot in general pose an associated minimisation problem. However, as we shall see later, (1) is equivalent to the minmax problem

$$\min_{u \in K} \max_{v \in K} \langle Au - f, u - v\rangle - \frac{1}{2\gamma} \|u - v\|_V^2 \tag{6}$$

for a given parameter $\gamma > 0$, regardless of whether A is symmetric or not. This problem formulation appears very natural since it resembles the notion of weak formulations in PDEs, which are well understood.

In order to approximate (6) we express u and the test function v by deep neural networks. This technique falls into the class of weak adversarial network (WAN) problems in the spirit of [62]; the maximisation for the test function acts as an adversarial force against the minimisation for the solution. In

addition to [62], see also e.g. [8, 10, 57, 35] for weak adversarial approaches for PDEs and related theory. Moreover, our approach is related to Physics Informed Neural Networks (PINNs); see, e.g., [41, 50]. More specifically it can be viewed as a weak PINNs-type approach with a hard constrained boundary condition.

In this work, we will provide a theoretical justification for the minmax problem, a discretised formulation of the problem amenable to computation, an error analysis as well as comprehensive numerical simulations. As we mentioned above, a special highlight of our work is that we can also tackle non-symmetric problems, e.g., A given as in (4) with $b_i \neq 0$ for at least one $i = 1, \dots, n$.

2 Analysis of the continuous problem

We begin with some theoretical results concerning the VI (1).

2.1 Basic properties and saddle points

Let us first of all address existence and uniqueness for (1). Since A is coercive and Lipschitz on $H^1(\Omega)$ and K is non-empty[1], closed and convex, well posedness follows from the classical Lions–Stampacchia theorem [52, §4:3, Theorem 3.1] (see also [52, §4:2, Theorem 2.3] for the case where $A = -\Delta$).

Proposition 2.1 (H^2-regularity)**.** *Let $A := -\Delta + \sum_{i=1}^n b_i \partial_{x_i} + c\mathrm{Id}$ with coefficients $b_i, c \in L^\infty(\Omega)$, $c \geq c_0 \geq 0$ for a constant c_0, such that the coercivity condition* (3b) *is satisfied, $f \in L^2(\Omega)$, $h \in H^{3/2}(\partial\Omega)$, $\psi \in H^2(\Omega)$ with $\psi|_{\partial\Omega} \leq h$, and let Ω be convex or $C^{1,1}$. Then the solution of* (1) *has the regularity $u \in H^2(\Omega)$. Furthermore, we have the* a priori *estimate $\|u\|_{H^2(\Omega)} \leq C^*$, where C^* is a constant (that depends in particular on f, ψ and h).*

Proof. In the case $h \equiv 0$, this follows by the Lewy–Stampacchia inequality [52, Theorem 3.3, §5:3] $f \leq Au \leq \max(f, A\psi)$ and from elliptic estimates which bound the H^2-norm of u by the L^2 norm of Au and the H^1-norm of u, see e.g. [52, Theorem 3.4, §5:3] for the case $\Omega \in C^{1,1}$, and [24, Proposition 2.5] for when Ω is convex. In the general inhomogeneous case of $h \in H^{3/2}(\partial\Omega)$, we can find $\tilde{h} \in H^2(\Omega)$ with $\tilde{h}|_{\partial\Omega} = h$ and make a substitution $\tilde{u} := u - \tilde{h}$ (like done in the proof of [52, Theorem 2.3, §4:2]) and resort to the above situation. □

For a direct proof in the setting when $A = -\Delta$, see [52, Proposition 2.2 and Corollary 2.3, §5:2]. We will now consider equivalent reformulations of the problem (1). Since the proof of the following result mostly follows the argumentation in [52, §1:3, p. 4], we have placed it in Appendix A.

Proposition 2.2. *If $u \in H^1(\Omega)$ satisfies $Au \in L^2(\Omega)$ and* (5)*, then it solves* (1)*. Conversely, if u solves* (1) *and satisfies $u \in C^0(\overline{\Omega})$ and $Au \in L^2(\Omega)$, and $\psi \in C^0(\overline{\Omega})$, then u solves* (5)*.*

The VI (1) has a saddle point reformulation (as discussed in [20, page 14]). Let us recall the definition first.

[1] By properties of the trace operator [59, Theorem 8.8, Chapter 1], there exists a function $w \in H^1(\Omega)$ with $w|_{\partial\Omega} = h$, and the function $\max(w, \psi) \in H^1(\Omega)$ belongs to K.

Definition 2.3. *Given a map* $\mathfrak{f}: X \times Y \to \mathbb{R}$*, a pair* $(x^*, y^*) \in X \times Y$ *is called a* saddle point *[63, §9.6] if it satisfies*

$$\mathfrak{f}(x^*, y) \leq \mathfrak{f}(x^*, y^*) \leq \mathfrak{f}(x, y^*) \quad \forall x \in X, \forall y \in Y,$$

or equivalently

$$\max_{y \in Y} \mathfrak{f}(x^*, y) = \mathfrak{f}(x^*, y^*) = \min_{x \in X} \mathfrak{f}(x, y^*) \quad \forall x \in X, \forall y \in Y.$$

Associated to the VI (1) we define a map $L: V \times V \to \mathbb{R}$ as

$$L(u, v) := \langle Au - f, u - v \rangle. \tag{7}$$

Note that $L(\cdot, v)$ is convex and $L(u, \cdot)$ is concave for fixed $u, v \in V$. From, e.g., [63, Corollary 9.16], if (u, v) is a saddle point then it satisfies

$$L(u, v) = \min_{\tilde{u} \in K} \sup_{\tilde{v} \in K} L(\tilde{u}, \tilde{v}) = \max_{\tilde{v} \in K} \inf_{\tilde{u} \in K} L(\tilde{u}, \tilde{v}). \tag{8}$$

Moreover, L has a saddle point if and only if the second equality holds. Note that we have supremum and infimum above as these optimising elements may not be attained in general. However if K is a bounded set, then the $\sup$ and $\inf$ can be replaced by $\max$ and $\min$. In addition to the above references we cite also [17, Chapter VI, §1] for more details. Now, regarding the connection of saddle points of L to the VI (1), we have the following result, see [20, Remark 2.7].

Proposition 2.4. *If* $u \in K$ *is a solution of* (1) *then* (u, u) *is a saddle point of* L *on* $K \times K$*. Conversely, if* (u, v) *is a saddle point of* L *on* $K \times K$ *then* $u = v$ *and* u *solves* (1)*.*

2.2 Minmax approach via the regularised gap function

As described above, the saddle point formulation of the VI is a $\min\sup$ problem involving $L(u, v) = \langle Au - f, u - v \rangle$. It turns out that modifying L by adding a quadratic term makes the problem more tractable and endows it with better properties. First, let us make the following assumption which allows us to be more general with the formulation of the $\min\sup$ problem.

Assumption 2.5. *Let* u^* *be a solution of* (1)*, and let* X *and* Y *be sets such that*

(i) $u^* \in X \cap Y$*,*

(ii) $K \cap X \subset K \cap Y$*,*

(iii) $K \cap Y$ *is convex and closed*[2]*.*

We have in mind two examples: the standard setting $X = Y = V$, in which case the intersections above involving K are simply equal to K, and secondly $Y = B^{H^2(\Omega)}(r)$ the closed ball in $H^2(\Omega)$ with center 0 and an appropriately chosen radius r (which is valid when u^* has the regularity $u^* \in H^2(\Omega)$ and we have an estimate for its norm in that space; see Proposition 2.1 for conditions that ensure this) and $X \subset Y$ is some subset. This latter option is useful because it allows us to use neural network approximation results in the error analysis later.

[2] The closedness is required for the unique solvability of the minmax problem (9).

Now, let us introduce, for a fixed $\gamma > 0$, the map $L_\gamma \colon V \times V \to \mathbb{R}$ defined

$$L_\gamma(u, v) := L(u, v) - \frac{1}{2\gamma} \|u - v\|_V^2 ,$$

and the *generalised regularised gap function* $G_\gamma \colon V \to \mathbb{R}$ given by

$$G_\gamma(u) := \sup_{v \in K \cap Y} L_\gamma(u, v) = \sup_{v \in K \cap Y} \left(L(u, v) - \frac{1}{2\gamma} \|u - v\|_V^2 \right).$$

When $X = Y = V$, $G_\gamma(\cdot) = \sup_{v \in K} L_\gamma(\cdot, v)$ is known in the literature as the *regularised gap function*. In addition, the case $\gamma = \infty$, i.e., the function $G_\infty(u) = \sup_{v \in K} L(u, v)$ is called a *gap function* (it is in general non-differentiable and the supremum may be infinite). For simplicity we will just refer to G_γ above without reference to 'generalised'. The quadratic term in G_γ serves to add differentiability and ensures that the supremum is finite. These concepts and properties when $X = Y$ is the whole space V are well known, see [6, 18, 5] for the originating works; we refer also to [42, 61].

Since $\|\cdot\|_V^2$ is strongly convex, it follows that $-L_\gamma(u, \cdot)$ is strongly convex for fixed u; this will be important later for existence. For now, we prove some basic properties.

Lemma 2.6. *The function G_γ satisfies the following properties:*

(i) G_γ *is finite everywhere on* V*,*

(ii) $G_\gamma(u) \geq 0$ *for* $u \in K \cap X$*,*

(iii) $u \in K \cap X$ *satisfies* $G_\gamma(u) = 0$ *if and only if* u *solves* (1)*,*

(iv) $G_\gamma \colon K \cap X \subset V \to \mathbb{R}$ *is lower semicontinuous.*

The proof is standard for the usual case $X = Y = V$, see e.g., [32, Theorem 3.1].

Proof. (i) This follows easily by Young's inequality with parameter $\epsilon > 0$ chosen carefully.

(ii) We have $G_\gamma(u) \geq L_\gamma(u, v)$ for all $v \in K \cap Y$ and we can select $v = u$ since $u \in K \cap X \subset K \cap Y$. The claim follows from $L_\gamma(u, u) = 0$.

(iii) If u solves (1), then $G_\gamma(u) \leq \sup_{v \in K \cap Y} -\frac{1}{2\gamma} \|u - v\|_V^2 \leq 0$, which combined with the non-negativity of G_γ gives $G_\gamma(u) = 0$.

For the converse, if $u \in K \cap X$ satisfies $G_\gamma(u) = 0$, we have

$$0 = \sup_{v \in K \cap Y} \left(L(u, v) - \frac{1}{2\gamma} \|u - v\|_V^2 \right) \geq \left(L(u, v) - \frac{1}{2\gamma} \|u - v\|_V^2 \right) \quad \forall v \in K \cap Y,$$

and with $\lambda \in (0, 1)$, selecting $v = (1 - \lambda)u + \lambda w$ where $w \in K \cap Y$ is arbitrary (by convexity, $v \in K \cap Y$ too), manipulating and taking $\lambda \to 0$ we get

$$\langle Au - f, u - w \rangle \leq 0 \quad \forall w \in K \cap Y.$$

Now we also know that (1) is uniquely solvable with u^* as solution, thus

$$\langle Au^* - f, u^* - v \rangle \leq 0 \quad \forall v \in K.$$

Picking $w = u^*$ (which by assumption belongs to Y) and $v = u$ and using coercivity, we find $u = u^*$.

(iv) This follows by the continuity of $u \mapsto L_\gamma(u, v)$ from $H^1(\Omega)$ into $\mathbb{R}$ for fixed v; the argumentation is the same as in [32, Lemma 3.1].

□

Due to Lemma 2.6 (ii) and (iii), to find the solution of (1), we can solve

$$\min_{u \in K \cap X} G_\gamma(u) \equiv \min_{u \in K \cap X} \sup_{v \in K \cap Y} L_\gamma(u, v) \tag{9}$$

(cf. (8)). Now, we would like to consider instead of the supremum above the maximum. First observe that $-L_\gamma(u, \cdot)$ is strongly convex (as mentioned above), continuous, and also proper thanks to the quadratic term. Since we also assumed that $K \cap Y$ is closed, then by standard optimisation theory the problem

$$\max_{v \in K \cap Y} L_\gamma(u, v)$$

has a unique solution. Thus instead of (9), we can consider

$$\min_{u \in K \cap X} G_\gamma(u) \equiv \min_{u \in K \cap X} \max_{v \in K \cap Y} L_\gamma(u, v). \tag{10}$$

A natural question now arises as to what we mean by a solution to this problem.

Definition 2.7. *Given a map* $\mathfrak{f} \colon X \times Y \to \mathbb{R}$ *(which need not be convex-concave), by a* (global) solution *or a* (global) minmax point *of the minmax problem*

$$\min_{x \in X} \max_{y \in Y} \mathfrak{f}(x, y), \tag{11}$$

following [36, Definition 9 and Remark 10], we mean a pair (x^*, y^*) *that satisfies*

$$\mathfrak{f}(x^*, y) \leq \mathfrak{f}(x^*, y^*) \quad \forall y \in Y \quad \textit{and} \quad \max_{y \in Y} \mathfrak{f}(x^*, y) \leq \max_{y \in Y} \mathfrak{f}(x, y) \quad \forall x \in X$$

or equivalently

$$\mathfrak{f}(x^*, y) \leq \mathfrak{f}(x^*, y^*) \leq \max_{\hat{y} \in Y} \mathfrak{f}(x, \hat{y}) \qquad \forall (x, y) \in X \times Y.$$

Observe that saddle points are global minmax points. In the context of finite dimensional problems, existence of solutions to (11) holds if $X \subset \mathbb{R}^{d_1}$ and $Y \subset \mathbb{R}^{d_2}$ are compact and $\mathfrak{f}$ is continuous [36, Proposition 11]. See also Proposition 2.6 of [34] (and the proceeding paragraph) for another existence result.

The next lemma is a crucial tool as it will give us an error estimate later on.

Lemma 2.8. *We have*

$$\left(C_a - \frac{1}{2\gamma}\right) \|u - u^*\|_V^2 + \langle Au^* - f, u - u^* \rangle \leq G_\gamma(u) \qquad \forall u \in V.$$

If $u \in K$*, the duality product can be omitted.*

Proof. For any $u \in V$ (not necessarily in $K \cap X$), since $u^* \in Y$, we have

$$\begin{aligned} G_\gamma(u) &= \max_{v \in K \cap Y} \langle Au - f, u - v \rangle - \frac{1}{2\gamma} \|u - v\|_V^2 \geq \langle Au - f, u - u^* \rangle - \frac{1}{2\gamma} \|u - u^*\|_V^2 \\ &= \langle A(u - u^*) + Au^* - f, u - u^* \rangle - \frac{1}{2\gamma} \|u - u^*\|_V^2 \\ &\geq \left(C_a - \frac{1}{2\gamma}\right) \|u - u^*\|_V^2 + \langle Au^* - f, u - u^* \rangle. \end{aligned}$$

□

From now on we will assume that γ is such that

$$C_a - \frac{1}{2\gamma} > 0, \qquad \text{i.e.,} \qquad \gamma > \frac{1}{2C_a}. \tag{12}$$

2.3 Relaxations of the problem

We now look at relaxing the minmax problem of interest (9) in ways that are more amenable to implementation, as optimising over a constraint set involving K is non-trivial. Let us define the loss function $L_o\colon L^2(\Omega) \to \mathbb{R}$ corresponding to the obstacle constraint and the loss function $L_b\colon L^2(\partial\Omega) \to \mathbb{R}$ for the boundary condition by

$$L_o(u) = \int_\Omega |(\psi - u)^+|^2 \qquad \text{and} \qquad L_b(u) = \int_{\partial\Omega} |u - h|^2 \tag{13}$$

respectively. These measure violations of the constraints.

Imposition of the boundary conditions The idea is the following. We take a function $\bar{h} \in V$ with $\bar{h}|_{\partial\Omega} = h$ (i.e., $\bar{h}$ is a lift or extension of the Dirichlet data to the interior of the domain) and posit that the solution and test functions are of the form $\bar{h} + z$ for $z \in H_0^1(\Omega)$, and given penalty parameters w_{o_1} and w_{o_2}, we solve

$$\min_{u \in H_0^1(\Omega)+\bar{h}} \max_{v \in H_0^1(\Omega)+\bar{h}} L_\gamma(u,v) + w_{o_1} L_o(u) - w_{o_2} L_o(v). \tag{P_{bc}}$$

That is, the boundary condition is satisfied exactly and violations of the obstacle constraint are penalised.

Full penalty approach In this case we seek to optimise over the entire Hilbert space and penalise violations of *both* the obstacle constraint and boundary condition. In contrast to the previous approach, the boundary condition is not enforced at the outset. This leads us to consider for penalty parameters w_{o_i}, w_{b_i} (for $i = 1, 2$) the problem

$$\min_{u \in V} \max_{v \in V} L_\gamma(u,v) + w_{o_1} L_o(u) + w_{b_1} L_b(u) - w_{o_2} L_o(v) - w_{b_2} L_b(v). \tag{$\mathsf{P}_{\mathrm{pen}}$}$$

Generalised approach Clearly, both of these approaches can be viewed as special cases of a single more general problem by choosing the function spaces and the weights for the penalty terms correctly. It becomes convenient later to take this viewpoint. Define for $i = 1, 2$ the penalty functions $R_i\colon H^1(\Omega) \to \mathbb{R}$ by

$$R_1(u) := w_{b_1} L_b(u) + w_{o_1} L_o(u) \qquad \text{and} \qquad R_2(v) = w_{b_2} L_b(v) + w_{o_2} L_o(v).$$

A generalised problem can be posed as

$$\min_{u \in X^{\mathrm{s}}} \max_{v \in X^{\mathrm{t}}} L_\gamma(u,v) + R_1(u) - R_2(v) \tag{$\mathsf{P}_{\mathrm{gen}}$}$$

where X^{s} and X^{t} are sets. It is (a discrete version of) this problem that we shall provide an error analysis for inasmuch as it possible to do so, in order to allow for the greatest generality possible.

3 Neural network approach

We wish to compute (approximate) solutions of the VI (1) using neural networks. The architecture we use is essentially the residual neural network considered in [16] consisting of the usual affine transformations and activations combined with skip connections, inspired by the original work [22]. Residual networks or ResNets have been empirically observed to be better at training deep networks and they avoid the vanishing gradient problem, see for example [23, 60] for some analysis.

Let us describe this special ResNet architecture precisely. Let $\mathfrak{d}, \mathfrak{w} \in \mathbb{N}$ be given positive integers (representing the depth and width of the network respectively). Given an initial weight $A_0 \in \mathbb{R}^{\mathfrak{w}\times n}$ and bias $b_0 \in \mathbb{R}^{\mathfrak{w}}$, and for $i = 1, \dots, \mathfrak{d}$ and $j = 1, 2$, weights $A_{ij} \in \mathbb{R}^{\mathfrak{w}\times\mathfrak{w}}$ and biases $b_{ij} \in \mathbb{R}^{\mathfrak{w}}$ determining the affine transforms

$$T_{ij}z := A_{ij}z + b_{ij} \quad \text{for } i = 1, \dots, \mathfrak{d} \text{ and } j = 1, 2,$$

and a weight $A_{\mathfrak{d}+1} \in \mathbb{R}^{1\times\mathfrak{w}}$ and bias $b_{\mathfrak{d}+1} \in \mathbb{R}$ for the final layer, define

$$\begin{aligned} T_0(x) &:= A_0x + b_0, \\ \mathfrak{B}_i(z) &:= \sigma \circ T_{i2} \circ \sigma \circ T_{i1}(z) + z \quad \text{for } i = 1, \dots, \mathfrak{d}, \\ T_{\mathfrak{d}+1}(z) &:= A_{\mathfrak{d}+1}z + b_{\mathfrak{d}+1}. \end{aligned}$$

Note that each $\mathfrak{B}_i$ for $i = 1, \dots, \mathfrak{d}$ has the so-called *block* structure; each block comprises two affine transformations and two activations with a residual connection, and we have that $\mathfrak{B}_i \colon \mathbb{R}^{\mathfrak{w}} \to \mathbb{R}^{\mathfrak{w}}$. With this notation at hand, the class of neural networks under consideration in this work consists of functions with the form

$$u \colon \mathbb{R}^n \to \mathbb{R}, \qquad u(x) = T_{\mathfrak{d}+1} \circ \mathfrak{B}_{\mathfrak{d}} \circ \cdots \circ \mathfrak{B}_1 \circ T_0(x). \tag{14}$$

The coefficients of the matrices and vectors appearing above determine the learnable parameters of the neural network. The neural network above has $\mathfrak{d}$ blocks and in total a depth of $2\mathfrak{d} + 2$ layers (if we understand each block to have two layers; hence there are $2\mathfrak{d}$ hidden layers). We will write the set of all such neural networks of block depth $\mathfrak{d}$, width $\mathfrak{w}$ and activation function σ as

$$\mathcal{F}_{\mathrm{DRR}}(\mathfrak{d}, \mathfrak{w}, \sigma) := \{u \colon \mathbb{R}^n \to \mathbb{R} : u \text{ is given by (14)}\}$$

(DRR for *D*eep *R*itz *R*esidual neural network, from the title of the work [16] from where this structure originated). Elements of this set differ by different choices of the learnable parameters, i.e., the weights and biases. By denoting all the learnable parameters of the network (14) as

$$\theta = \{(A_0, b_0), (A_{\mathfrak{d}+1}, b_{\mathfrak{d}+1})\} \bigcup_{i=1}^{\mathfrak{d}} \{(A_{i1}, b_{i1}), (A_{i2}, b_{i2})\},$$

we can view any $u \in \mathcal{F}_{\mathrm{DRR}}(\mathfrak{d}, \mathfrak{w}, \sigma)$ as a function of the parameters, i.e., $u = u(\theta)$. The element θ belongs to the *weight space* (or *parameter space*)

$$\begin{aligned} \Theta := \Big\{ \theta = \{(A_0, b_0), (A_{\mathfrak{d}}, b_{\mathfrak{d}})\} \bigcup_{i=1}^{\mathfrak{d}} \{(A_{i1}, b_{i1}), (A_{i2}, b_{i2})\} \\ : \quad (A_0, b_0) \in \mathbb{R}^{\mathfrak{w}\times n} \times \mathbb{R}^{\mathfrak{w}}, (A_{\mathfrak{d}+1}, b_{\mathfrak{d}+1}) \in \mathbb{R}^{1\times\mathfrak{w}} \times \mathbb{R}, (A_{ij}, b_{ij}) \in \mathbb{R}^{\mathfrak{w}\times\mathfrak{w}} \times \mathbb{R}^{\mathfrak{w}} \Big\}. \end{aligned}$$

In total, the network has

$$\mathfrak{m} := n\mathfrak{w} + \mathfrak{w} + 2(\mathfrak{w}^2 + \mathfrak{w})\mathfrak{d} + \mathfrak{w} + 1 = 2\mathfrak{d}\mathfrak{w}^2 + (n + 2\mathfrak{d} + 2)\mathfrak{w} + 1$$

parameters to be determined hence $\Theta \subset \mathbb{R}^{\mathfrak{m}}$ can be viewed as a subset of Euclidean space. Note that the first layer is chosen so that the input $x \in \mathbb{R}^n$ is transformed to a $\mathfrak{w}$-dimensional object (which is necessary to apply the block structure). In the case that $n < \mathfrak{w}$, we could alternatively have padded the input with zeros.

The standard feedforward neural network structure ubiquitous in machine learning has a very similar architecture as described above except (more or less) without the addition of the z term in the definition of $\mathfrak{B}$. Let us write this set of neural networks as $\mathcal{F}_{\mathrm{FFN}}(\mathfrak{d}, \mathfrak{w}, \sigma)$ where an element of this space has the structure

$$u\colon \mathbb{R}^n \to \mathbb{R} \qquad u(x) = T_{\mathfrak{d}+1} \circ \sigma \circ T_{\mathfrak{d}} \circ \sigma \circ \cdots \circ T_1 \circ \sigma \circ T_0(x).$$

Note that $\mathcal{F}_{\mathrm{DRR}}(\mathfrak{d}, \mathfrak{w}, \sigma)$ has $2\mathfrak{d} + 2$ layers while $\mathcal{F}_{\mathrm{FFN}}(\mathfrak{d}, \mathfrak{w}, \sigma)$ has $\mathfrak{d} + 2$ layers; this mismatch is of little consequence as it is just a notational choice.

For convenience, we write $\mathcal{F}$ or $\mathcal{F}(\mathfrak{d}, \mathfrak{w}, \sigma)$ for a set of neural networks of arbitrary architecture (with depth $\mathfrak{d}$, width $\mathfrak{w}$ and activation σ; the depth and width are ambiguous), typically $\mathcal{F} \in \{\mathcal{F}_{\mathrm{DRR}}, \mathcal{F}_{\mathrm{FFN}}\}$ is what we have in mind.

Neural networks with homogeneous BC. Given a neural network class $\mathcal{F}$, it is useful to consider another set of functions $\mathcal{F}_0$ that have the same type of architecture as $\mathcal{F}$ and that are zero on the boundary.

One way to practically construct such a set is by taking a function $\eta \in C^1(\overline{\Omega})$ satisfying $\eta|_{\partial\Omega} = 0$, then given $v \in \mathcal{F}$, the function $v\eta \in \mathcal{F}_0$. That is, given $\tilde{u} \in \mathcal{F}$ this amounts to appending an extra final layer $M_\eta \circ \tilde{u}$ where $M_\eta\colon C^1(\overline{\Omega}) \to C^1(\overline{\Omega})$ defined by $M_\eta(\tilde{u}) := \tilde{u}\eta$ is the pointwise multiplication operator. Let us label this subset of $\mathcal{F}_0$ with a subscript η, e.g., as

$$\mathcal{F}_{0,\eta} := \{M_\eta \circ \tilde{u} : u \in \mathcal{F}\}. \tag{15}$$

Thus in particular, $\mathcal{F}_{\mathrm{DRR},0,\eta}(\mathfrak{d}, \mathfrak{w}, \sigma) := \{u\colon \mathbb{R}^n \to \mathbb{R} : u(x) = M_\eta \circ T_{\mathfrak{d}+1} \circ \mathfrak{B}_{\mathfrak{d}} \circ \cdots \circ \mathfrak{B}_1 \circ T_0(x)\}$. The function η is easy to construct when the domain is an interval or rectangle. For more complex geometries it is non-trivial and the desired C^1 property may not hold; see e.g. [56] for a way to construct such a function η (with lower regularity).

Neural networks with additional final layer. We can generalise this structure to cater to a more general final layer that could be different to pointwise multiplication. Indeed, given a operator $M\colon C^1(\overline{\Omega}) \to C^1(\overline{\Omega})$, let us define

$$\mathcal{F}(\mathfrak{d}, \mathfrak{w}, \sigma, M) := \{M \circ u : u \in \mathcal{F}(\mathfrak{d}, \mathfrak{w}, \sigma)\}. \tag{16}$$

Note that $\mathcal{F}_{0,\eta}(\mathfrak{d}, \mathfrak{w}, \sigma) \equiv \mathcal{F}(\mathfrak{d}, \mathfrak{w}, \sigma, M_\eta)$ in this notation.

We finish with a simple but useful result which follows by calculus arguments, see Appendix A for the proof.

Lemma 3.1. *Let $u \in \mathcal{F}_{\mathrm{DRR}}(\mathfrak{w}, \mathfrak{d}, \sigma, M) \cup \mathcal{F}_{\mathrm{FFN}}(\mathfrak{w}, \mathfrak{d}, \sigma, M)$ where $\sigma \in C^1(\mathbb{R})$ and $M\colon C^1(\bar{\Omega}) \to C^1(\bar{\Omega})$ is continuous. Then the map $\theta \mapsto u(\theta, \cdot)$ is continuous from $\mathbb{R}^{\mathfrak{m}}$ to $C^1(\overline{\Omega})$.*

Note that this result applies to elements of the space $\mathcal{F}_{0,\eta}$ for both architectures. It also implicitly tells us that if $\sigma \in C^1(\mathbb{R})$ and $M \in C^0(C^1(\bar{\Omega}), C^1(\bar{\Omega}))$, every function in $\mathcal{F}_{\mathrm{DRR}}(\mathfrak{w}, \mathfrak{d}, \sigma, M) \cup \mathcal{F}_{\mathrm{FFN}}(\mathfrak{w}, \mathfrak{d}, \sigma, M)$ is $C^1(\overline{\Omega})$ (which implies that ($\widehat{\mathsf{P}}_\ell$) below is well defined) and hence in $H^1(\Omega)$, i.e., $\mathcal{F}(\mathfrak{w}, \mathfrak{d}, \sigma, M) \subset C^1(\overline{\Omega}) \subset H^1(\Omega)$ for both architectures. Though ReLU is not C^1, it is well known that we still have $\mathcal{F}(\mathfrak{w}, \mathfrak{d}, \mathrm{ReLU}) \subset H^1(\Omega)$ for both architectures.

3.1 Discretised problems

We need discrete versions of the objective function appearing in (10). We shall henceforth require the regularity $f \in L^2(\Omega) \cap C^0(\Omega)$. For simplicity, we assume that $A = -\Delta + k\mathrm{Id}$ for a constant $k \geq 0$, generating the pairing

$$\langle Au, u - v \rangle = \int_\Omega \nabla u \cdot \nabla(u - v) + ku(u - v)$$

(other cases can be easily handled with appropriate modifications). Take a set of collocation points $\{x_i\}_{i=1}^N$ and define $\hat{L} \colon C^1(\overline{\Omega}) \times C^1(\overline{\Omega}) \to \mathbb{R}$ by

$$\begin{aligned}\hat{L}(u, v) := \frac{|\Omega|}{N} \sum_{i=1}^N \nabla u(x_i) \cdot (\nabla u(x_i) - \nabla v(x_i)) + ku(x_i)(u(x_i) - v(x_i)) \\ - \frac{|\Omega|}{N} \sum_{i=1}^N f(x_i)(u(x_i) - v(x_i)).\end{aligned}$$

This is the discrete version of L. Similarly, the discrete version of L_γ is $\hat{L}_\gamma \colon C^1(\overline{\Omega}) \times C^1(\overline{\Omega}) \to \mathbb{R}$ given by

$$\hat{L}_\gamma(u, v) := \hat{L}(u, v) - \frac{|\Omega|}{2\gamma N} \sum_{i=1}^N (u(x_i) - v(x_i))^2 + |\nabla u(x_i) - \nabla v(x_i)|^2.$$

Remark 3.2. *One should bear in mind that $\hat{L}$ and $\hat{L}_\gamma$ (and all other discretised quantities that we shall introduce) are also functions of the grid points, i.e., $\hat{L}(u, v) = \hat{L}^{\{x_i\}}(u, v)$. A different choice of $\{x_i\}$ gives a different $\hat{L}$. We usually omit this dependence for aesthetic reasons.*

Let $\mathcal{X}^{\mathrm{s}}$ and $\mathcal{X}^{\mathrm{t}}$ be arbitrary subsets of $C^1(\bar{\Omega})$. Later, we will take these sets as sets of neural networks representing the solution and test function space respectively (with potentially different widths, depths and activations, etc., cf. the Petrov–Galerkin method) of the form (16), intersected with a ball of a carefully chosen radius. A discretised problem corresponding to (10) is

$$\min_{u \in \mathcal{X}^{\mathrm{s}}} \max_{v \in \mathcal{X}^{\mathrm{t}}} \hat{L}_\gamma(u, v). \tag{$\widehat{\mathsf{P}}$}$$

In general, $\mathcal{X}^{\mathrm{s}}$ and $\mathcal{X}^{\mathrm{t}}$ need not be subsets of K and it could be that they are not rich enough for the above problem to be a good approximation (see Remark 3.5 below). In view of this and the relaxations we surveyed in Section 2.3, let us consider some alternatives.

3.1.1 Imposition of the boundary conditions

To formulate a discrete version of (P_{bc}), let us first assume the regularity $\psi \in C^0(\Omega)$ and define the discrete obstacle constraint loss corresponding to L_o as

$$\hat{L}_o(u) := \frac{|\Omega|}{N} \sum_{i=1}^{N} |(\psi(x_i) - u(x_i))^+|^2.$$

For a given function $\tilde{h} \in C^1(\overline{\Omega})$ satisfying $\tilde{h}|_{\partial\Omega} = h$ and taking

$$\mathcal{F}_0^{\mathfrak{s}} := \mathcal{F}_0(\mathfrak{w}^{\mathfrak{s}}, \mathfrak{b}^{\mathfrak{s}}, \sigma^{\mathfrak{s}}) \qquad \text{and} \qquad \mathcal{F}_0^{\mathfrak{t}} := \mathcal{F}_0(\mathfrak{w}^{\mathfrak{t}}, \mathfrak{b}^{\mathfrak{t}}, \sigma^{\mathfrak{t}})$$

(as mentioned, a practical way to realise these sets is to use $\mathcal{F}_{0,\eta}$ as defined in (15)), we pose

$$\min_{u \in \mathcal{F}_0^{\mathfrak{s}} + \tilde{h}} \max_{v \in \mathcal{F}_0^{\mathfrak{t}} + \tilde{h}} \hat{L}_\gamma(u, v) + w_{o_1} \hat{L}_o(u) - w_{o_2} \hat{L}_o(v). \tag{$\widehat{\mathrm{P}}_{\mathrm{bc}}$}$$

Remark 3.3. *The problem* ($\widehat{\mathrm{P}}_{\mathrm{bc}}$) *can be formulated as*

$$\min_{\tilde{u} \in \mathcal{F}_0^{\mathfrak{s}}} \max_{\tilde{v} \in \mathcal{F}_0^{\mathfrak{t}}} \hat{L}_\gamma(\tilde{u} + \tilde{h}, \tilde{v} + \tilde{h}) + w_{o_1} \hat{L}_o(\tilde{u} + \tilde{h}) - w_{o_2} \hat{L}_o(\tilde{v} + \tilde{h}). \tag{17}$$

Then if $\tilde{u}$ denotes a solution of (17)*, to recover a solution of* ($\widehat{\mathrm{P}}_{\mathrm{bc}}$) *we set* $u := \tilde{u} + \tilde{h}$.

3.1.2 Full penalty approach

Here we wish to formulate the approach of ($\mathrm{P}_{\mathrm{pen}}$) in the discrete setting. In order to measure the boundary loss for the discretised problem, we introduce $\{x_i^b\}_{i=1}^{N_b}$ as a set of boundary collocation points, and assuming $\psi \in C^0(\Omega),\ h \in C^0(\partial\Omega)$, defining

$$\hat{L}_b(u) := \frac{|\partial\Omega|}{N_b} \sum_{i=1}^{N_b} (u(x_i^b) - h(x_i^b))^2,$$

we can pose

$$\min_{u \in \mathcal{F}^{\mathrm{s}}(\mathfrak{b}^{\mathrm{s}}, \mathfrak{w}^{\mathrm{s}}, \sigma^{\mathrm{s}})} \max_{v \in \mathcal{F}^{\mathrm{t}}(\mathfrak{b}^{\mathrm{t}}, \mathfrak{w}^{\mathrm{t}}, \sigma^{\mathrm{t}})} \hat{L}_\gamma(u, v) + w_{o_1} \hat{L}_o(u) + w_{b_1} \hat{L}_b(u) - w_{o_2} \hat{L}_o(v) - w_{b_2} \hat{L}_b(v). \tag{$\widehat{\mathrm{P}}_{\mathrm{pen}}$}$$

3.1.3 Generalised approach

Now let us consider the generalised problem ($\mathrm{P}_{\mathrm{gen}}$). Defining the discrete versions $\hat{R}_1$ and $\hat{R}_2$ of R_1 and R_2 (by using $\hat{L}_b$ and $\hat{L}_o$), we can consider

$$\min_{u \in \mathcal{X}^{\mathrm{s}}} \max_{v \in \mathcal{X}^{\mathrm{t}}} \hat{L}_\gamma(u, v) + \hat{R}_1(u) - \hat{R}_2(v) \tag{$\widehat{\mathrm{P}}_{\mathrm{gen}}$}$$

given sets of neural networks $\mathcal{X}^{\mathrm{s}}$ and $\mathcal{X}^{\mathrm{t}}$. This structure generalises all of the above-mentioned approaches. We will denote by $\hat{u}$ and $\widehat{u}_A$ an exact solution to ($\widehat{\mathrm{P}}_{\mathrm{gen}}$) and a computed solution to ($\widehat{\mathrm{P}}_{\mathrm{gen}}$) via a numerical algorithm $\mathcal{A}$, respectively. In our numerical implementation, we will use a alternating gradient descent ascent algorithm (GDA) to solve our discrete problem, see Algorithm 1 on page 33 (GDA can be thought of as an inexact version of Uzawa's algorithm).

Define $\widehat{G}_\gamma \colon \mathcal{X}^{\mathrm{s}} \to \mathbb{R}$ by

$$\widehat{G}_\gamma(u) := \max_{v \in \mathcal{X}^{\mathrm{t}}} \hat{L}_\gamma(u, v) + \hat{R}_1(u) - \hat{R}_2(v). \tag{18}$$

This plays the role of a discrete version of the gap function associated to ($\widehat{\mathrm{P}}_{\mathrm{gen}}$). We remark that if $\hat{L}_\gamma(u,u) + \hat{R}_1(u) - \hat{R}_2(u) \geq 0$ and $\mathcal{X}^{\mathrm{s}} \subseteq \mathcal{X}^{\mathrm{t}}$, then it is easy to see that $\widehat{G}_\gamma(u) \geq 0$ for all $u \in \mathcal{X}^{\mathrm{s}}$.

Remark 3.4. *Suppose that $\mathcal{F}_0^{\mathrm{s}} \subseteq \mathcal{F}_0^{\mathrm{t}}$ and $w_{o_1} = w_{o_2} = 0$. If $u \in \mathcal{F}_0^{\mathrm{s}}$ satisfies*

$$\frac{|\Omega|}{N} \sum_{i=1}^{N} \nabla u(x_i) \cdot (\nabla u(x_i) - \nabla v(x_i)) - \frac{|\Omega|}{N} \sum_{i=1}^{N} f(x_i)(u(x_i) - v(x_i)) \leq 0 \quad \forall v \in \mathcal{F}_0^{\mathrm{t}},$$

then it is not difficult to see that $\widehat{G}_\gamma(u) = 0$. That is, solutions of the above VI solve the discrete problem ($\widehat{\mathrm{P}}_{\mathrm{bc}}$).

Remark 3.5 (Exact satisfaction of the constraints)**.** *In some circumstances it is possible to choose $\mathcal{X}^{\mathrm{s}}$ and $\mathcal{X}^{\mathrm{t}}$ such that both are subsets of K. This is in a sense the ideal setting as all elements are feasible and therefore there is no need to expend effort in trying to meet the constraints. Furthermore the error analysis of Section 3.2 is greatly simplified in this setting.*

A common scenario where $\mathcal{X}^{\mathrm{s}}, \mathcal{X}^{\mathrm{t}} \subset K$ can be achieved with ease is when the VI has the homogeneous boundary condition $h \equiv 0$ (a typical setting in the literature) and when, with only a small loss of generality[3], the obstacle is also zero. Then we can solve

$$\min_{u \in \mathcal{F}^{\mathrm{s}}(\mathfrak{d}^{\mathrm{s}}, \mathfrak{w}^{\mathrm{s}}, \sigma^{\mathrm{s}}, Q \circ M_\eta)} \max_{v \in \mathcal{F}^{\mathrm{t}}(\mathfrak{d}^{\mathrm{t}}, \mathfrak{w}^{\mathrm{t}}, \sigma^{\mathrm{t}}, Q \circ M_\eta)} \hat{L}_\gamma(u, v)$$

where M_η is as before and Q is the map $Q(u) = u^2$. Satisfying the constraints in a more general setting (with non-zero boundary conditions) appears to be a non-trivial undertaking. We could consider a class of neural networks that satisfies also the obstacle condition in addition to the boundary condition. Take $u \in \mathcal{F}_{0,\eta}(\mathfrak{d}, \mathfrak{w}, \sigma)$, a function $\tilde{h} \in C^1(\overline{\Omega})$ with $\tilde{h}|_{\partial\Omega} = h$ (like described above) and make the transformation $u \mapsto \max(u + \tilde{h}, \psi)$. Label the set of such functions $\mathcal{F}_h^\psi(\mathfrak{d}, \mathfrak{w}, \sigma)$, which is a subset of K. One could then consider

$$\min_{u \in \mathcal{F}_h^\psi(\mathfrak{d}^{\mathrm{s}}, \mathfrak{w}^{\mathrm{s}}, \sigma^{\mathrm{s}})} \max_{v \in \mathcal{F}_h^\psi(\mathfrak{d}^{\mathrm{t}}, \mathfrak{w}^{\mathrm{t}}, \sigma^{\mathrm{t}})} \hat{L}_\gamma(u, v),$$

however, from experience, this does not work as well as hoped due in part to the nonsmooth $\max$ operation which hinders training.

Before we proceed to the error analysis, let us briefly discuss existence of solutions to these neural network problems.

3.1.4 Remarks on existence

It is well known that sets of neural networks may not be closed (nor convex), hampering the use of standard theory to deduce existence of optimal points for the discretised problems. One way to work around this issue is to use the notion of *quasi-minimisation* [54] (see also [12, 28]). An element

[3] If the obstacle ψ belongs to $H_0^1(\Omega)$ then one can always transform the VI to a VI with zero obstacle by $u \mapsto u - \psi$; this essentially moves the obstacle into the source term.

$\bar{u} \in U \subset X$ of a set U in a Hilbert space X is said to be a *quasi-minimiser* of the functional $J\colon U \to \mathbb{R}$ if it satisfies $J(\bar{u}) \leq \inf_{u\in U} J(u) + \epsilon$ for some $\epsilon > 0$. One could then weaken the notion of global minmax points (see Definition 2.7 with the corresponding function $\mathfrak{f}$) and define a *quasi-minimax point* as a point (x^*, y^*) that satisfies

$$\mathfrak{f}(x^*, y) - \epsilon \leq \mathfrak{f}(x^*, y^*) \quad \forall y \in Y \quad \text{and} \quad \max_{y\in Y} \mathfrak{f}(x^*, y) \leq \max_{y\in Y} \mathfrak{f}(x, y) + \epsilon \quad \forall x \in X$$

for some $\epsilon > 0$, and then study the associated theory. However, as this is not the focus of our work, let us content ourselves with giving existence results under a simplified setting where the associated parameter space to the sets of neural networks is sufficiently regular.

Let us consider a general problem

$$\min_{u\in\mathcal{X}^{\mathrm{s}}} \max_{v\in\mathcal{X}^{\mathrm{t}}} \hat{\ell}(u, v) \tag{$\widehat{\mathsf{P}}_{\hat{\ell}}$}$$

under the following assumption.

Assumption 3.6 (Assumptions on ($\widehat{\mathsf{P}}_{\hat{\ell}}$))**.** *Let $\hat{\ell}\colon C^1(\overline{\Omega}) \times C^1(\overline{\Omega}) \to \mathbb{R}$ be a given map and assume that $\mathcal{X}^{\mathrm{s}}$ and $\mathcal{X}^{\mathrm{t}}$ are of the form*

$$\mathcal{X}^{\mathrm{s}} \subset \mathcal{F}(\mathfrak{d}^{\mathrm{s}}, \mathfrak{w}^{\mathrm{s}}, \sigma^{\mathrm{s}}, M^{\mathrm{s}}), \quad \mathcal{X}^{\mathrm{t}} \subset \mathcal{F}(\mathfrak{d}^{\mathrm{t}}, \mathfrak{w}^{\mathrm{t}}, \sigma^{\mathrm{t}}, M^{\mathrm{t}}),$$

with associated parameter spaces $\Theta^{\mathfrak{s}}$ and Θ^{t} respectively and where $\sigma^{\mathfrak{s}}, \sigma^{\mathrm{t}} \in C^1(\mathbb{R})$ and $M^{\mathrm{s}}, M^{\mathrm{t}}\colon C^1(\overline{\Omega}) \to C^1(\overline{\Omega})$ are given. Furthermore, assume that

(i) $\hat{\ell}\colon C^1(\overline{\Omega}) \times C^1(\overline{\Omega}) \to \mathbb{R}$ *is continuous,*

(ii) $M^{\mathrm{s}}, M^{\mathrm{t}} \in C^0(C^1(\overline{\Omega}), C^1(\overline{\Omega}))$*,*

(iii) $\Theta^{\mathfrak{s}}, \Theta^{\mathrm{t}}$ *are non-empty and compact.*

Here, we take the perspective that a *solution* to problem ($\widehat{\mathsf{P}}_{\hat{\ell}}$) is associated with solving the finite-dimensional problem

$$\min_{\theta\in\Theta^{\mathfrak{s}}} \max_{\vartheta\in\Theta^{\mathrm{t}}} \hat{\ell}(u(\theta, \cdot), v(\vartheta, \cdot)).$$

To be precise, in Assumption 3.6 we enforced conditions on the parameter spaces $\Theta^{\mathfrak{s}}$ and Θ^{t} associated to $\mathcal{X}^{\mathrm{s}}$ and $\mathcal{X}^{\mathrm{t}}$ which constrain the latter sets: elements of $\mathcal{X}^{\mathrm{s}}$ and $\mathcal{X}^{\mathrm{t}}$ have a limitation on how large the weights can be.

Proposition 3.7. *Let Assumption 3.6 hold. Then there exists a solution $\hat{u} \in \mathcal{X}^{\mathrm{s}}$ to the problem ($\widehat{\mathsf{P}}_{\hat{\ell}}$).*

Proof. As discussed above, we have by definition of our solution concept that

$$\min_{u\in\mathcal{X}^{\mathrm{s}}} \max_{v\in\mathcal{X}^{\mathrm{t}}} \hat{\ell}(u, v) = \min_{\theta\in\Theta^{\mathfrak{s}}} \max_{\vartheta\in\Theta^{\mathrm{t}}} \hat{\ell}(u(\theta, \cdot), v(\vartheta, \cdot)).$$

Using Lemma 3.1 and the continuity of $\hat{\ell}$, it follows that $(\theta, \vartheta) \to \hat{\ell}(u_\theta, v_\vartheta)$ is continuous. Applying [36, Proposition 11], we get the result. □

We now address ($\widehat{\mathsf{P}}_{\mathrm{bc}}$) and ($\widehat{\mathsf{P}}_{\mathrm{pen}}$) (note that the former is not a special case of the latter because the architectures are potentially different).

Corollary 3.8. *We have the following.*

(i) In the context of ($\widehat{\mathrm{P}}_{\mathrm{bc}}$)*, let* $\sigma^{\mathrm{s}}, \sigma^{\mathrm{t}} \in C^1(\mathbb{R})$ *and let the parameter spaces* Θ^{s} *and* Θ^{t} *associated to* $\mathcal{F}_0^{\mathrm{s}}$ *and* $\mathcal{F}_0^{\mathrm{t}}$ *respectively be non-empty and compact. Then the problem* ($\widehat{\mathrm{P}}_{\mathrm{bc}}$) *possesses a solution.*

(ii) In the context of ($\widehat{\mathrm{P}}_{\mathrm{pen}}$)*, let* $\sigma^{\mathfrak{s}}, \sigma^{\mathfrak{t}} \in C^1(\mathbb{R})$ *and let the parameter spaces* $\Theta^{\mathfrak{s}}$ *and* $\Theta^{\mathfrak{t}}$ *associated to* $\mathcal{F}^{\mathrm{s}}(\mathfrak{d}^{\mathfrak{s}}, \mathfrak{w}^{\mathfrak{s}}, \sigma^{\mathfrak{s}})$ *and* $\mathcal{F}^{\mathrm{t}}(\mathfrak{d}^{\mathfrak{t}}, \mathfrak{w}^{\mathfrak{t}}, \sigma^{\mathfrak{t}})$ *respectively be non-empty and compact. Then the problem* ($\widehat{\mathrm{P}}_{\mathrm{pen}}$) *possesses a solution.*

(iii) In the context of (18)*, let* $\theta \mapsto (u(\theta, \cdot), v(\theta, \cdot))$ *be continuous from* $\mathbb{R}^{\mathfrak{m}}$ *to* $C^1(\overline{\Omega}) \times C^1(\overline{\Omega})$ *for all* $u \in \mathcal{X}^{\mathrm{s}}$ *and* $v \in \mathcal{X}^{\mathrm{t}}$ *and let the parameter spaces* $\Theta^{\mathfrak{s}}$ *and* $\Theta^{\mathfrak{t}}$ *associated to* $\mathcal{X}^{\mathrm{s}}$ *and* $\mathcal{X}^{\mathrm{t}}$ *respectively be non-empty and compact. Then the problem* ($\widehat{\mathrm{P}}_{\mathrm{pen}}$) *possesses a solution.*

Proof. For (i), we consider the reformulation (17) of ($\widehat{\mathrm{P}}_{\mathrm{bc}}$). Thanks to Proposition 3.7, it suffices to show that $M_\eta \in C^0(C^1(\overline{\Omega}), C^1(\overline{\Omega}))$ and that the resulting loss function is continuous from $C^1(\overline{\Omega}) \times C^1(\overline{\Omega})$ to $\mathbb{R}$. The former is clear because $\eta \in C^1(\overline{\Omega})$. For the latter, we begin by setting $\tilde{h} = 0$ for simplicity. Define the pointwise evaluations $S_i(u) := u(x_i)$ and $T_i(u) := \nabla u(x_i)$. Then we can write

$$\begin{aligned}\hat{L}(u, v) = \frac{|\Omega|}{N} \sum_{i=1}^{N} (T_i(u), T_i(u) - T_i(v))_{\mathbb{R}^n} + \frac{|\Omega|}{N} \sum_{i=1}^{N} k S_i(u)(S_i(u) - S_i(v)) \\ - \frac{|\Omega|}{N} \sum_{i=1}^{N} f(x_i)(S_i(u) - S_i(v)).\end{aligned}$$

As $S_i \colon C^1(\overline{\Omega}) \to \mathbb{R}$ and $T_i \colon C^1(\overline{\Omega}) \to \mathbb{R}^n$ are clearly continuous, the composition of continuous maps being continuous gives continuity of $\hat{L} \colon C^1(\overline{\Omega}) \times C^1(\overline{\Omega}) \to \mathbb{R}$. Writing also

$$\hat{L}_o(u) = \frac{|\Omega|}{N} \sum_{i=1}^{N} |(S_i(\psi) - S_i(u))^+|^2,$$

the continuity of the functions $(\cdot)^+, |\cdot|^2 \colon \mathbb{R} \to \mathbb{R}$ yields continuity of this object too. The remaining terms in the objective function in (17) can be tackled similarly. The claims (ii) and (iii) follow similarly. □

It is clear that existence for the problem ($\widehat{\mathrm{P}}$) can also be handled with a similar argument like above.

Note that when transiting from the continuous problem (9) to any of the associated discrete problems above, the (strong) convexity and concavity properties for the former are lost. This is due to the structure of the mapping from the neural network weights to the neural network output $(\theta_1, \theta_2) \mapsto (u_{\theta_1}, v_{\theta_2})(x)$. As a consequence, the discrete problems are non-convex in the θ_1 variable and non-concave in the θ_2 variable which imposes major difficulties concerning the characterisation and efficient computation of solutions. For an overview in the smooth setting we refer to the recent works [36, 51] and the references therein. Clearly, the design of tailored numerical solvers for such neural network based problems is a very important research task which, however, goes beyond the scope of our present work. Subsequently, we focus on a comprehensive error analysis regarding the true and a computed solution for (1).

3.2 Error analysis

Recall that we consider the general problem ($\widehat{\mathrm{P}}_{\mathrm{gen}}$), as described in Section 3.1.3. As mentioned before, by choosing the sets $\mathcal{X}^{\mathrm{s}}$ and $\mathcal{X}^{\mathrm{t}}$ and the weights appearing in R_1 and R_2 appropriately, we

can cover both cases ($\widehat{\mathsf{P}}_{\mathrm{bc}}$) and ($\widehat{\mathsf{P}}_{\mathrm{pen}}$) in this formulation. From the error estimate in Lemma 2.8 and the minimisation property Lemma 2.6 (iii),

$$\left(C_a - \frac{1}{2\gamma}\right) \|\widehat{u}_A - u^*\|_V^2 \leq G_\gamma(\widehat{u}_A) - G_\gamma(u^*) - \langle Au^* - f, u^* - \widehat{u}_A\rangle, \tag{19}$$

so we need to focus on estimating the right-hand side in an appropriate way. First let us recall (from Section 3.1.3) that we denote by $\hat{u}$ a solution of ($\widehat{\mathsf{P}}_{\mathrm{gen}}$), so that

$$\widehat{G}_\gamma(\hat{u}) \leq \widehat{G}_\gamma(u) \quad \forall u \in \mathcal{X}^{\mathrm{s}}. \tag{20}$$

Define $H_\gamma \colon V \to \mathbb{R}$ by

$$H_\gamma(u) := \max_{v \in \mathcal{X}^{\mathrm{t}}} L_\gamma(u, v) + R_1(u) - R_2(v),$$

which can be thought of as the continuous version of $\widehat{G}_\gamma$ (existence can be shown in a similar way to Proposition 3.7, utilising the continuity of $L_\gamma + R_1 - R_2$). We begin with the following decomposition of the right-hand side of (19): for arbitrary $\bar{u} \in \mathcal{X}^{\mathrm{s}}$, we have

$$\begin{aligned} G_\gamma(\widehat{u}_A) - G_\gamma(u^*) \leq{}& \underbrace{G_\gamma(\widehat{u}_A) - H_\gamma(\widehat{u}_A)}_{I} + \underbrace{H_\gamma(\widehat{u}_A) - \widehat{G}_\gamma(\widehat{u}_A) + \widehat{G}_\gamma(\bar{u}) - H_\gamma(\bar{u})}_{II} \\ &+ \underbrace{\widehat{G}_\gamma(\widehat{u}_A) - \widehat{G}_\gamma(\hat{u})}_{III} + \underbrace{H_\gamma(\bar{u}) - H_\gamma(u^*)}_{IV} + \underbrace{H_\gamma(u^*) - G_\gamma(u^*)}_{V} \end{aligned}$$

where we used $\widehat{G}_\gamma(\hat{u}) \leq \widehat{G}_\gamma(\bar{u})$ by (20). We need the following two lemmas, wherein $\mathfrak{R} \colon V^* \to V$ is the Riesz map.

Lemma 3.9. *For all $u, v \in V$,*

$$L_\gamma(u, z) - L_\gamma(u, v) = \frac{1}{2\gamma} \|v - z\|_V^2 + \left(\mathfrak{R}(Au - f) + \frac{z - u}{\gamma}, v - z\right)_V. \tag{21}$$

Proof. This follows easily from the definition of L_γ and the three-point identity

$$\|v - u\|^2 - \|z - u\|^2 = \|v - z\|^2 + 2(z - u, v - z).$$

□

Lemma 3.10. *For all $u, v \in H^1(\Omega)$, we have for $i = 1, 2$ the estimate*

$$\begin{aligned} R_i(u) - R_i(v) \leq{}& 2w_{b_i}(\|h\|_{L^2(\partial\Omega)} + \|u\|_{L^2(\partial\Omega)}) \|u - v\|_{L^2(\partial\Omega)} \\ &+ w_{o_i}(\|u + v\|_{L^2(\Omega)} + 2\|\psi\|_{L^2(\Omega)}) \|u - v\|_{L^2(\Omega)} \\ &- w_{b_i} \|u - v\|_{L^2(\partial\Omega)}^2. \end{aligned} \tag{22}$$

Proof. Let us just set $i = 1$. We start with

$$R_1(u) - R_1(v) = w_{b_1} \int_{\partial\Omega} (u - h)^2 - (v - h)^2 + w_{o_1} \int_\Omega |(\psi - u)^+|^2 - |(\psi - v)^+|^2.$$

By straightforward computations and the trace inequality [59, Theorem 6.7, Chapter 1], we estimate

$$\begin{aligned}\int_{\partial\Omega}(u-h)^2-(v-h)^2 &= \int_{\partial\Omega}2(u-h)(u-v)-|u-v|^2\\ &\leq 2(\|h\|_{L^2(\partial\Omega)}+\|u\|_{L^2(\partial\Omega)})\|u-v\|_{L^2(\partial\Omega)}-\|u-v\|^2_{L^2(\partial\Omega)},\end{aligned}$$

while for the second integral, using the fact that $(\cdot)^+$ is Lipschitz, we get

$$\begin{aligned}\int_\Omega|(u-\psi)^+|^2-|(v-\psi)^+|^2 &\leq \int_\Omega|u-v||(\psi-u)^+ + (\psi-v)^+)|\\ &\leq (\|u+v\|_{L^2(\Omega)}+2\|\psi\|_{L^2(\Omega)})\|u-v\|_{L^2(\Omega)}.\end{aligned}$$

□

From now on we assume

$$\begin{aligned}&\exists M^{\mathrm{s}},B^{\mathrm{s}}>0: \quad \|u\|_V\leq M^{\mathrm{s}} \quad\text{and}\quad \|u\|_{L^2(\partial\Omega)}\leq B^{\mathrm{s}} \quad \forall u\in\mathcal{X}^{\mathrm{s}},\\ &\exists M^{\mathrm{t}},B^{\mathrm{t}}>0: \quad \|u\|_V\leq M^{\mathrm{t}} \quad\text{and}\quad \|u\|_{L^2(\partial\Omega)}\leq B^{\mathrm{t}} \quad \forall u\in\mathcal{X}^{\mathrm{t}}.\end{aligned} \tag{23}$$

Proposition 3.11 (Estimate on I)**.** *We have*

$$\begin{aligned}&G_\gamma(\widehat{u}_A)-H_\gamma(\widehat{u}_A)\\ &\leq \max_{v\in K\cap Y}\min_{w\in\mathcal{X}^{\mathrm{t}}}\Big(K_1\|w-v\|_V+K_2\|w-v\|_{L^2(\Omega)}+K_3\|w-v\|_{L^2(\partial\Omega)}-w_{b_2}\|w-v\|^2_{L^2(\partial\Omega)}\\ &\quad-\frac{1}{2\gamma}\|w-v\|_V^2\Big)-R_1(\widehat{u}_A)\end{aligned}$$

where

$$K_1:=C_bM^{\mathrm{s}}+\|f\|_{V^*}+\frac{1}{\gamma}(M^{\mathrm{s}}+M^{\mathrm{t}}),\quad K_2:=w_{o_2}(r+M^{\mathrm{t}}+2\|\psi\|_{L^2(\Omega)}),\quad K_3:=2w_{b_2}(\|h\|_{L^2(\partial\Omega)}+B^{\mathrm{t}}).$$

Proof. We begin with

$$\begin{aligned}G_\gamma(\widehat{u}_A)-H_\gamma(\widehat{u}_A) &= -R_1(\widehat{u}_A)+\max_{v\in K\cap Y}L_\gamma(\widehat{u}_A,v)-\max_{w\in\mathcal{X}^{\mathrm{t}}}(L_\gamma(\widehat{u}_A,w)-R_2(w))\\ &= -R_1(\widehat{u}_A)+\max_{v\in K\cap Y}\min_{w\in\mathcal{X}^{\mathrm{t}}}(L_\gamma(\widehat{u}_A,v)-L_\gamma(\widehat{u}_A,w)+R_2(w)).\end{aligned}$$

Now from (21), we can estimate the middle two terms as

$$\begin{aligned}L_\gamma(\widehat{u}_A,v)-L_\gamma(\widehat{u}_A,w) &= \frac{1}{2\gamma}\|w-v\|^2+(\mathfrak{R}(A\widehat{u}_A-f)+\frac{v-\widehat{u}_A}{\gamma},w-v)_V\\ &\leq \frac{1}{2\gamma}\|w-v\|^2+(\|A\widehat{u}_A-f\|_{V^*}+\frac{1}{\gamma}\|w-\widehat{u}_A\|_V)\|w-v\|_V-\frac{1}{\gamma}\|w-v\|_V^2\\ &\leq -\frac{1}{2\gamma}\|w-v\|^2+(C_b\|\widehat{u}_A\|_V+\|f\|_{V^*}+\frac{1}{\gamma}(M^{\mathrm{t}}+\|\widehat{u}_A\|_V))\|w-v\|_V.\end{aligned}$$

Plugging this in, writing $R_2(w)=R_2(w)-R_2(v)$ and using the estimate on R_2 from (22), we get the result. □

Proposition 3.12 (Estimate on II)**.** *For all $\bar{u} \in \mathcal{X}^{\mathrm{s}}$, we have*

$$\begin{aligned} &H_\gamma(\widehat{u}_A) - \widehat{G}_\gamma(\widehat{u}_A) + \widehat{G}_\gamma(\bar{u}) - H_\gamma(\bar{u}) \\ &\quad \le 2 \sup_{\substack{u \in \mathcal{X}^{\mathrm{s}} \\ v \in \mathcal{X}^{\mathrm{t}}}} |L_\gamma(u,v) + R_1(u) - R_2(v) - (\hat{L}_\gamma(u,v) + \hat{R}_1(u) - \hat{R}_2(v))|. \end{aligned}$$

Proof. If we set

$$\begin{aligned} &\hat{v} \in \operatorname*{arg\,max}_{v \in \mathcal{X}^{\mathrm{t}}} \big(L_\gamma(\widehat{u}_A, v) + R_1(\widehat{u}_A) - R_2(v)\big), \qquad && \tilde{v} \in \operatorname*{arg\,max}_{v \in \mathcal{X}^{\mathrm{t}}} \big(\hat{L}_\gamma(\widehat{u}_A, v) + \hat{R}_1(\widehat{u}_A) - \hat{R}_2(v)\big), \\ &\bar{v} \in \operatorname*{arg\,max}_{v \in \mathcal{X}^{\mathrm{t}}} \big(\hat{L}_\gamma(\bar{u}, v) + \hat{R}_1(\bar{u}) - \hat{R}_2(v)\big), && v^* \in \operatorname*{arg\,max}_{v \in \mathcal{X}^{\mathrm{t}}} \big(L_\gamma(\bar{u}, v) + R_1(\bar{u}) - R_2(v)\big), \end{aligned}$$

the left-hand side of the desired inequality equals

$$\begin{aligned} H_\gamma(\widehat{u}_A) - \widehat{G}_\gamma(\widehat{u}_A) + \widehat{G}_\gamma(\bar{u}) - H_\gamma(\bar{u}) = {} & L_\gamma(\widehat{u}_A, \hat{v}) + R_1(\widehat{u}_A) - R_2(\hat{v}) - (\hat{L}_\gamma(\widehat{u}_A, \tilde{v}) + \hat{R}_1(\widehat{u}_A) - \hat{R}_2(\tilde{v})) \\ & + \hat{L}_\gamma(\bar{u}, \bar{v}) + \hat{R}_1(\bar{u}) - \hat{R}_2(\bar{v}) - (L_\gamma(\bar{u}, v^*) + R_1(\bar{u}) - R_2(v^*)). \end{aligned}$$

Now since $\tilde{v} \in \mathcal{X}^{\mathrm{t}}$ is the maximiser and $\hat{v}$ belongs to $\mathcal{X}^{\mathrm{t}}$, and similarly with v^* and $\bar{v}$, we obtain

$$\begin{aligned} \hat{L}_\gamma(\widehat{u}_A, \tilde{v}) + \hat{R}_1(\widehat{u}_A) - \hat{R}_2(\tilde{v}) &\ge \hat{L}_\gamma(\widehat{u}_A, \hat{v}) + \hat{R}_1(\widehat{u}_A) - \hat{R}_2(\hat{v}), \\ L_\gamma(\bar{u}, v^*) + R_1(\bar{u}) - R_2(v^*) &\ge L_\gamma(\bar{u}, \bar{v}) + R_1(\bar{u}) - R_2(\bar{v}), \end{aligned}$$

which we can insert above to find

$$\begin{aligned} H_\gamma(\widehat{u}_A) - \widehat{G}_\gamma(\widehat{u}_A) + \widehat{G}_\gamma(\bar{u}) - H_\gamma(\bar{u}) \le {} & L_\gamma(\widehat{u}_A, \hat{v}) + R_1(\widehat{u}_A) - R_2(\hat{v}) - (\hat{L}_\gamma(\widehat{u}_A, \hat{v}) + \hat{R}_1(\widehat{u}_A) - \hat{R}_2(\hat{v})) \\ & + \hat{L}_\gamma(\bar{u}, \bar{v}) + \hat{R}_1(\bar{u}) - \hat{R}_2(\bar{v}) - (L_\gamma(\bar{u}, \bar{v}) + R_1(\bar{u}) - R_2(\bar{v})). \end{aligned}$$

□

Now for the fourth difference the following estimate holds true.

Proposition 3.13 (Estimate on IV)**.** *For all $\bar{u} \in \mathcal{X}^{\mathrm{s}}$, we have*

$$\begin{aligned} H_\gamma(\bar{u}) - H_\gamma(u^*) \le {} & K_4 \|\bar{u} - u^*\|_V + K_5 \|\bar{u} - u^*\|_{L^2(\Omega)} + K_6 \|\bar{u} - u^*\|_{L^2(\partial\Omega)} - w_{b_1} \|\bar{u} - u^*\|^2_{L^2(\partial\Omega)} \\ & - \frac{1}{2\gamma} \|u^* - \bar{u}\|^2_V \end{aligned}$$

where

$$\begin{aligned} &K_4 := C_b(M^{\mathrm{s}} + M^{\mathrm{t}}) + C_1 + \frac{1}{\gamma}(\|u^*\|_V + M^{\mathrm{t}}), \quad K_5 := w_{o_1}(M^{\mathrm{s}} + \|u^*\|_V + 2\|\psi\|_{L^2(\Omega)}), \\ &K_6 := w_{b_1} 2(\|h\|_{L^2(\partial\Omega)} + B^{\mathrm{s}}). \end{aligned}$$

Proof. The left-hand side satisfies

$$H_\gamma(\bar{u}) - H_\gamma(u^*) \le \max_{v \in \mathcal{X}^{\mathrm{t}}} \left(L_\gamma(\bar{u}, v) + R_1(\bar{u}) - L_\gamma(u^*, v) - R_1(u^*)\right).$$

We have

$$L_\gamma(\bar{u}, v) - L_\gamma(u^*, v) = L(\bar{u}, v) - L(u^*, v) + \frac{1}{2\gamma}\left(\|u^* - v\|^2_V - \|\bar{u} - v\|^2_V\right).$$

Firstly, let us estimate the first two terms on the right-hand side:

$$L(\bar{u},v)-L(u^*,v)=\langle A\bar{u}-f,\bar{u}-v\rangle-\langle Au^*-f,u^*-v\rangle=\langle A(\bar{u}-u^*),\bar{u}-v\rangle+\langle Au^*-f,\bar{u}-u^*\rangle$$
$$\leq C_b(M^{\mathrm{s}}+M^{\mathrm{t}})\left\|\bar{u}-u^*\right\|_V+C_1\left\|\bar{u}-u^*\right\|_V.$$

Now for the squared norm terms, we again use the three-point equality to obtain

$$\left\|u^*-v\right\|_V^2-\left\|\bar{u}-v\right\|_V^2=2(u^*-v,u^*-\bar{u})-\left\|u^*-\bar{u}\right\|_V^2\leq 2(\left\|u^*\right\|_V+M^{\mathrm{t}})\left\|u^*-\bar{u}\right\|_V-\left\|u^*-\bar{u}\right\|_V^2.$$

Putting everything together and using again the estimate (22), we obtain

$$\begin{aligned}&H_\gamma(\bar{u})-H_\gamma(u^*)\\&\quad\leq C_b(M^{\mathrm{s}}+M^{\mathrm{t}})\left\|\bar{u}-u^*\right\|_V+C_1\left\|\bar{u}-u^*\right\|_V+\frac{1}{\gamma}(\left\|u^*\right\|_V+M^{\mathrm{t}})\left\|u^*-\bar{u}\right\|_V-\frac{1}{2\gamma}\left\|u^*-\bar{u}\right\|_V^2\\&\qquad+w_{b_1}2(\left\|h\right\|_{L^2(\partial\Omega)}+\left\|\bar{u}\right\|_{L^2(\partial\Omega)})\left\|\bar{u}-u^*\right\|_{L^2(\partial\Omega)}+w_{o_1}(M^{\mathrm{s}}+\left\|u^*\right\|_V+2\left\|\psi\right\|_{L^2(\Omega)})\left\|\bar{u}-u^*\right\|_{L^2(\Omega)}\\&\qquad-w_{b_1}\left\|\bar{u}-u^*\right\|_{L^2(\partial\Omega)}^2.\end{aligned}$$

□

Combining the previous results, we have shown

$$\begin{aligned}\left(C_a-\frac{1}{2\gamma}\right)\left\|\widehat{u}_A-u^*\right\|_V^2\leq{}&\max_{v\in K\cap Y}\min_{w\in\mathcal{X}^{\mathrm{t}}}K_1\left\|w-v\right\|_V+K_2\left\|w-v\right\|_{L^2(\Omega)}+K_3\left\|w-v\right\|_{L^2(\partial\Omega)}\\&+\inf_{u\in\mathcal{X}^{\mathrm{s}}}\left(K_4\left\|u-u^*\right\|_V+K_5\left\|u-u^*\right\|_{L^2(\Omega)}+K_6\left\|u-u^*\right\|_{L^2(\partial\Omega)}\right)\\&+2\sup_{\substack{u\in\mathcal{X}^{\mathrm{s}}\\v\in\mathcal{X}^{\mathrm{t}}}}|L_\gamma(u,v)+R_2(u)+R_1(v)-(\hat{L}_\gamma(u,v)+\hat{R}_2(u)+\hat{R}_1(v))|\\&+\widehat{G}_\gamma(\widehat{u}_A)-\widehat{G}_\gamma(\hat{u})\\&+H_\gamma(u^*)-G_\gamma(u^*)+\langle Au^*-f,u^*-\widehat{u}_A\rangle-R_1(\widehat{u}_A).\end{aligned}\tag{24}$$

The quantity comprising the first two lines on the right-hand side above

$$\begin{aligned}\xi_{\mathrm{app}}:={}&\max_{v\in K\cap Y}\min_{w\in\mathcal{X}^{\mathrm{t}}}K_1\left\|w-v\right\|_V+K_2\left\|w-v\right\|_{L^2(\Omega)}+K_3\left\|w-v\right\|_{L^2(\partial\Omega)}\\&+\inf_{u\in\mathcal{X}^{\mathrm{s}}}\left(K_4\left\|u-u^*\right\|_V+K_5\left\|u-u^*\right\|_{L^2(\Omega)}+K_6\left\|u-u^*\right\|_{L^2(\partial\Omega)}\right)\end{aligned}$$

is known as the *approximation error*. It is related to how well the spaces $\mathcal{X}^{\mathrm{s}}$ and $\mathcal{X}^{\mathrm{t}}$ approximate u^* and the set $K\cap Y$ respectively. The *statistical error* is the quantity

$$\xi_{\mathrm{stat}}:=\sup_{\substack{u\in\mathcal{X}^{\mathrm{s}}\\v\in\mathcal{X}^{\mathrm{t}}}}|L_\gamma(u,v)+R_2(u)+R_1(v)-(\hat{L}_\gamma(u,v)+\hat{R}_2(u)+\hat{R}_1(v))|,$$

i.e., it is a measure of the error arising from the numerical approximation of the integral. We shall estimate both of these in the next subsections. The term

$$\xi_{\mathrm{opt}}=\widehat{G}_\gamma(\widehat{u}_A)-\widehat{G}_\gamma(\hat{u})$$

is called the *optimisation error*; as explained in Section 3.1.4, this is not something we shall explore in this paper. Finally, the two terms $\langle Au^*-f,u^*-\widehat{u}_A\rangle-R_1(\widehat{u}_A)$ and $H_\gamma(u^*)-G_\gamma(u^*)$ essentially arise due to the presence of K and the inexactness of our approximating spaces. Let us deal with these terms first.

3.2.1 Bounds on the error arising from the constraint set

Proposition 3.14. *If $\widehat{u}_A|_{\partial\Omega} = h$ (which is the case in the situation of ($\widehat{\mathrm{P}}_{\mathrm{bc}}$)) and $u^* \in H^2(\Omega)$, then for every $\epsilon > 0$, if the penalty weight w_{o_1} satisfies $w_{o_1} \geq \frac{1}{4\epsilon} \|Au^* - f\|^2_{L^2(\Omega)}$, we have*

$$\langle Au^* - f, u^* - \widehat{u}_A\rangle - R_1(\widehat{u}_A) \leq \epsilon.$$

Proof. We have that $\max(\psi, \widehat{u}_A)$ belongs to K (since $\widehat{u}_A|_{\partial\Omega} = h$ and $\psi|_{\partial\Omega} \leq h$). Thus, using Young's inequality with $\epsilon > 0$ and noting that $\max(\psi, \widehat{u}_A) - \widehat{u}_A = (\psi - \widehat{u}_A)^+$,

$$\begin{aligned}\langle Au^* - f, u^* - \widehat{u}_A\rangle &= \langle Au^* - f, u^* - \max(\psi, \widehat{u}_A)\rangle + \langle Au^* - f, \max(\psi, \widehat{u}_A) - \widehat{u}_A\rangle \\ &\leq \|Au^* - f\|_{L^2(\Omega)} \left\|(\psi - \widehat{u}_A)^+\right\|_{L^2(\Omega)} \\ &\leq \epsilon + \frac{1}{4\epsilon} \|Au^* - f\|^2_{L^2(\Omega)} \left\|(\psi - \widehat{u}_A)^+\right\|^2_{L^2(\Omega)}.\end{aligned}$$

The claim then follows from

$$\langle Au^* - f, u^* - \widehat{u}_A\rangle - R_1(\widehat{u}_A) \leq \epsilon + \left(\frac{1}{4\epsilon} \|Au^* - f\|^2_{L^2(\Omega)} - w_{o_1}\right) \left\|(\psi - \widehat{u}_A)^+\right\|^2_{L^2(\Omega)}.$$

□

Proposition 3.15 (Estimate on V)**.** *For every $\epsilon > 0$, there exists an $\alpha_0 > 0$ such that if the penalty parameters w_{o_2}, w_{b_2} satisfy $\min(w_{o_2}, w_{b_2}) \geq \alpha_0$, then*

$$H_\gamma(u^*) - G_\gamma(u^*) \leq \epsilon - \frac{1}{2\gamma} \|u^* - v^*\|^2_V - R_2(v^*),$$

where $v^* \in \arg\max_{w\in H^1(\Omega)} L_\gamma(u^*, w) - R_2(w)$.

Proof. Since $G_\gamma(u^*) = 0$ and $\mathcal{X}^{\mathrm{t}} \subset H^1(\Omega)$, we have, setting $v^* \in \arg\max_{w\in H^1(\Omega)} L_\gamma(u^*, w) - R_2(w)$,

$$\begin{aligned}H_\gamma(u^*) - G_\gamma(u^*) &= \max_{w\in\mathcal{X}^{\mathrm{t}}} L_\gamma(u^*, w) - R_2(w) \leq \max_{w\in H^1(\Omega)} L_\gamma(u^*, w) - R_2(w) \\ &= \langle Au^* - f, u^* - v^*\rangle - \frac{1}{2\gamma} \|u^* - v^*\|^2_V - R_2(v^*) \\ &\leq \langle Au^* - f, P_K(v^*) - v^*\rangle - \frac{1}{2\gamma} \|u^* - v^*\|^2_V - R_2(v^*) \\ &\leq \|Au^* - f\|_{V^*} \|P_K(v^*) - v^*\|_V - \frac{1}{2\gamma} \|u^* - v^*\|^2_V - R_2(v^*). \qquad (25)\end{aligned}$$

Now, in fact $v^* = v_\alpha$ where α represents the penalty parameters in R_2. We will now show that v_α converges strongly in V to some $\hat{v} \in K$ as $\alpha \to \infty$. If this were the case, then using $\hat{v} = P_K(\hat{v})$, the right-hand side of the estimate

$$\|P_K(v_\alpha) - v_\alpha\|_V \leq \|P_K(v_\alpha) - P_K(\hat{v})\|_V + \|\hat{v} - v_\alpha\|_V \leq 2\|\hat{v} - v_\alpha\|_V$$

can be made arbitrarily small if we take α large enough. This would then allow us to control the first term on the right-hand side of the (25), leading to the desired result. Writing occasionally R_2^α instead of R_2 to emphasise the dependence of R_2 on α and setting

$$J_\alpha(w) := R_2^\alpha(w) - L_\gamma(u^*, w),$$

we have by definition of v_α that $J_\alpha(v_\alpha) \leq J_\alpha(w)$ for all $w \in H^1(\Omega)$; taking $w = u^* \in K$ and manipulating gives

$$R_2^\alpha(v_\alpha) + \frac{1}{4\gamma} \|u^* - v_\alpha\|_V^2 \leq \gamma \|Au^* - f\|_{V^*}^2 . \tag{26}$$

This uniform bound implies the existence of $\hat{v} \in V$ such that, for a subsequence (that we relabelled), $v_\alpha \rightharpoonup \hat{v}$ in V and strongly in $L^2(\Omega)$. Since $R_2^\alpha(v_\alpha) = w_{o_2} \int_\Omega |(\psi - v_\alpha)^+|^2 + w_{b_2} \int_{\partial\Omega} (v_\alpha - h)^2$, if we divide the above inequality by w_{o_2}, disregard the boundary term and vice versa with w_{b_2}, we find $\int_\Omega |(\psi - v_\alpha)^+|^2 + \int_{\partial\Omega}(v_\alpha - h)^2 \to 0$ in the limit $w_{b_2}, w_{o_2} \to \infty$. Using the convergence result for v_α and the fact that the trace operator is linear and continuous as well as the weak lower semicontinuity of the functionals involved, we then get $\hat{v} \in K$. Now let $J(w) := -L_\gamma(u^*, w)$. We have $J_\alpha(w) \geq J(w)$ and so, using weak lower semicontinuity,

$$\begin{aligned}\liminf_{\alpha\to\infty} J_\alpha(v_\alpha) \geq \liminf_{\alpha\to\infty} J(v_\alpha) &= \liminf_{\alpha\to\infty} \frac{1}{2\gamma} \|u^* - v_\alpha\|_V^2 - \langle Au^* - f, u^* - v_\alpha\rangle \\ &\geq \frac{1}{2\gamma} \|u^* - \hat{v}\|_V^2 - \langle Au^* - f, u^* - \hat{v}\rangle = J(\hat{v}).\end{aligned}$$

On the other hand, since $J_\alpha(v_\alpha) \leq J_\alpha(\hat{v})$ and using the fact that $\hat{v} \in K$,

$$\liminf_{\alpha\to\infty} J_\alpha(v_\alpha) \leq \liminf_{\alpha\to\infty} J_\alpha(\hat{v}) = \liminf_{\alpha\to\infty} -L_\gamma(u^*, \hat{v}) = J(\hat{v}).$$

Hence, $\liminf_{\alpha\to\infty} J(v_\alpha) = J(\hat{v})$ and therefore for a subsequence (relabelled), we get $J(v_\alpha) \to J(\hat{v})$. That is,

$$\frac{1}{2\gamma} \|u^* - v_\alpha\|_V^2 - L(u^*, v_\alpha) \to \frac{1}{2\gamma} \|u^* - \hat{v}\|_V^2 - L(u^*, \hat{v}),$$

and since $L(u^*, v_\alpha) \to L(u^*, \hat{v})$, it follows that $\|u^* - v_\alpha\|_V \to \|u^* - \hat{v}\|_V$. The weak convergence $u^* - v_\alpha \rightharpoonup u^* - \hat{v}$ together with the above norm convergence then yields $v_\alpha \to \hat{v}$ in $H^1(\Omega)$. Now we have $J_\alpha(v_\alpha) \leq J_\alpha(k) = J(k)$ for all $k \in K$, since $K \subset H^1(\Omega)$ and v_α is the minimiser. Taking the limit inferior in this inequality and using the first chain of inequalities above, $J(\hat{v}) \leq J(k)$ for all $k \in K$, i.e., $\hat{v}$ is the minimiser of J over K (which is uniquely determined). The subsequence principle then tells us that the entire sequence $\{v_\alpha\}$ converges. □

The proof reveals an error rate. Indeed, we see from (26) and the definition of R_2 that

$$\left\|(\psi - v_\alpha)^+\right\|_{L^2(\Omega)} = \mathcal{O}(1/\sqrt{\alpha}) \qquad \text{and} \qquad \|v_\alpha - h\|_{L^2(\partial\Omega)} = \mathcal{O}(1/\sqrt{\alpha}).$$

3.2.2 Bounds on the approximation error

To control the approximation error we need to ensure that the two network spaces $\mathcal{X}^{\mathrm{s}}$ and $\mathcal{X}^{\mathrm{t}}$ are sufficiently rich. This relates to universal approximation theorems for neural networks. The strength of the approximation results strongly depends on the norm used, and some care is required in ensuring that one can apply them to ξ_{app} since we need uniformity in $K \cap Y$ in order to bound the maxmin term there. In order to keep our theory as general as reasonably possible and bearing in mind that approximation results for different architectures are rapidly evolving, we make the next assumption and provide a scenario (see Proposition 3.18) in which it is satisfied.

Assumption 3.16 (Assumptions on the density of neural network spaces in $H^1(\Omega)$)**.**

(i) There exists a neural network architecture $\mathcal{F}^{\mathrm{s}}$ with activation function σ^{s} satisfying: for every $w \in H^2(\Omega)$ and every $\epsilon > 0$, there exists $\mathfrak{d}^{\mathrm{s}}, \mathfrak{w}^{\mathrm{s}} \in \mathbb{N}$ and $u \in \mathcal{F}^{\mathrm{s}}(\mathfrak{d}^{\mathrm{s}}, \mathfrak{w}^{\mathrm{s}}, \sigma^{\mathrm{s}})$ such that

$$\|u - w\|_V \leq \epsilon.$$

(ii) There exists a neural network architecture $\mathcal{F}^{\mathrm{t}}$ with activation function σ^{t} satisfying: for every $k \in Y$ and every $\epsilon > 0$, there exist $\mathfrak{d}^{\mathrm{t}}, \mathfrak{w}^{\mathrm{t}} \in \mathbb{N}$ independent of k and $v \in \mathcal{F}^{\mathrm{t}}(\mathfrak{d}^{\mathrm{t}}, \mathfrak{w}^{\mathrm{t}}, \sigma^{\mathrm{t}})$ such that

$$\|v - k\|_V \leq \epsilon.$$

We also need the space Y to be such that $K \cap Y$ is bounded in V:

Assumption 3.17. *Assume that there exists $\bar{R} > 0$ such that $\|v\|_V \leq \bar{R}$ for all $v \in K \cap Y$.*

In the next proposition, we show that these assumptions can be satisfied. To keep matters simple we focus only on the two activation functions ReLU and $\tanh$; others are possible too.

Proposition 3.18. *Let $h \in H^{3/2}(\partial\Omega)$, $\psi \in H^2(\Omega)$, and let Ω be convex or $C^{1,1}$. Let $\tilde{R}$ be a constant satisfying*

$$\tilde{R} \geq \|u^*\|_V,$$

set $Y := B^{H^2(\Omega)}(\tilde{R})$, the closed ball in $H^2(\Omega)$ with center 0 and radius $\tilde{R}$, and choose $\mathcal{F}^{\mathrm{t}}$ and $\mathcal{F}^{\mathrm{s}}$ as $\mathcal{F}_{\mathrm{FFN}}$ with ReLU or $\tanh$ as the activation. Then Assumption 3.16 and Assumption 3.17 are satisfied

Proof. Proposition 4.8 of [21] implies the following: let $\mu > 0$ be arbitrary; then there exist constants $\mathfrak{d}^{\mathrm{t}}, C, \theta$ and $\bar{\epsilon}$ such that for every $\epsilon < \bar{\epsilon}$ and every $k \in B^{H^2(\Omega)}(1)$, there exists a neural network $v \in \mathcal{F}_{\mathrm{FFN}}$ with at most $\mathfrak{d}^{\mathrm{t}}$ layers and at most

$$\begin{cases} C\epsilon^{-n/(1-\mu)} & : \text{if activation is } \tanh, \\ C\epsilon^{-n} & : \text{if activation is } \mathrm{ReLU}, \end{cases}$$

non-zero weights bounded in absolute value by $C\epsilon^{-\theta}$ satisfying $\|v - k\|_{H^1(\Omega)} \leq \epsilon$. Now if $g \in B^{H^2(\Omega)}(R)$, the above gives us a neural network $\tilde{v}$ with $\|\tilde{v} - g/R\|_{H^1(\Omega)} \leq \epsilon/R$ as long as $\epsilon \leq R\bar{\epsilon}$. Setting $v := R\tilde{v}$, we get $\|v - g\|_{H^1(\Omega)} \leq \epsilon$ where $v \in \mathcal{F}_{\mathrm{FFN}}$ is also a neural network (with the same architecture as $\tilde{v}$ except the final layer has been scaled by R) with at most

$$\begin{cases} C(R^{-1}\epsilon)^{-n/(1-\mu)} & : \text{if activation is } \tanh, \\ C(R^{-1}\epsilon)^{-n} & : \text{if activation is } \mathrm{ReLU}, \end{cases}$$

non-zero weights.

The key point that gives us the existence of the neural network architecture that is independent of the particular element k that is being approximated is the following: if a network with an arbitrary number of neurons and layers has at most M non-zero weights, it can be viewed as a network with M non-zero weights and at most $M + 1$ neurons and layers. Since the above gives us at most $C(R^{-1}\epsilon)^{-n/(1-\mu)}$ or $C(R^{-1}\epsilon)^{-n}$ non-zero weights and this number is independent of k, it can be viewed as a neural network in $\mathcal{F}_{\mathrm{FFN}}$ with a fixed (independent of k) size.

Hence Assumption 3.16 (ii) holds if we choose $Y = B^{H^2(\Omega)}(\tilde{R})$ and $\mathcal{F}^{\mathrm{t}} = \mathcal{F}_{\mathrm{FFN}}$ as the standard feedforward neural network architecture. This is a valid choice of Y (i.e., it satisfies Assumption 2.5 by

the assumptions on the domain and the data, see Proposition 2.1). Assumption 3.16 (i) also holds by the above argument since $u^* \in H^2(\Omega)$.

As $Y = B^{H^2(\Omega)}(\tilde{R})$, Assumption 3.17 follows from the continuous embedding $H^2(\Omega) \hookrightarrow V$ with $\bar{R} = \tilde{R}$. □

Since in our implementation we use the special ResNet structure of [16] for which approximation results still appear to be incomplete, we shall postpone the study of approximation theorems for our specific architecture to a later work. For approximation results involving the standard ResNet structure, we refer to e.g. [2, 45, 43, 44]; note that networks with one hidden layer possess universal approximation power in L^1 for ResNets [44].

Now, take $P \geq 1 + \|u^*\|_V$ and $Q \geq 1 + \bar{R}$ and fix $\mathcal{X}^{\mathrm{s}} := \mathcal{F}^{\mathrm{s}} \cap B^{H^1(\Omega)}(P)$ and $\mathcal{X}^{\mathrm{t}} := \mathcal{F}^{\mathrm{t}} \cap B^{H^1(\Omega)}(Q)$, where $\mathcal{F}^{\mathrm{s}}$ and $\mathcal{F}^{\mathrm{t}}$ are as in Assumption 3.16. Due to this, if we set $R^* := \max(P, Q, C_T P, C_T Q)$, we get that the constants in (23) satisfy $\max(M^{\mathrm{s}}, M^{\mathrm{t}}, B^{\mathrm{s}}, B^{\mathrm{t}}) \leq R^*$. Observe that all the constants K_i depend on R^*. Define $\kappa_1 := \max(1, K_4 + K_5 + C_T K_6)$. Under Assumption 3.16 (i), for every $\epsilon \in (0, 1)$, there exists $\tilde{u} \in \mathcal{X}^{\mathrm{s}}$ such that

$$\|\tilde{u} - u^*\|_V \leq \frac{\epsilon}{3\kappa_1}. \tag{27}$$

Note further that since $\kappa_1 \geq 1$ and $\tilde{u} \in \mathcal{X}^{\mathrm{s}}$,

$$\|\tilde{u}\|_V \leq \|\tilde{u} - u^*\|_V + \|u^*\|_V \leq \frac{\epsilon}{3\kappa_1} + \|u^*\|_V \leq 1 + \|u^*\|_V \leq P.$$

In a similar way, let $\kappa_2 := \max(1, K_1 + K_2 + C_T K_3)$. Under Assumption 3.16 (ii), for $\epsilon \in (0, 1)$, we can show

$$\max_{k \in K \cap Y} \min_{v \in \mathcal{X}^{\mathrm{t}}} \|k - v\|_V \leq \frac{\epsilon}{3(K_1 + K_2 + C_T K_3)}. \tag{28}$$

Indeed, fixing an arbitrary $k \in K \cap Y$, by Assumption 3.16 (ii) we get the existence of $\mathfrak{d}^{\mathrm{t}}, \mathfrak{w}^{\mathrm{t}} \in \mathbb{N}$ *uniform in* k and $v \in \mathcal{F}(\mathfrak{d}^{\mathrm{t}}, \mathfrak{w}^{\mathrm{t}}, \sigma^{\mathrm{t}})$ with $\|k - v\|_V \leq \epsilon/(3\kappa_2)$. This implies, because $\kappa_2 \geq 1$ and by the bound on k from Assumption 3.17,

$$\|v\|_V \leq \|v - k\|_V + \|k\|_V \leq \frac{\epsilon}{3\kappa_2} + \bar{R} \leq 1 + \bar{R} \leq Q.$$

This shows that $v \in B^{H^1(\Omega)}(Q)$ and thus $v \in \mathcal{X}^{\mathrm{t}}$. Then (28) follows from

$$(K_1 + K_2 + C_T K_3) \|k - v\|_V \leq \kappa_2 \|k - v\|_V \leq \frac{\epsilon}{3}.$$

Theorem 3.19. *Let Assumption 3.16 and Assumption 3.17 hold. For any $\epsilon \geq 0$, there exist architectures $\mathcal{F}^{\mathrm{s}}, \mathcal{F}^{\mathrm{t}}$, activations $\sigma^{\mathrm{s}}, \sigma^{\mathrm{t}}$, numbers $\mathfrak{d}^{\mathrm{s}}, \mathfrak{w}^{\mathrm{s}}, \mathfrak{d}^{\mathrm{t}}, \mathfrak{w}^{\mathrm{t}} \in \mathbb{N}$ such that, with $\mathcal{X}^{\mathrm{s}} = \mathcal{F}^{\mathrm{s}}(\mathfrak{d}^{\mathrm{s}}, \mathfrak{w}^{\mathrm{s}}, \sigma^{\mathrm{s}})$ and $\mathcal{X}^{\mathrm{t}} = \mathcal{F}^{\mathrm{t}}(\mathfrak{d}^{\mathrm{t}}, \mathfrak{w}^{\mathrm{t}}, \sigma^{\mathrm{t}})$, we have*

$$\xi_{\mathrm{app}} \leq \epsilon.$$

Proof. Fix $\epsilon > 0$. We have from the estimate (24) and making use of (27) and (28),

$$\xi_{\mathrm{app}} \leq \max_{k \in K \cap Y} \min_{v \in \mathcal{X}^{\mathrm{t}}} (K_1 + K_2 + C_T K_3) \|k - v\|_V + \inf_{u \in \mathcal{X}^{\mathrm{s}}} (K_4 + K_5 + C_T K_6) \|\bar{u} - u^*\|_V \leq \epsilon.$$

□

At this point let us remark that if we had $\widehat{u}_A \in K$ the above error analysis could have been greatly simplified.

3.2.3 The $(\widehat{\mathrm{P}}_{\mathrm{bc}})$ setting

In the case of $(\widehat{\mathrm{P}}_{\mathrm{bc}})$ we wish to be able to choose $\mathcal{F}^{\mathrm{s}}$ to be $\mathcal{F}^{\mathrm{s}}_{0,\eta} + \tilde{h}$, or at least $\mathcal{F}^{\mathrm{s}}_{0} + \tilde{h}$. In practice, the above results imply that we can 'almost' do this. Indeed, recalling $\tilde{u}$ from (27), we see that

$$\|\tilde{u} - h\|_{H^{1/2}(\partial\Omega)} \leq C_T \|\tilde{u} - u^*\|_V \leq \frac{C_T \epsilon}{3\kappa_1},$$

and setting $\hat{u} := (\tilde{u} - \tilde{h}) + \tilde{h}$, which satisfies $\|\hat{u} - u^*\|_V = \|\tilde{u} - u^*\|_V \leq \epsilon/(3\kappa_1)$, we can conclude that $\mathcal{F}^{\mathrm{s}}$ can be taken to be $\mathcal{F}^{\mathrm{s}}_{\approx 0} + \tilde{h}$ where $\mathcal{F}^{\mathrm{s}}_{\approx 0}$ means a set of neural networks that are almost zero on the boundary. A similar argument applies for $\mathcal{F}^{\mathrm{t}}$ too. To be able to make this exact, we need to know if Sobolev functions that are zero on the boundary can be approximated by neural networks that are zero on the boundary and we need to do so in a uniform way (the width and depth of the network should be independent of the target function) for $\mathcal{F}^{\mathrm{t}}_0$. [15, Theorem 2] tells us that $H^1_0(\Omega)$ functions can be realised by ReLU neural networks of depth $\lceil \log_2(d+1) \rceil + 1$. In view of works such as [21], the desired uniform approximation result appears reasonable, however, we are not aware at present of any literature with quantitative rates that supply such a result.

3.2.4 Bounds on the statistical error

In this section, we bound the following quantity

$$\sup_{\substack{u \in \mathcal{F}^{\mathrm{s}} \\ v \in \mathcal{F}^{\mathrm{t}}}} \left| L_\gamma(u,v) + R_1(u) - R_2(v) - (\hat{L}_\gamma(u,v) + \hat{R}_1(u) - \hat{R}_2(v)) \right|.$$

in the context of the problem $(\widehat{\mathrm{P}}_{\mathrm{bc}})$, i.e., when the boundary condition is met and the penalty terms only contain contributions of the obstacle loss L_o. We carry out the calculations for neural neworks of the special ResNet structure $\mathcal{F}_{\mathrm{DRR}}$. A statistical error analysis for $\mathcal{F}_{\mathrm{FFN}}$ involving different loss functions can be found in [35, §4], whose arguments we adapt.

We first need the concept of Rademacher complexity and covering numbers.

Definition 3.20 (Rademacher complexity)**.** *Let $\mathcal{F}$ be a family of functions from Ω into $\mathbb{R}$ and let P be a distribution over Ω and $\{X_i\}_{i=1}^N$ be independent identically distributed (iid) samples from P. The* Rademacher complexity *of $\mathcal{F}$ associated with the distribution P and sample size N is defined as*

$$\mathcal{R}(\mathcal{F}) := \mathbb{E}_{\{X_i\}_{i=1}^N} \mathbb{E}_{\{\sigma_i\}_{i=1}^N} \left[\sup_{u \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^N \sigma_i u(X_i) \right],$$

where $\{\sigma_i\}_{i=1}^N$ are iid random variables such that $\mathbb{P}[\sigma_i = 1] = \mathbb{P}[\sigma_i = -1] = \frac{1}{2}$.

Definition 3.21 (Covering number)**.** *Given $\varepsilon > 0$, we say that $\mathcal{A} \subset \mathbb{R}^n$ is an* ε-cover of $\mathcal{B} \subset \mathbb{R}^n$ with respect to a metric ρ *if for all $v' \in \mathcal{B}$, there exists $v \in \mathcal{A}$ such that $\rho(v, v') \leq \varepsilon$.*

The ε-covering number of $\mathcal{B}$*, denoted as $\mathfrak{C}(\varepsilon, \mathcal{B}, \rho)$,is the minimum cardinality among all ε-covers of $\mathcal{B}$ with respect to the metric ρ.*

Rademacher complexity is useful because it bounds the statistical error from above as the next result shows. Below, we use the notation $U(\Omega)$ to denote the uniform distribution on Ω.

Lemma 3.22. *Assume that the collocation points $\{x_i\}_{i=1}^N$ (see Section 3.1) are iid drawn from $U(\Omega)$. Then*

$$\sup_{\substack{u\in\mathcal{F}^{\mathrm{s}}\\ v\in\mathcal{F}^{\mathrm{t}}}} |L_\gamma(u,v)+R_1(u)-R_2(v)-(\hat{L}_\gamma(u,v)+\hat{R}_1(u)-\hat{R}_2(v))| \le 2|\Omega|\sum_{i=1}^{10}\mathcal{R}(\mathcal{F}_i),$$

where

$$\begin{aligned}
\mathcal{F}_1 &= \{\|\nabla u\|^2 : u\in\mathcal{F}^{\mathrm{s}}\}, & \mathcal{F}_2 &= \{\nabla u\cdot\nabla v : u\in\mathcal{F}^{\mathrm{s}}, v\in\mathcal{F}^{\mathrm{t}}\},\\
\mathcal{F}_3 &= \{ku^2 : u\in\mathcal{F}^{\mathrm{s}}\}, & \mathcal{F}_4 &= \{kuv : u\in\mathcal{F}^{\mathrm{s}}, v\in\mathcal{F}^{\mathrm{t}}\},\\
\mathcal{F}_5 &= \{fu : u\in\mathcal{F}^{\mathrm{s}}\}, & \mathcal{F}_6 &= \{fv : v\in\mathcal{F}^{\mathrm{t}}\},\\
\mathcal{F}_7 &= \left\{\frac{1}{2\gamma}(u-v)^2 : u\in\mathcal{F}^{\mathrm{s}}, v\in\mathcal{F}^{\mathrm{t}}\right\}, & \mathcal{F}_8 &= \left\{\frac{1}{2\gamma}|\nabla u-\nabla v|^2 : u\in\mathcal{F}^{\mathrm{s}}, v\in\mathcal{F}^{\mathrm{t}}\right\},\\
\mathcal{F}_9 &= \{w_{o_1}|(\psi-u)^+|^2 : u\in\mathcal{F}^{\mathrm{s}}\}, & \mathcal{F}_{10} &= \{w_{o_2}|(\psi-v)^+|^2 : u\in\mathcal{F}^{\mathrm{t}}\}.
\end{aligned}$$

Proof. We can write L_γ as

$$\begin{aligned}
&L_\gamma(u,v)\\
&\quad= |\Omega|\mathbb{E}_{X\sim U(\Omega)}\big[\nabla u(X)\cdot(\nabla u(X)-\nabla v(X))+ku(X)(u(X)-v(X))-f(X)(u(X)-v(X))\big]\\
&\qquad-|\Omega|\mathbb{E}_{X\sim U(\Omega)}\left[\frac{1}{2\gamma}\big((u(X)-v(X))^2+|\nabla u(X)-\nabla v(X)|^2\big)\right].
\end{aligned}$$

With this in mind, we can decompose

$$\begin{aligned}
|L_\gamma(u,v)+R_1(u)-R_2(v)-(\hat{L}_\gamma(u,v)+\hat{R}_1(u)-\hat{R}_2(v))| \le{}& \sum_{i=1}^{8}|L_{\gamma,i}(u,v)-\hat{L}_{\gamma,i}(u,v)|\\
&+\sum_{i=1}^{2}|R_i(u,v)-\hat{R}_i(u,v)|
\end{aligned}$$

where

$$\begin{aligned}
L_{\gamma,1}(u,v) &= |\Omega|\mathbb{E}_{X\sim U(\Omega)}\left[\|\nabla u(X)\|^2\right], & L_{\gamma,2}(u,v) &= |\Omega|\mathbb{E}_{X\sim U(\Omega)}\left[\nabla u(X)\cdot\nabla v(X)\right],\\
L_{\gamma,3}(u,v) &= k|\Omega|\mathbb{E}_{X\sim U(\Omega)}\left[u(X)^2\right], & L_{\gamma,4}(u,v) &= k|\Omega|\mathbb{E}_{X\sim U(\Omega)}\left[u(X)v(X)\right],\\
L_{\gamma,5}(u,v) &= |\Omega|\mathbb{E}_{X\sim U(\Omega)}\left[f(X)u(X)\right], & L_{\gamma,6}(u,v) &= |\Omega|\mathbb{E}_{X\sim U(\Omega)}\left[f(X)v(X)\right],\\
L_{\gamma,7}(u,v) &= \frac{|\Omega|}{2\gamma}\mathbb{E}_{X\sim U(\Omega)}\left[(u(X)-v(X))^2\right], & L_{\gamma,8}(u,v) &= \frac{|\Omega|}{2\gamma}\mathbb{E}_{X\sim U(\Omega)}\left[|\nabla u(X)-\nabla v(X)|^2\right],\\
R_1(u,v) &= |\Omega|\mathbb{E}_{X\sim U(\Omega)}\left[|(\psi(X)-u(X)^+|^2\right], & R_2(u,v) &= |\Omega|\mathbb{E}_{X\sim U(\Omega)}\left[|(\psi(X)-v(X)^+|^2\right],
\end{aligned}$$

and

$$\hat{L}_{\gamma,1}(u,v) = \frac{|\Omega|}{N}\sum_{i=1}^{N}\|\nabla u(x_i)\|^2, \qquad \hat{L}_{\gamma,2}(u,v) = \frac{|\Omega|}{N}\sum_{i=1}^{N}\nabla u(x_i)\cdot\nabla v(x_i),$$

$$\hat{L}_{\gamma,3}(u,v) = \frac{k|\Omega|}{N}\sum_{i=1}^{N}u(x_i)^2, \qquad \hat{L}_{\gamma,4}(u,v) = \frac{k|\Omega|}{N}\sum_{i=1}^{N}u(x_i)v(x_i),$$

$$\hat{L}_{\gamma,5}(u,v) = \frac{|\Omega|}{N}\sum_{i=1}^{N}f(x_i)u(x_i), \qquad \hat{L}_{\gamma,6}(u,v) = \frac{|\Omega|}{N}\sum_{i=1}^{N}f(x_i)v(x_i),$$

$$\hat{L}_{\gamma,7}(u,v) = \frac{|\Omega|}{2\gamma N}\sum_{i=1}^{N}(u(x_i)-v(x_i))^2, \qquad \hat{L}_{\gamma,8}(u,v) = \frac{|\Omega|}{2\gamma N}\sum_{i=1}^{N}|\nabla u(x_i)-\nabla v(x_i)|^2,$$

$$\hat{R}_1(u,v) = \frac{|\Omega|}{N}\sum_{i=1}^{N}|(\psi(x_i)-u(x_i))^+|^2, \qquad \hat{R}_2(u,v) = \frac{|\Omega|}{N}\sum_{i=1}^{N}|(\psi(x_i)-v(x_i))^+|^2.$$

Now if we let $F_{u,v}$ denote a function of u and v and their first derivatives, we can write

$$\left|\,|\Omega|\mathbb{E}_{X\sim U(\Omega)}\left[F_{u,v}(X)\right] - \frac{|\Omega|}{N}\sum_{i=1}^{N}F_{u,v}(x_i)\right| = \frac{|\Omega|}{N}\left|\mathbb{E}_{\{X_i\}_{i=1}^N}\left[\sum_{i=1}^{N}F_{u,v}(X_i)-F_{u,v}(x_i)\right]\right|,$$

where $\{X_i\}_{i=1}^N$ are iid random variables drawn from $U(\Omega)$ and $\mathbb{E}_{\{X_i\}_{i=1}^N}$ means the (multiple) expectation with respect to $X_1,\cdots,X_N$. Taking the supremum over $\mathcal{F}^{\mathrm{s}}$ and $\mathcal{F}^{\mathrm{t}}$ and taking the expectation with respect to $\{x_i\}_{i=1}^N$ each of which are drawn from $U(\Omega)$, we have by Jensen's inequality

$$\begin{aligned}&\mathbb{E}_{\{x_i\}_{x=1}^N}\left[\sup_{\substack{u\in\mathcal{F}^{\mathrm{s}}\\ v\in\mathcal{F}^{\mathrm{t}}}}\frac{|\Omega|}{N}\left|\mathbb{E}_{\{X_i\}_{i=1}^N}\left[\sum_{i=1}^{N}F_{u,v}(X_i)-F_{u,v}(x_i)\right]\right|\right]\\ &\quad\le \frac{|\Omega|}{N}\mathbb{E}_{\{x_i\}_{i=1}^N}\mathbb{E}_{\{X_i\}_{i=1}^N}\left[\sup_{\substack{u\in\mathcal{F}^{\mathrm{s}}\\ v\in\mathcal{F}^{\mathrm{t}}}}\left|\sum_{i=1}^{N}F_{u,v}(X_i)-F_{u,v}(x_i)\right|\right].\end{aligned}$$

Let $\{\sigma_i\}_{i=1}^N$ be iid random variables such that $\mathbb{P}[\sigma_i=1]=\mathbb{P}[\sigma_i=-1]=\frac{1}{2}$ for all i. Note that for all i, $F_{u,v}(X_i)-F_{u,v}(x_i)$ and $F_{u,v}(x_i)-F_{u,v}(X_i)$ are equal in distribution and so the right-hand side

above is equal to

$$
\begin{aligned}
&\frac{|\Omega|}{2N}\mathbb{E}_{\{x_i\}_{i=1}^N}\mathbb{E}_{\{X_i\}_{i=1}^N}\left[\sup_{\substack{u\in\mathcal{F}^{\mathrm{s}}\\ v\in\mathcal{F}^{\mathrm{t}}}}\left|\sum_{i=1}^N F_{u,v}(X_i)-F_{u,v}(x_i)\right|\right]\\
&\quad+\frac{|\Omega|}{2N}\mathbb{E}_{\{x_i\}_{i=1}^N}\mathbb{E}_{\{X_i\}_{i=1}^N}\left[\sup_{\substack{u\in\mathcal{F}^{\mathrm{s}}\\ v\in\mathcal{F}^{\mathrm{t}}}}\left|\sum_{i=1}^N F_{u,v}(x_i)-F_{u,v}(X_i)\right|\right]\\
&\quad=\frac{|\Omega|}{N}\mathbb{E}_{\{x_i\}_{i=1}^N}\mathbb{E}_{\{X_i\}_{i=1}^N}\mathbb{E}_{\{\sigma_i\}_{i=1}^N}\left[\sup_{\substack{u\in\mathcal{F}^{\mathrm{s}}\\ v\in\mathcal{F}^{\mathrm{t}}}}\left|\sum_{i=1}^N \sigma_i(F_{u,v}(X_i)-F_{u,v}(x_i))\right|\right]\\
&\quad\le\frac{|\Omega|}{N}\mathbb{E}_{\{x_i\}_{i=1}^N}\mathbb{E}_{\{X_i\}_{i=1}^N}\mathbb{E}_{\{\sigma_i\}_{i=1}^N}\left[\sup_{\substack{u\in\mathcal{F}^{\mathrm{s}}\\ v\in\mathcal{F}^{\mathrm{t}}}}\left|\sum_{i=1}^N \sigma_i F_{u,v}(X_i)\right|\right]\\
&\qquad+\frac{|\Omega|}{N}\mathbb{E}_{\{x_i\}_{i=1}^N}\mathbb{E}_{\{X_i\}_{i=1}^N}\mathbb{E}_{\{\sigma_i\}_{i=1}^N}\left[\sup_{\substack{u\in\mathcal{F}^{\mathrm{s}}\\ v\in\mathcal{F}^{\mathrm{t}}}}\left|\sum_{i=1}^N \sigma_i F_{u,v}(x_i)\right|\right]\\
&\quad\le\frac{2|\Omega|}{N}\mathbb{E}_{\{x_i\}_{i=1}^N}\mathbb{E}_{\{\sigma_i\}_{i=1}^N}\left[\sup_{\substack{u\in\mathcal{F}^{\mathrm{s}}\\ v\in\mathcal{F}^{\mathrm{t}}}}\sum_{i=1}^N \sigma_i F_{u,v}(x_i)\right]\\
&\quad=2|\Omega|\mathcal{R}(\mathcal{F}),
\end{aligned}
$$

where $\mathcal{F}=\{F_{u,v}:u\in\mathcal{F}^{\mathrm{s}},v\in\mathcal{F}^{\mathrm{t}}\}$ is a function class parameterised by $\Theta^{\mathrm{s}}\times\Theta^{\mathrm{t}}$ where Θ^{s}, Θ^{t} are the network weights of u, v respectively. The above can be employed to show that for each i, $\sup_{\substack{u\in\mathcal{F}^{\mathrm{s}}\\ v\in\mathcal{F}^{\mathrm{t}}}}|L_{\gamma,i}(u,v)-\hat{L}_{\gamma,i}(u,v)|+|R_i(u,v)-\hat{R}_i(u,v)|$ is bounded by the Rademacher complexity of the function class $\mathcal{F}_i$. The claim follows. □

Now, [35, Lemma 4.8] shows that to bound the Rademacher complexity of a function class, we can bound the associated covering number which for Lipschitz functions we can bound by the covering number of the parameter space [35, Lemma 4.6]. Hence, our strategy is to first show that the set of neural networks defined by (14) and the first derivative of each element of the set are bounded and Lipschitz and from that conclude that each of the $\mathcal{F}_i$ (being a certain function of (14)) is also bounded and Lipschitz. This can then be used to bound their respective Rademacher complexities using the aforementioned lemmata.

From now on we assume that the activation function σ of our neural network architecture is sufficiently regular, see Assumption 3.23 — $\tanh$ is a valid activation function. The next lemma shows that in this case, neural networks given by (14) are Lipschitz in their parameters in the $C^1(\Omega)$ norm. For the proof, it is more convenient to rewrite the definition of our neural network architecture in a recursive manner and also to index by network layers rather than by network blocks. For $l=1,\dots,2\mathfrak{d}$, let

$$
\begin{aligned}
f^{(0)}(x)&=A^{(0)}x+b^{(0)},\\
f^{(l)}(x)&=\sigma(A^{(l)}f^{(l-1)}(x)+b^{(l)})+f^{(l-2)}(x)\mathbf{1}_{\{l\text{ even}\}},\\
u(x)&=A^{(2\mathfrak{d}+1)}f^{(2\mathfrak{d})}(x)+b^{(2\mathfrak{d}+1)},
\end{aligned}
$$

where $A^{(0)}=A_0$, $b^{(0)}=b_0$, $A^{(2\mathfrak{d}+1)}=A_{\mathfrak{d}+1}$, $b^{(2\mathfrak{d}+1)}=b_{\mathfrak{d}+1}$, $A^{(l)}=A_{ij}$ and $b^{(l)}=b_{ij}$ with $i=\lceil l/2\rceil$, $j=1$ for l odd and $j=2$ otherwise, where $A_0,A_{ij},A_{\mathfrak{d}+1},b_0,b_{ij}$ and $b_{\mathfrak{d}+1}$ are as in

the definition of (14). We also use the notation $A^{(l)} = (a_{qj}^{(l)})_{q,j=1}^{\mathfrak{w}}$ and $b^{(l)} = (b_1^{(l)}, \dots, b_{\mathfrak{w}}^{(l)})$. Each $f^{(i)} = (f_1^{(i)}, \dots, f_{\mathfrak{w}}^{(i)})$ and u are functions of their network weights θ. For calculations involving two different sets of network weights θ and $\tilde{\theta}$, we adorn a variable with a tilde (e.g. $\tilde{f}^{(i)}$, $\tilde{b}^{(i)}$) to indicate that the function or variable is with respect to $\tilde{\theta}$. Moreover, let n_i denote the number of network weights in the ith layer and N_i to be the total number of weights up to and including the ith layer. Without loss of generality, we assume $\mathfrak{w} \geq n_0$. Below, we let $C_\Omega \geq 1$ be a constant such that $|x| \leq C_\Omega$ for all $x \in \Omega$.

Assumption 3.23. *Suppose that*

(i) $\sigma \in C^1(\mathbb{R}) \cap W^{1,\infty}(\mathbb{R})$ *and* σ *and* σ' *are Lipschitz continuous with Lipschitz constants* L_σ *and* $L_{\sigma'}$ *respectively,*

(ii) Θ *is bounded by* B_θ.

Lemma 3.24. *Under Assumption 3.23,* u *defined by* (14) *satisfies* $\|u\|_{C^0(\Omega)} \leq B_u$, *and* $\left\|\partial_{x_p} u\right\|_{C^0(\Omega)} \leq B_{u'}$ *where*

$$B_u := (\mathfrak{w}+1)\bar{C}^2, \qquad B_{u'} := 2^{\mathfrak{d}}\mathfrak{w}^{2\mathfrak{d}+1}\bar{C}^{4\mathfrak{d}+2}, \qquad \bar{C} := \max(1, \|\sigma\|_{L^\infty(\Omega)}, \|\sigma'\|_{L^\infty(\Omega)}, L_\sigma, L_{\sigma'}, B_\theta).$$

Furthermore, the map $\theta \mapsto u$ *is Lipschitz from* $\ell^2(\mathbb{R})$ *into* $C^1(\Omega)$ *with*

$$\|u - \tilde{u}\|_{C^0(\Omega)} \leq L_u \|\theta - \tilde{\theta}\|_2, \qquad \left\|\partial_{x_p} u - \partial_{x_p}\tilde{u}\right\|_{C^0(\Omega)} \leq L_{u'}\|\theta - \tilde{\theta}\|_2,$$

where

$$L_u := \sqrt{\mathfrak{m}} 2^{2\mathfrak{d}-1}\mathfrak{w}^{2\mathfrak{d}+1}\bar{C}^{4\mathfrak{d}+1}C_\Omega, \qquad L_{u'} := \sqrt{\mathfrak{m}}\left(\sum_{k=0}^{2\mathfrak{d}} 2^k\right) 2^{4\mathfrak{d}-2}\mathfrak{w}^{4\mathfrak{d}+1}\bar{C}^{8\mathfrak{d}+1}C_\Omega.$$

Proof. First note that u is bounded because σ is bounded:

$$|u| \leq \sum_{j=1}^{\mathfrak{w}} |a_j^{(2\mathfrak{d}+1)}||f_j^{2\mathfrak{d}}| + |b^{(2\mathfrak{d}+1)}| \leq (\mathfrak{w}+1)\bar{C}^2.$$

We now prove the Lipschitzness of u. We have,

$$|f_q^{(i)} - \tilde{f}_q^{(i)}| \leq \left|\sigma\left(\sum_{j=1}^{\mathfrak{w}} a_{qj}^{(i)} f_j^{(i-1)} + b_q^{(i)}\right) - \sigma\left(\sum_{j=1}^{\mathfrak{w}} \tilde{a}_{qj}^{(i)} \tilde{f}_j^{(i-1)} + \tilde{b}_q^{(i)}\right)\right| + |f_q^{(i-2)} - \tilde{f}_q^{(i-2)}|\mathbf{1}_{\{i \text{ even}\}}.$$

We can bound the first term on the right-hand side using the same method as in the proof of [35, Lemma 4.9] to derive the following recurrence relation:

$$|f_q^{(i)} - \tilde{f}_q^{(i)}| \leq \bar{C}^2 \sum_{j=1}^{\mathfrak{w}} |f_j^{(i-1)} - \tilde{f}_j^{(i-1)}| + \bar{C}F^{(i-1)} \sum_{j=1}^{\mathfrak{w}} |a_{qj}^{(i)} - \tilde{a}_{qj}^{(i)}| + \bar{C}|b_q^{(i)} - \tilde{b}_q^{(i)}| + |f_q^{(i-2)} - \tilde{f}_q^{(i-2)}|\mathbf{1}_{\{i \text{ even}\}},$$

where $F^{(i)}$ is a constant satisfying $F^{(i)} \leq \|\sigma\|_{L^\infty(\Omega)}$ for $i \geq 1$ and

$$F^{(0)} = \sup_q |f_q^{(0)}| = \sup_q \left|\sum_{j=1}^{n} a_{qj}^{(0)} x_j + b_q^{(0)}\right| \leq \sup_q \sum_{j=1}^{n} |a_{qj}^{(0)}||x_j| + |b_q^{(0)}| \leq n_0 \bar{C} C_\Omega.$$

For $i = 0$,

$$|f_q^{(0)} - \tilde{f}_q^{(0)}| \leq \left| \sum_{j=1}^{n} a_{qj}^{(0)} x_j + b_q^{(0)} - \sum_{j=1}^{n} \tilde{a}_{qj}^{(0)} x_j - \tilde{b}_q^{(0)} \right| \leq \sum_{j=1}^{n} |x_j||a_{qj}^{(0)} - \tilde{a}_{qj}^{(0)}| + |b_q^{(0)} - \tilde{b}_q^{(0)}|$$
$$\leq C_\Omega \sum_{j=1}^{n_0} |\theta_j - \tilde{\theta}_j|.$$

Assume for $i \geq 1$ that

$$|f_q^{(i)} - \tilde{f}_q^{(i)}| \leq 2^{i-1} \mathfrak{w}^i \bar{C}^{2i} C_\Omega \sum_{j=1}^{N_i} |\theta_j - \tilde{\theta}_j|, \tag{29}$$

then

$$\begin{aligned}
&|f_q^{(i+1)} - \tilde{f}_q^{(i+1)}| \\
&\quad \leq \bar{C}^2 \sum_{j=1}^{\mathfrak{w}} |f_q^{(i)} - \tilde{f}_q^{(i)}| + \bar{C}^2 \sum_{j=1}^{\mathfrak{w}} |a_{qj}^{(i+1)} - \tilde{a}_{qj}^{(i+1)}| + \bar{C}|b_q^{(i+1)} - \tilde{b}_q^{(i+1)}| + |f_q^{(i-1)} - \tilde{f}_q^{(i-1)}| \mathbf{1}_{\{i+1 \text{ even}\}} \\
&\quad \leq 2^{i-1} \mathfrak{w}^i \bar{C}^{2i} \bar{C}^2 C_\Omega \sum_{j=1}^{\mathfrak{w}} \sum_{k=1}^{N_i} |\theta_k - \tilde{\theta}_k| + \bar{C}^2 \sum_{j=1}^{n_{i+1}} |\theta_j - \tilde{\theta}_j| + 2^{i-2} \mathfrak{w}^{i-1} \bar{C}^{2(i-1)} C_\Omega \sum_{j=1}^{N_{i-1}} |\theta_j - \tilde{\theta}_j| \\
&\quad \leq 2^i \mathfrak{w}^{i+1} \bar{C}^{2(i+1)} C_\Omega \sum_{j=1}^{N_{i+1}} |\theta_j - \tilde{\theta}_j|.
\end{aligned}$$

Hence, (29) is true for $i = 1, \ldots, 2\mathfrak{d}$ and so

$$\begin{aligned}
|u(x) - \tilde{u}(x)| &\leq \bar{C} \sum_{j=1}^{\mathfrak{w}} |f_j^{(2\mathfrak{d})} - \tilde{f}_j^{(2\mathfrak{d})}| + \bar{C} \sum_{j=1}^{\mathfrak{w}} |a_j^{(2\mathfrak{d}+1)} - \tilde{a}_j^{(2\mathfrak{d}+1)}| + |b^{(2\mathfrak{d}+1)} - \tilde{b}^{(2\mathfrak{d}+1)}| \\
&\leq 2^{2\mathfrak{d}-1} \mathfrak{w}^{2\mathfrak{d}} \bar{C}^{4\mathfrak{d}+1} C_\Omega \sum_{j=1}^{\mathfrak{w}} \sum_{k=1}^{N_{2\mathfrak{d}}} |\theta_k - \tilde{\theta}_k| + \bar{C} \sum_{j=1}^{N_{2\mathfrak{d}+1}} |\theta_j - \tilde{\theta}_j| \\
&\leq 2^{2\mathfrak{d}-1} \mathfrak{w}^{2\mathfrak{d}+1} \bar{C}^{4\mathfrak{d}+1} C_\Omega \sum_{j=1}^{\mathfrak{m}} |\theta_j - \tilde{\theta}_j| \\
&\leq \sqrt{\mathfrak{m}} 2^{2\mathfrak{d}-1} \mathfrak{w}^{2\mathfrak{d}+1} \bar{C}^{4\mathfrak{d}+1} C_\Omega \|\theta - \tilde{\theta}\|_2,
\end{aligned}$$

where the last line follows from Hölder's inequality. For the spatial derivatives of u we have,

$$\partial_{x_p} f_q^{(i)} = \sum_{j=1}^{\mathfrak{w}} a_{qj}^{(i)} \partial_{x_p} f_j^{(i-1)} \sigma' \left(\sum_{j=1}^{\mathfrak{w}} a_{qj}^{(i)} f_j^{(i)} + b_q^{(i)} \right) + \partial_{x_p} f_q^{(i-2)} \mathbf{1}_{\{i \text{ even}\}},$$

and so for i even

$$\begin{aligned}
|\partial_{x_p} f_q^{(i)}| &\leq \bar{C}^2 \sum_{j=1}^{\mathfrak{w}} |\partial_{x_p} f_j^{(i-1)}| + |\partial_{x_p} f_q^{(i-2)}| \leq \bar{C}^4 \sum_{j=1}^{\mathfrak{w}} \sum_{k=1}^{\mathfrak{w}} |\partial_{x_p} f_k^{(i-2)}| + |\partial_{x_p} f_q^{(i-2)}| \\
&\leq 2\mathfrak{w} \bar{C}^4 \sum_{j=1}^{\mathfrak{w}} |\partial_{x_p} f_j^{(i-2)}|.
\end{aligned}$$

Iterating the above gives, for i even,

$$|\partial_{x_p} f_q^{(i)}| \leq 2^{i/2} \mathfrak{w}^{i-1} \bar{C}^{2i} \sum_{j=1}^{\mathfrak{w}} |\partial_{x_p} f_j^{(0)}| \leq 2^{i/2} \mathfrak{w}^{i-1} \bar{C}^{2i} \sum_{j=1}^{\mathfrak{w}} |a_{jp}^{(0)}| \leq 2^{i/2} \mathfrak{w}^{i} \bar{C}^{2i+1}. \tag{30}$$

In a similar way, for i odd,

$$|\partial_{x_p} f_q^{(i)}| \leq \bar{C}^2 \sum_{j=1}^{\mathfrak{w}} |\partial_{x_p} f_j^{(i-1)}| \leq \bar{C}^2 \sum_{j=1}^{\mathfrak{w}} 2^{(i-1)/2} \mathfrak{w}^{i-1} \bar{C}^{2(i-1)+1} \leq 2^{i/2} \mathfrak{w}^{i} \bar{C}^{2i+1}. \tag{31}$$

Therefore,

$$|\partial_{x_p} u(x)| \leq \left| \partial_{x_p} \left(\sum_{j=1}^{\mathfrak{w}} a_j^{(2\mathfrak{d}+1)} f_j^{(2\mathfrak{d})} + b^{(2\mathfrak{d}+1)} \right) \right| \leq \sum_{j=1}^{\mathfrak{w}} |a_j^{(2\mathfrak{d}+1)}| |\partial_{x_p} f_j^{(2\mathfrak{d})}| \leq 2^{\mathfrak{d}} \mathfrak{w}^{2\mathfrak{d}+1} \bar{C}^{4\mathfrak{d}+2}.$$

We now show that the derivatives of u are Lipschitz. We have

$$\begin{aligned} |\partial_{x_p} f_q^{(i)} - \partial_{x_p} \tilde{f}_q^{(i)}| \leq & \left| \partial_{x_p} \sigma \left(\sum_{j=1}^{\mathfrak{w}} a_{qj}^{(i)} f_j^{(i)} + b_q^{(i)} \right) - \partial_{x_p} \sigma \left(\sum_{j=1}^{\mathfrak{w}} a_{qj}^{(i)} f_j^{(i)} + b_q^{(i)} \right) \right| \\ & + |\partial_{x_p} f_q^{(i-2)} - \partial_{x_p} \tilde{f}_q^{(i-2)}| \mathbf{1}_{\{i \text{ even}\}}. \end{aligned}$$

We can bound the first term on the right-hand side of the above in the same manner as in the proof of [35, Lemma 4.11] and use (29), (30) and (31) to show that

$$\begin{aligned} & |\partial_{x_p} f_q^{(i)} - \partial_{x_p} \tilde{f}_q^{(i)}| \qquad (32) \\ & \leq \bar{C} \sum_{j=1}^{\mathfrak{w}} |a_{qj}^{(i)}| |\partial_{x_p} f_j^{(i-1)} - \partial_{x_p} \tilde{f}_j^{(i-1)}| + |\partial_{x_p} f_j^{(i-1)}| |a_{qj}^{(i)} - \tilde{a}_{qj}^{(i)}| \\ & \quad + \bar{C} \left(\sum_{j=1}^{\mathfrak{w}} |\tilde{a}_{qj}^{(i)}| |\partial_{x_p} \tilde{f}_j^{(i-1)}| \right) \left(\sum_{j=1}^{\mathfrak{w}} |f_j^{(i-1)}| |a_{qj}^{(i)} - \tilde{a}_{qj}^{(i)}| + |\tilde{a}_{qj}^{(i)}| |f_j^{(i-1)} - \tilde{f}_j^{(i-1)}| + |b_q^{(i)} - \tilde{b}_q^{(i)}| \right) \\ & \quad + |\partial_{x_p} f_q^{(i-2)} - \partial_{x_p} \tilde{f}_q^{(i-2)}| \qquad (33) \\ & \leq \bar{C}^2 \sum_{j=1}^{\mathfrak{w}} |\partial_{x_p} f_j^{(i-1)} - \partial_{x_p} \tilde{f}_j^{(i-1)}| + 2^{2(i-1)} \mathfrak{w}^{2i} \bar{C}^{4i} C_\Omega \sum_{j=1}^{N_i} |\theta_j - \tilde{\theta}_j| + |\partial_{x_p} f_q^{(i-2)} - \partial_{x_p} \tilde{f}_q^{(i-2)}|. \end{aligned}$$

For $i = 0$, we have $|\partial_{x_p} f_q^{(0)} - \partial_{x_p} \tilde{f}_q^{(0)}| \leq |a_{qp}^{(0)} - \tilde{a}_{qp}^{(0)}| \leq \sum_{j=1}^{n_0} |\theta_j - \tilde{\theta}_j|$. Assume for $i \geq 1$ that

$$|\partial_{x_p} f_q^{(i)} - \partial_{x_p} \tilde{f}_q^{(i)}| \leq 2^{2(i-1)} \left(\sum_{k=0}^{i} 2^k \right) \mathfrak{w}^{2i} \bar{C}^{4i} C_\Omega \sum_{j=1}^{N_i} |\theta_j - \tilde{\theta}_j|,$$

then

$$
\begin{aligned}
&|\partial_{x_p} f_q^{(i+1)} - \partial_{x_p} \tilde{f}_q^{(i+1)}| \\
&\quad \leq \bar{C}^2 \sum_{j=1}^{\mathfrak{w}} |\partial_{x_p} f_j^{(i)} - \partial_{x_p} \tilde{f}_j^{(i)}| + 2^{2i} \mathfrak{w}^{2(i+1)} \bar{C}^{4(i+1)} C_\Omega \sum_{j=1}^{N_{i+1}} |\theta_j - \tilde{\theta}_j| + |\partial_{x_p} f_q^{(i-1)} - \partial_{x_p} \tilde{f}_q^{(i-1)}| \\
&\quad \leq 2^{2(i-1)} \left(\sum_{k=0}^{i} 2^k \right) \mathfrak{w}^{2i} \bar{C}^{4i+2} C_\Omega \sum_{j=1}^{m} \sum_{k=1}^{N_i} |\theta_k - \tilde{\theta}_k| + 2^{2i} \mathfrak{w}^{2(i+1)} \bar{C}^{4(i+1)} C_\Omega \sum_{j=1}^{N_{i+1}} |\theta_j - \tilde{\theta}_j| \\
&\qquad + 2^{2(i-2)} \left(\sum_{k=0}^{i-1} 2^k \right) \mathfrak{w}^{2(i-1)} \bar{C}^{4(i-1)} C_\Omega \sum_{j=1}^{N_{i-1}} |\theta_j - \tilde{\theta}_j| \\
&\quad \leq 2^{2i} \left(\sum_{k=0}^{i+1} 2^k \right) \mathfrak{w}^{2(i+1)} \bar{C}^{4(i+1)} C_\Omega \sum_{j=1}^{N_{i+1}} |\theta_j - \tilde{\theta}_j|.
\end{aligned}
$$

Hence, by induction

$$
\begin{aligned}
|\partial_{x_p} u(x) - \partial_{x_p} \tilde{u}(x)| &\leq \sum_{j=1}^{\mathfrak{w}} |\partial_{x_p} f_j^{(2\mathfrak{d})}| |a_j^{(2\mathfrak{d}+1)} - \tilde{a}_j^{(2\mathfrak{d}+1)}| + |\tilde{a}_j^{(2\mathfrak{d}+1)}| |\partial_{x_p} f_j^{(2\mathfrak{d})} - \partial_{x_p} \tilde{f}_j^{(2\mathfrak{d})}| \\
&\leq 2^{4\mathfrak{d}-2} \left(\sum_{k=0}^{2\mathfrak{d}} 2^k \right) \mathfrak{w}^{4\mathfrak{d}+1} \bar{C}^{8\mathfrak{d}+1} C_\Omega \sum_{j=1}^{\mathfrak{m}} |\theta_j - \tilde{\theta}_j| \\
&\leq 2^{4\mathfrak{d}-2} \sqrt{\mathfrak{m}} \left(\sum_{k=0}^{2\mathfrak{d}} 2^k \right) \mathfrak{w}^{4\mathfrak{d}+1} \bar{C}^{8\mathfrak{d}+1} C_\Omega \|\theta_j - \tilde{\theta}\|_2.
\end{aligned}
$$

□

In the setting of ($\widehat{\mathrm{P}}_{\mathrm{bc}}$), we use the same neural network architecture for both the solution and the test function, in particular, we take $u = \bar{u}\eta + \bar{h}$ where $\bar{u} \in \mathcal{F}(\mathfrak{d}, \mathfrak{w}, \sigma)$, $\eta \in C^1(\bar{\Omega})$ satisfying $\eta|_{\partial\Omega} = 0$ and $\bar{h} \in C^1(\bar{\Omega})$ satisfies the boundary condition. In practice, we take $\bar{h}$ to be also a neural network of the form (14). Assume also that η and $\|\nabla\eta\|^2$ are bounded by B_η and $B_{\eta'}$ respectively and the source term f and obstacle ψ are bounded by B_f and B_ψ respectively. Then, u is bounded with constant $B_u B_\eta + B_u =: D_1$ is and Lipschitz with constant $L_u B_\eta + L_u =: D_2$, and $\partial_{x_p} u$ is bounded by $B_{u'} B_\eta + B_u B_{\eta'} + B_{u'} =: D_3$ and is Lipschitz with constant $L_{u'} B_\eta + L_u B_{\eta'} + L_{u'} =: D_4$. It follows that for any $g \in \mathcal{F}_i$, $i = 1, \ldots, 10$, we have $|g| \leq B_i$ where

$$
\begin{aligned}
&B_1 = B_2 = nD_3^2, && B_3 = B_4 = kD_1^2, && B_5 = B_6 = B_f D_1, \\
&B_7 = 2\gamma^{-1} D_1^2, && B_8 = 2\gamma^{-1} B_1, && B_9 = 2w_{o_1}(B_\psi^2 + D_1^2), \\
&B_{10} = 2w_{o_2}(B_\psi^2 + D_1^2), &&&&
\end{aligned}
$$

and g is Lipschitz with constant L_i where

$$
\begin{aligned}
&L_1 = L_2 = 2nD_3 D_4, && L_3 = L_4 = 2kD_1 D_2, && L_5 = L_6 = B_f D_2, \\
&L_7 = 4\gamma^{-1} D_1 D_2, && L_8 = 4\gamma^{-1} n D_3 D_4, && L_9 = 2w_{o_1}(B_\psi + D_1) D_2, \\
&L_{10} = 2w_{o_2}(B_\psi + D_1) D_2. &&&&
\end{aligned}
$$

With this we can prove the following theorem which gives a bound of the statistical error. In the proof, we denote generic constants which may differ line-by-line and on its dependencies by $K(\cdot)$. Recall also that n is the dimension of the domain $\Omega \subset \mathbb{R}^n$.

Theorem 3.25 (Statistical error estimate for $(\widehat{\mathsf{P}}_{\mathrm{bc}})$)**.** *Consider the $\mathcal{F}_{\mathrm{DRR}}$ case. Let $f \in C^0(\Omega) \cap L^\infty(\Omega)$, the activation function $\sigma \in C^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ be Lipschitz and Θ be bounded. Then we have*

$$\mathbb{E}_{\{x_i\}_{i=1}^N}\left[\sup_{\substack{u\in\mathcal{F}^{\mathrm{s}}\\ v\in\mathcal{F}^{\mathrm{t}}}} |L_\gamma(u,v)+R_1(u)-R_2(v)-(\hat{L}_\gamma(u,v)+\hat{R}_1(u)-\hat{R}_2(v))|\right] \leq \frac{K(\eta,f)n^{\frac{3}{2}}\mathfrak{m}\sqrt{\sum_{k=0}^{2\mathfrak{d}}2^k}2^{\frac{9\mathfrak{d}-2}{2}}\mathfrak{w}^{7\mathfrak{d}+3}\bar{C}^{14\mathfrak{d}+9}\sqrt{C_\Omega}}{N^{\frac{1}{4}}}.$$

A similar bound holds for the $\mathcal{F}_{\mathrm{FFN}}$ case.

Recall again that we do not have boundary penalty terms in this formulation.

Proof. We address the DRR setting; the standard FFN case follows by the obvious modifications. Let $\mathcal{F}$ be an arbitrary function class such that for $f \in \mathcal{F}$ we have $\|f\|_{L^\infty(\Omega)} \leq B$ and f is L-Lipschitz with respect to the parameter θ. Then by [35, Lemmas 4.8, 4.6, 4.5] in that order, we have

$$\begin{aligned}\mathcal{R}(\mathcal{F}) &\leq \inf_{0<\delta<B/2}\left(4\delta+\frac{12}{\sqrt{N}}\int_\delta^{B/2}\sqrt{\log \mathfrak{C}(\varepsilon,\mathcal{F},\|\cdot\|_\infty)}\,\mathrm{d}\varepsilon\right)\\ &\leq \inf_{0<\delta<B/2}\left(4\delta+\frac{12}{\sqrt{N}}\int_\delta^{B/2}\sqrt{\mathfrak{m}\log\left(\frac{2L\bar{C}\sqrt{\mathfrak{m}}}{\varepsilon}\right)}\,\mathrm{d}\varepsilon\right)\\ &\leq \frac{4}{\sqrt{N}}+\frac{6\sqrt{\mathfrak{m}}B}{\sqrt{N}}\sqrt{\log(2L\bar{C}\sqrt{N\mathfrak{m}})},\end{aligned}$$

where in the last line we set $\delta = 1/\sqrt{N}$. It is clear that $B_i \leq KB_1$ and $L_i \leq KL_1$, $i = 1,\dots,10$ for some constant K so it suffices to bound $\mathcal{R}(\mathcal{F}_1)$ only. Using the inequality $(a+b)^2 \leq 2(a^2+b^2)$, we have

$$B_1 \leq 4n(B_u^2B_{\eta'}^2+B_{u'}^2B_\eta^2+B_{u'}^2) \leq K(\eta)n(B_u^2+B_{u'}^2) \leq K(\eta)nB_{u'}^2 = K(\eta)n2^{2\mathfrak{d}}\mathfrak{w}^{4\mathfrak{d}+2}\bar{C}^{8\mathfrak{d}+4}$$

and

$$L_1 \leq K(\eta)b(B_u+B_{u'})(L_u+L_{u'}) \leq K(\eta)nB_{u'}L_{u'} = K(\eta)n\sqrt{\mathfrak{m}}\left(\sum_{k=0}^{2\mathfrak{d}}2^k\right)2^{5\mathfrak{d}-2}\mathfrak{w}^{6\mathfrak{d}+2}\bar{C}^{12\mathfrak{d}+3}C_\Omega.$$

Therefore,

$$\begin{aligned}\mathcal{R}(\mathcal{F}_1) &\leq \frac{4}{\sqrt{N}}+\frac{K(\eta)\sqrt{\mathfrak{m}}n2^{2\mathfrak{d}}\mathfrak{w}^{4\mathfrak{d}+2}\bar{C}^{8\mathfrak{d}+4}}{\sqrt{N}}\sqrt{\log\left(2\left(\sum_{k=0}^{2\mathfrak{d}}2^k\right)2^{5\mathfrak{d}-2}\mathfrak{w}^{6\mathfrak{d}+2}\bar{C}^{12\mathfrak{d}+4}C_\Omega\sqrt{N}\mathfrak{m}\right)}\\ &\leq \frac{4}{\sqrt{N}}+\frac{K(\eta)}{N^{\frac{1}{4}}}n^{\frac{3}{2}}\mathfrak{m}\sqrt{\sum_{k=0}^{2\mathfrak{d}}2^k}2^{\frac{9\mathfrak{d}-2}{2}}\mathfrak{w}^{7\mathfrak{d}+3}\bar{C}^{14\mathfrak{d}+9}\sqrt{C_\Omega}\\ &\leq \frac{K(\eta)n^{\frac{3}{2}}\mathfrak{m}\sqrt{\sum_{k=0}^{2\mathfrak{d}}2^k}2^{\frac{9\mathfrak{d}-2}{2}}\mathfrak{w}^{7\mathfrak{d}+3}\bar{C}^{14\mathfrak{d}+9}\sqrt{C_\Omega}}{N^{\frac{1}{4}}}.\end{aligned}$$

Recalling that the left-hand side of the statement of the theorem is bounded by $\sum_{i=1}^{10}\mathcal{R}(\mathcal{F}_i)$ completes the proof. □

The theorem tells us that the statistical error can be made arbitrarily small if the number of grid points N chosen is large enough, and it also indicates that the error may grow if the width and depth of the network increase.

4 Numerical details and examples

We numerically solve the discrete minmax problem ($\widehat{\mathrm{P}}_{\mathrm{bc}}$) for various examples via the approach detailed in Section 3.1.1. Given a concrete obstacle problem, the first step is to construct the function $\tilde{h}$, which is supposed to (approximately) satisfy the boundary condition. We find such a function by simply minimising the loss[4]

$$\min_{w \in \mathcal{F}_{\mathrm{DRR}}(\mathfrak{d},\mathfrak{w},\tanh)} L_b(w) + L_o(w),$$

so that the output $\tilde{h}$ is itself a neural network. We stop training when $\tilde{h}$ satisfies the boundary condition up to some error threshold. With this $\tilde{h}$ fixed, we then solve ($\widehat{\mathrm{P}}_{\mathrm{bc}}$) by applying the gradient descent ascent (GDA) scheme in Algorithm 1.

The GDA approach alternates between two steps: in the descent step it keeps the test function fixed and updates the weights of the solution candidate by one gradient step, in the ascent step it keeps the solution candidate fixed and updates the test function by one gradient step. Since this basic form of the GDA approach converges to the minmax point if the function is convex-concave (see Section 3.1.4) and since the continuous formulation of the minmax problem is indeed convex-concave, one can at least heuristically hope that the GDA algorithm converges to the solution under suitable conditions.

Our implementation and codebase can be found at [4]. It is written in `Python` and uses the `PyTorch` framework. The gradient with respect to the weights $\theta_i \in \Theta_i$ that appear in lines 4 and 4 of Algorithm 1 are calculated using standard automatic differentiation libraries implemented in `PyTorch` [49]. The optimization steps are each one step of the built-in optimizer `AdamW` [46]. `AdamW`, which belongs to the family of stochastic gradient descent methods, is an adaptive gradient method that adjusts the learning rate of each parameter of the neural network individually such that the learning rates of parameters that typically have a larger gradient during training are slowed down more than the ones which typically have a small gradient. The major hyperparameters (which are detailed below) were found with the use of the `Optuna` package [3]. The results of the numerical experiments that are given below were performed on an NVIDIA A100 80GB PCIe GPU on `Python` version 3.12.4. By using the package `Ray` we were able to parallelise up to 10 training sessions on a single GPU; the training was done on a cluster with 4 GPUs.

For all our examples we use one of two sets of training hyperparameters, depending on whether the example is in 1D or 2D, see Tables 1 and 2 for the most important parameters. While it is certainly possible to improve the results by using individual hyperparameters for each example, we do not do this for simplicity.

We now present the examples. Along the way, we shall explore and analyse certain aspects related to the choice of weights and parameters.

[4] We included the obstacle constraint loss in the minimisation problem in order to give the minmax problem ($\widehat{\mathrm{P}}_{\mathrm{bc}}$) a good start in the sense that $\tilde{h} \in K$ (again, up to error).

Algorithm 1 Alternating gradient descent ascent algorithm in the $h \equiv 0$ setting

Input: Parameters including initial learning rates $\lambda^{\mathfrak{s}}, \lambda^{\mathfrak{t}}$, learning rate schedulers $\mathrm{LR}^{\mathfrak{s}}, \mathrm{LR}^{\mathfrak{t}}$, weights w_{o_1}, w_{o_2}, γ, number of collocation points N and number of epochs M
Output: Neural network solution $\widehat{u}_A$
Initialize weights $\theta^{\mathfrak{s}}$, $\theta^{\mathfrak{t}}$ of neural networks $u_{\theta^{\mathfrak{s}}}$, $v_{\theta^{\mathfrak{t}}}$ (with zero boundary conditions) representing solution and test function respectively
for epoch $= 0, 1, \ldots, M$ **do**
    if epoch is odd **then**
        $\theta_t \leftarrow$ UPDATETESTFUNCTION($\theta^{\mathfrak{s}}, \theta^{\mathfrak{t}}, \lambda^{\mathfrak{t}}$):
        $\lambda^{\mathfrak{t}} \leftarrow \mathrm{LR}^{\mathfrak{t}}(\lambda^{\mathfrak{t}}$, epoch)
    else
        $\theta_s \leftarrow$ UPDATESOLUTION($\theta^{\mathfrak{s}}, \theta^{\mathfrak{t}}, \lambda^{\mathfrak{s}}$):
        $\lambda^{\mathfrak{s}} \leftarrow \mathrm{LR}^{\mathfrak{s}}(\lambda^{\mathfrak{s}}$, epoch)
    end if
end for
Set $\widehat{u}_A := u_{\theta^{\mathfrak{s}}}$

function UPDATETESTFUNCTION($\theta^{\mathfrak{s}}, \theta^{\mathfrak{t}}, \lambda^{\mathfrak{t}}$)
    Sample grid points in the interior $X = \{x_1, \ldots, x_N\}$
    $(\hat{L}^X, \hat{L}_o^X, \hat{D}^X) \leftarrow$ COMPUTELOSSES($X, \theta^{\mathfrak{s}}, \theta^{\mathfrak{t}}$)
    $g \leftarrow \nabla_{\theta^{\mathfrak{t}}}(-\hat{L}^X(\theta^{\mathfrak{t}}) - \hat{D}^X(\theta^{\mathfrak{t}}) + \hat{L}_o^X(\theta^{\mathfrak{t}}))$
    $\theta^{\mathfrak{t}} \leftarrow \mathtt{AdamW}(\theta^{\mathfrak{t}}, \lambda^{\mathfrak{t}}, g)$
    return $\theta^{\mathfrak{t}}$
end function

function UPDATESOLUTION($\theta^{\mathfrak{s}}$,$\theta^{\mathfrak{t}}$)
    Sample grid points in the interior $X = \{x_1, \ldots, x_N\}$
    $(\hat{L}^X, \hat{L}_o^X, \hat{D}^X) \leftarrow$ COMPUTELOSSES($X, \theta^{\mathfrak{s}}, \theta^{\mathfrak{t}}$)
    $g \leftarrow \nabla_{\theta^{\mathfrak{s}}}(\hat{L}^X(\theta^{\mathfrak{s}}) + \hat{D}^X(\theta^{\mathfrak{s}}) + \hat{L}_o^X(\theta^{\mathfrak{s}}))$
    $\theta^{\mathfrak{s}} \leftarrow \mathtt{AdamW}(\theta^{\mathfrak{s}}, \lambda^{\mathfrak{s}}, g)$
    return $\theta^{\mathfrak{s}}$
end function

function COMPUTELOSSES($X, u_{\theta^{\mathfrak{s}}}, v_{\theta^{\mathfrak{t}}}$)
    Approximate via Monte Carlo integration
    - the loss $L(u_{\theta^{\mathfrak{s}}}, v_{\theta^{\mathfrak{t}}})$ by $\hat{L}^X$
    - the gap term loss $\|u_{\theta^{\mathfrak{s}}} - v_{\theta^{\mathfrak{t}}}\|^2$ by $\hat{D}^X$
    - the obstacle loss $\|(\psi - u_{\theta^{\mathfrak{t}}})^+\|^2 + \|(\psi - v_{\theta^{\mathfrak{t}}})^+\|^2$ by $\hat{L}_o^X$

    return $(\hat{L}^X, \hat{L}_o^X, \hat{D}^X)$
end function

Parameter	**Value**
width $(\mathfrak{w})$	80
depth $(\mathfrak{d})$	4
activation (σ)	$\tanh$
architecture $\mathcal{F}$	$\mathcal{F}_{\mathrm{DRR}}$

Table 1: Architecture for the solution and test function.

Parameter	**1D**	**2D**
n_interior (N)	1024	1024
n_boundary (N^b)	2	256
Epochs (M)	12000	12000
lr_soln (λ^{s})	0.002	0.003
lr_testfn (λ^{t})	0.001	0.0047
$\mathrm{LR}^{\mathrm{s}}, \mathrm{LR}^{\mathrm{t}}$	CosineAnnealingWarmRestarts	
T_0	2001	2001
T_mult	2	2
weight_soln_obs (w_{o_1})	8000	5000
weight_testfn_obs (w_{o_2})	1500	5000
weight_gap_term $((2\alpha)^{-1})$	0.0001	0.0005

Table 2: Parameters for the 1D and 2D examples. Above, T_0 and T_mult are parameters in the learning rate scheduler.

4.1 1D Examples

4.1.1 Example 1: A benchmark example from [13, Example 1, §3]

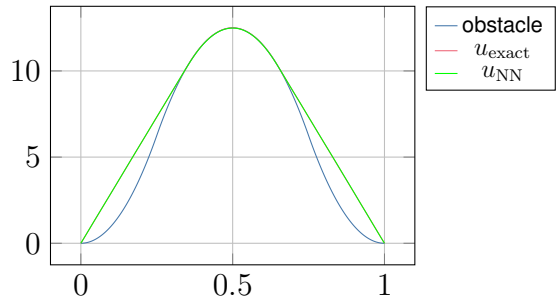
Here we set $\Omega = (0,1)$, $h \equiv 0$,

$$\psi(x) := \begin{cases} 100x^2 & : x \in [0, 0.25], \\ 100x(1-x) - 12.5 & : x \in (0.25, 0.5], \\ \psi(1-x) & : x \in (0.5, 1], \end{cases}$$

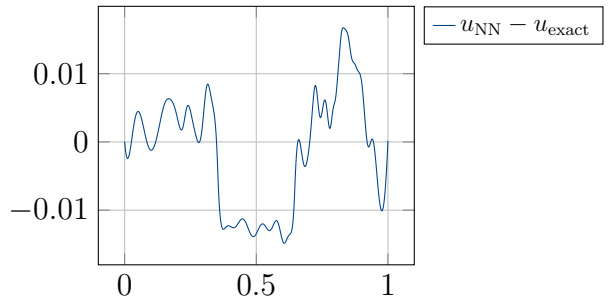
$f \equiv 0$, and solve problem (1) with $Au = -u_{xx}$. The exact solution is

$$u(x) = \begin{cases} (100 - 50\sqrt{2})x & : 0 \le x < \frac{1}{2\sqrt{2}}, \\ 100x(1-x) - 12.5 & : \frac{1}{2\sqrt{2}} \le x < 1 - \frac{1}{2\sqrt{2}}, \\ (50\sqrt{2} - 100)(1-x) & : 1 - \frac{1}{2\sqrt{2}} \le x \le 1. \end{cases}$$

The results are displayed in Figure 1. Figure 1a shows that our learned solution more or less coincides with the true solution, at least visually. A plot of the obstacle is also included for convenience. In Figure 1b we plot the difference of the true and learned solutions, which confirms that our approach produces a good result.



(a) Learned and exact solution.



(b) Difference between learned and exact solution.

Figure 1: **Example 1**. The difference between the learned solution u_{NN} and the true solution u_{exact} is smaller than 0.02. On the coincidence set $[1/2\sqrt{2}, 1 - 1/2\sqrt{2}] \approx [0.354, 0.646]$, u_{NN} violates the obstacle condition nearly constantly. The maximal difference is comparable to the one obtained in [13].

In Figure 2 we plot the evolution of the L^2 and L^∞ errors during training for a number of different runs or initialisations (i.e., different random seeds). The variance in the evolutions for the different runs is explained by this and the Monte Carlo integration. We see that the variance between the runs is very small.

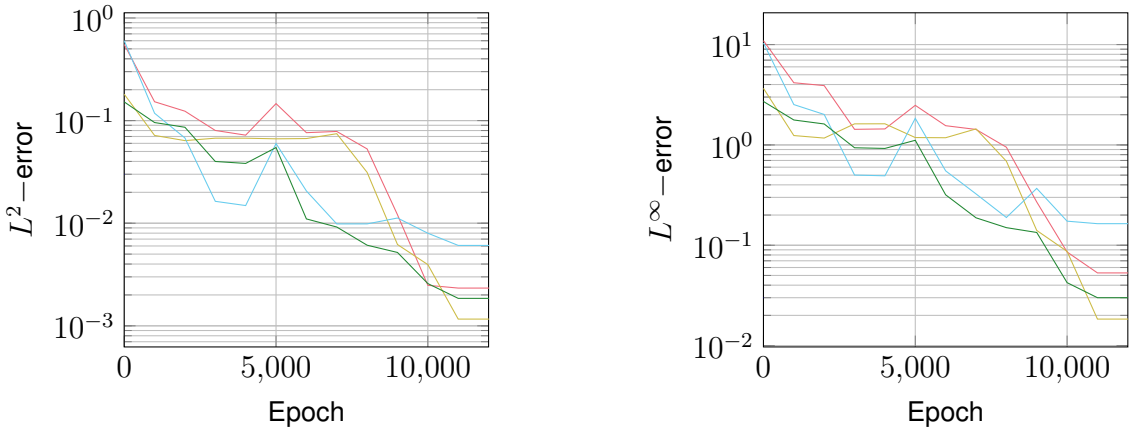


Figure 2: Training trajectories for Example 1.

4.1.2 Example 2: a non-symmetric case

Let us proceed with an example where the elliptic operator associated to the VI is non-symmetric. As we mentioned, being able to handle non-symmetric VIs is one of the advantages of our work which distinguishes us from other existent work in the literature to the best of our knowledge. We set $\Omega = (-2, 2)$, $h \equiv 0$, $\psi(x) = 1 - x^2$, define

$$f(x) := \begin{cases} (4 - 2\sqrt{3}) & : x \in [-2, -2 + \sqrt{3}), \\ -(4 - 2\sqrt{3}) & : x \in [2 - \sqrt{3}, 2], \\ -(2\sqrt{3} - 2) & : x \in [-2 + \sqrt{3}, 2 - \sqrt{3}], \\ 0 & : \text{otherwise}, \end{cases}$$

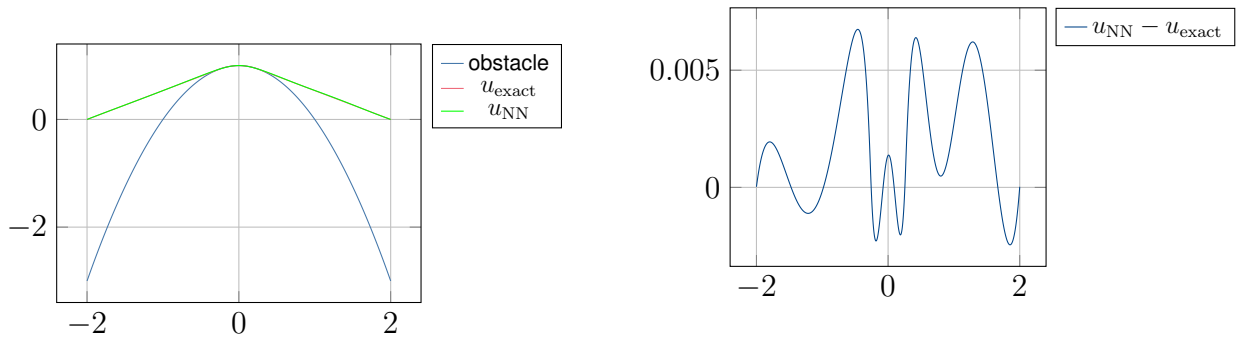
and solve the non-symmetric version of (1) with $Au = -u_{xx} + u_x$. The exact solution is

$$u(x) = \begin{cases} (4 - 2\sqrt{3})(x + 2) & : -2 \le x < -2 + \sqrt{3}, \\ 1 - x^2 & : -2 + \sqrt{3} \le x < 2 - \sqrt{3}, \\ (4 - 2\sqrt{3})(2 - x) & : 2 - \sqrt{3} \le x < 2. \end{cases}$$

This is a modification of the (symmetric) example in [64, §4.1]. We again observe a pleasing result, see Figure 3.

Obstacle weight experiment Recall that we enforce the obstacle condition via penalty terms for the solution and test function with weights w_{o_1} and w_{o_2} respectively. In Figure 5 we study the effect that different weights $w_{o_1} = w_{o_2}$ have on the error in the context of Example 2. We see that there is an optimal value for the penalty, which is reasonable: for small weights the solution is not punished enough for violating the obstacle constraint and for large weights, during the training phase, the solution candidate is pushed above the obstacle so that halting at the correct solution is overridden by the "momentum"that arises from the obstacle term.

A weight of 900 leads to an optimum in the L^2-sense, but once gradients are taken into account, 500 appears to be the optimal choice (with 900 not too distant). By examining the individual errors that produced each mean, 500 has a lower variance in L^2 and the variances are comparable in H^1, but the lowest error for 900 bests the lowest error for 500, across both measures of error. That said, we emphasise that both weights produce low errors: they differ only on an order of 10^{-2}.



(a) Learned and exact solution.

(b) Difference between learned and exact solution.

Figure 3: **The non-symmetric Example 2**. The difference has a magnitude smaller than 0.007.

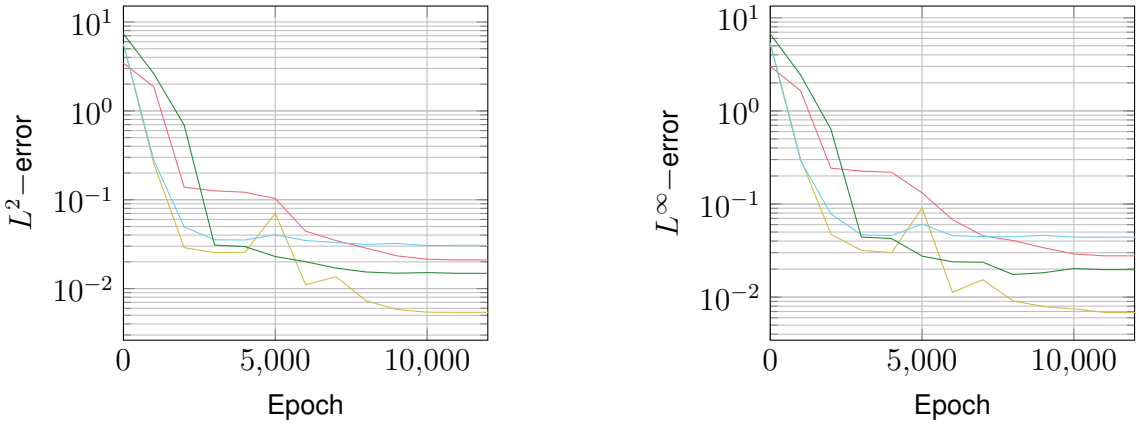


Figure 4: Training trajectories for Example 2. Compared to Figure 2, the convergence happens earlier.

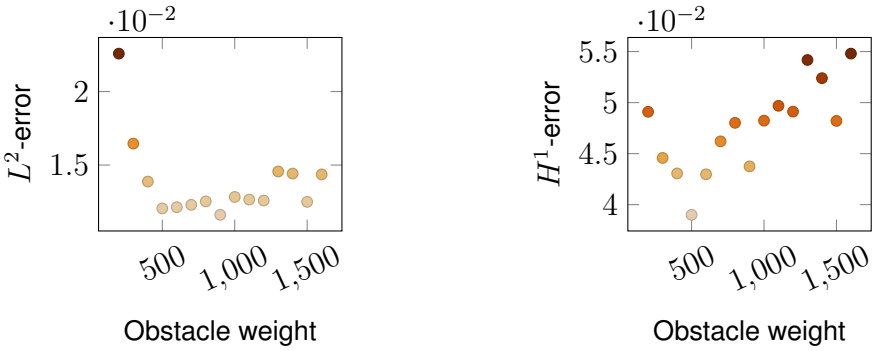


Figure 5: Mean (over 50 seeds) of L^2 and H^1 errors for Example 2 for different values of $w_{o_1} = w_{o_2}$.

4.1.3 Example 3: a piecewise smooth case

Our final one-dimensional example is taken from [14, §2]; we call this 'piecewise' because the solution is composed of five separate pieces and to the unsuspecting eye appears to be a piecewise affine

function. The setting here is $\Omega = (-1, 1)$, $h \equiv 0$, $f \equiv 0$, and with $\alpha = 0.4$,

$$\psi(x) := \begin{cases} \varphi\left(x + \frac{1}{2}\right)\left(\frac{3}{2} - 12|x + \frac{1}{2}|^{2-\alpha}\right) - \frac{1}{2} & : x \in (-1, 0], \\ \varphi\left(x - \frac{1}{2}\right)\left(\frac{3}{2} - 12|x - \frac{1}{2}|^{2-\alpha}\right) - \frac{1}{2} & : x \in (0, 1), \end{cases}$$

where $\varphi \in C_c^\infty(\mathbb{R})$ satisfies

$$0 \leq \varphi \leq 1, \quad \varphi = 1 \text{ in } (-0.3, 0.3), \quad \operatorname{supp}(\varphi) \subset [-0.4, 0.4].$$

We solve the problem (1) with $Au = -u_{xx}$. The exact solution is

$$u(x) = \begin{cases} \psi(-\beta - 0.5)\frac{x+1}{0.5-\beta} & : x \in (-1, -\beta - 0.5), \\ \psi(x) & : x \in [-0.5 - \beta, -0.5), \\ 1 & : x \in [-0.5, 0.5), \\ \psi(x) & : x \in [0.5, 0.5 + \beta), \\ \psi(\beta + 0.5)\frac{x-1}{\beta-0.5} & : x \in [\beta + 0.5, 1), \end{cases} \tag{34}$$

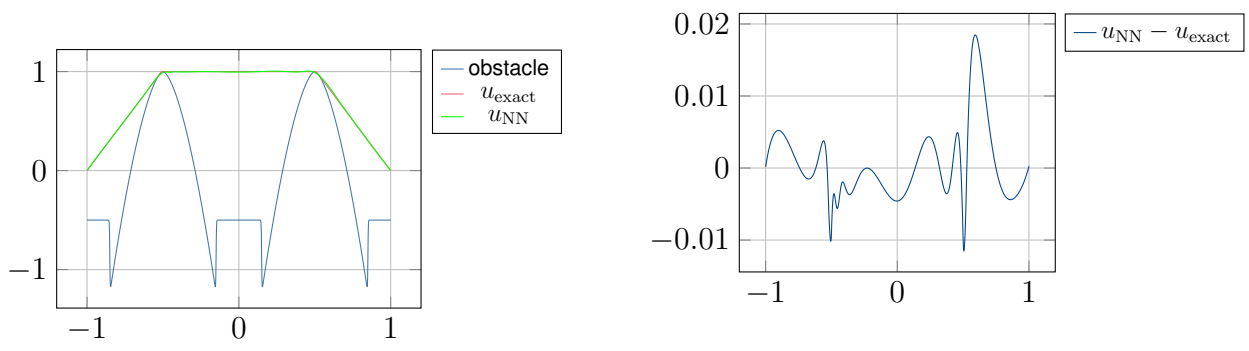
where the constant β is the unique solution of the equation

$$\psi(-\beta - 0.5) = (0.5 - \beta)\psi'(-\beta - 0.5), \quad \beta \in (0, 0.3).$$

In practice, we take $\beta = 0.02376$ as an approximate solution of the equation, and as for the function φ, we use

$$\varphi(x) := \frac{\mu(0.4 - |x|)}{\mu(|x| - 0.3) + \mu(0.4 - |x|)}, \qquad \text{where} \qquad \mu(x) := \begin{cases} \exp(-1/x) & : x > 0, \\ 0 & : x \leq 0. \end{cases}$$

The results can be seen in Figure 6.



(a) Learned and exact solution.

(b) Difference between learned and exact solution.

Figure 6: **Piecewise smooth case of Example 3**. The difference to the exact solution has a magnitude smaller than 0.03. Compared to the small average difference the violation of the obstacle condition at the peaks (for $x = 0.5$ and $x = -0.5$) is relatively large.

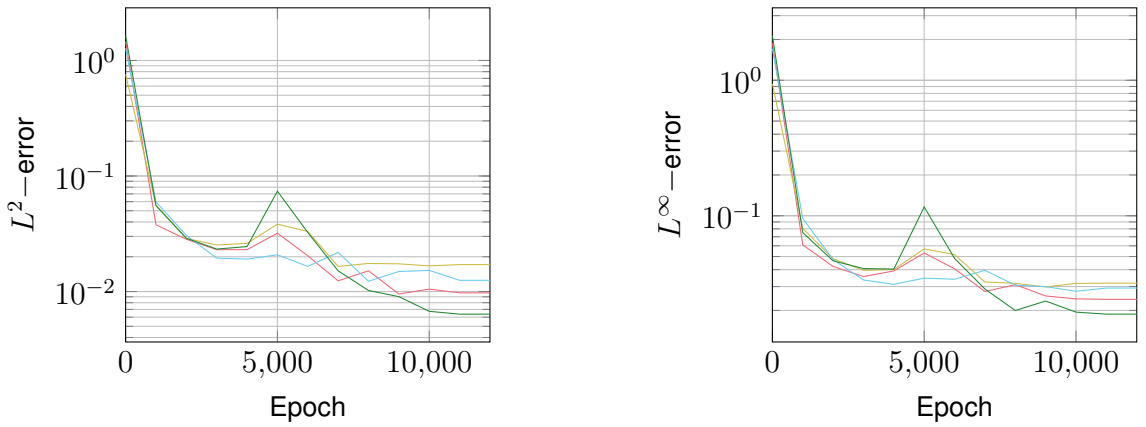


Figure 7: Training trajectories for Example 3.

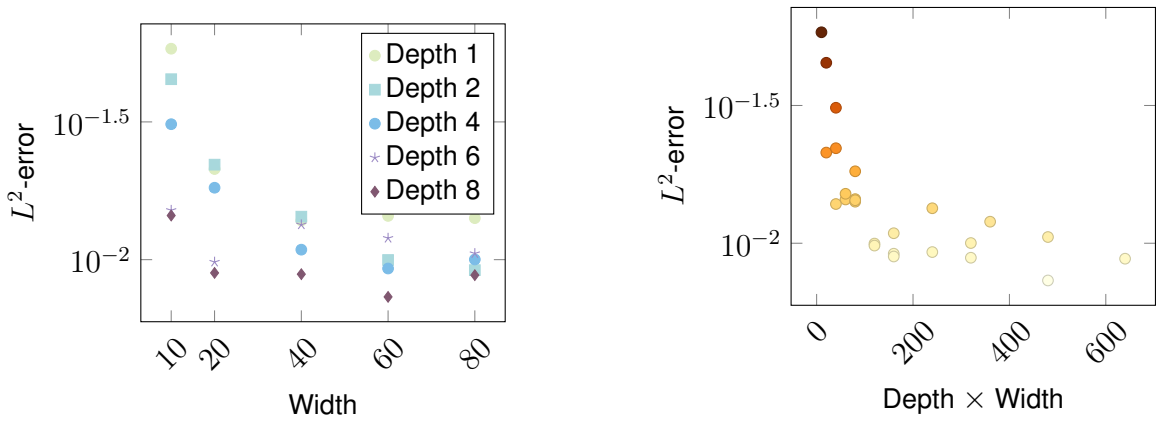


Figure 8: Effect of neural network size on the solution of Example 3 (means over 4 seeds are plotted).

Mesh independence experiment Let us look at how the L^2 error for Example 3 behaves as we change the size (i.e., width and depth) of the neural networks parametrising the solution and test function (both networks share the same architecture). In Figure 8 (left) we visualise the error for each depth as the width varies over the horizontal axis. We see that when the depth is sufficiently large (≥ 6), the errors for all widths are within a margin of approximately 10^{-2}. Likewise, when the width is large enough (≥ 40), the errors stay below approximately 0.015. This can be interpreted as a kind of mesh independence: the size of the network determines the number of learnable parameters, and the figure shows that the quality of the approximation is (roughly) independent of the number of learnable parameters. It is also illustrative to see how the error varies with respect to total number of learnable weights (which is a function of the width multiplied by the depth), see Figure 8 (right). The trend is clearly that the more weights the better the solution, but we again see a threshold level of weights beyond which the error is acceptable.

4.2 2D Examples

4.2.1 Example 4: an example from optimal control

We take this example from [48, §7.1]. Set $\Omega = (0,1)^2$, $A = -\Delta$, $\psi \equiv 0$, $h \equiv 0$, and using the intermediary quantities

$$\begin{aligned} z_1(x) &:= -4096x^6 + 6144x^5 - 3072x^4 + 512x^3, \\ z_2(x) &:= -244.140625x^6 + 585.9375x^5 - 468.75x^4 + 125x^3, \\ \zeta &:= \begin{cases} z_1(x-0.5)z_2(y) & : x \in (0.5,1) \text{ and } y \in (0,0.8), \\ 0 & : \text{otherwise}, \end{cases} \end{aligned}$$

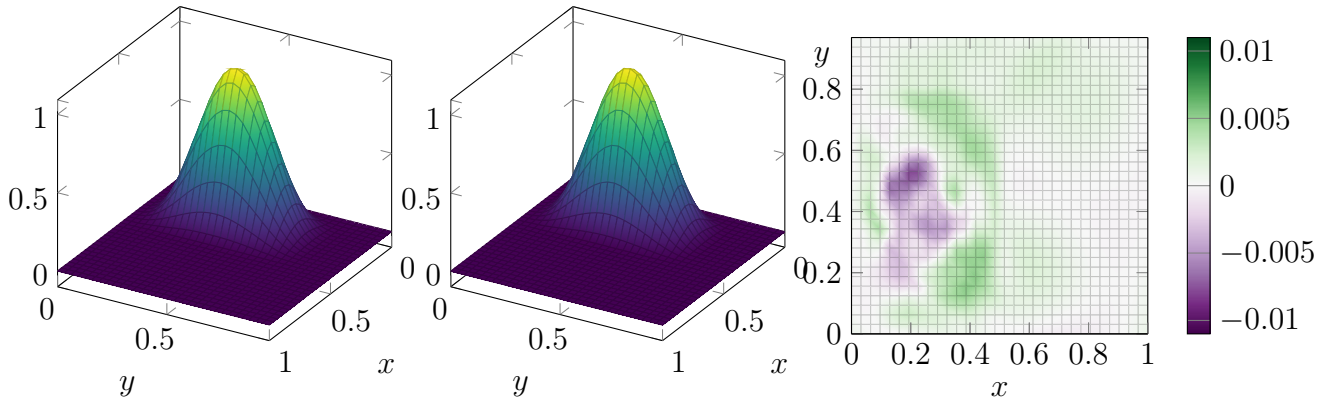
we define the source term

$$f(x,y) := -\zeta - \begin{cases} z_1(x)z_2''(y) + z_1''(x)z_2(y) & : x < 0.5 \text{ and } y < 0.8, \\ 0 & : \text{otherwise}, \end{cases}$$

and consider the VI (1) with $Au = -\Delta u$. The exact solution is

$$u(x,y) = \begin{cases} z_1(x)z_2''(y) + z_1''(x)z_2(y) & : x < 0.5 \text{ and } y < 0.8, \\ 0 & : \text{otherwise}. \end{cases}$$

We visualise the results in Figure 9. Since the obstacle is the zero function, we do not plot it explicitly.



(a) Learned solution u_{NN}. (b) Exact solution u_{exact}. (c) Difference $u_{\mathrm{NN}} - u_{\mathrm{exact}}$.

Figure 9: **Example 4**. Similarly to the 1D examples the solution candidate rather slightly violates the obstacle conditions (this is illustrated by the purple area in the third plot).

Mesh independence experiment We again take a look at how the network sizes affect the quality of the solution, now for Example 4, see Figure 11. Note that it is not the case that the bigger the network, the better the solution. Larger networks may require a longer training period (number of epochs), whereas smaller networks may suffer from underfitting. The optimal for this particular example and run appears to be a network architecture of depth 2 and width 40; we again recall the fact that we choose our architecture to provide a good result for all the examples that we considered and a case-by-case fine-tuning would likely improve our results, but this is not the focus of the paper.

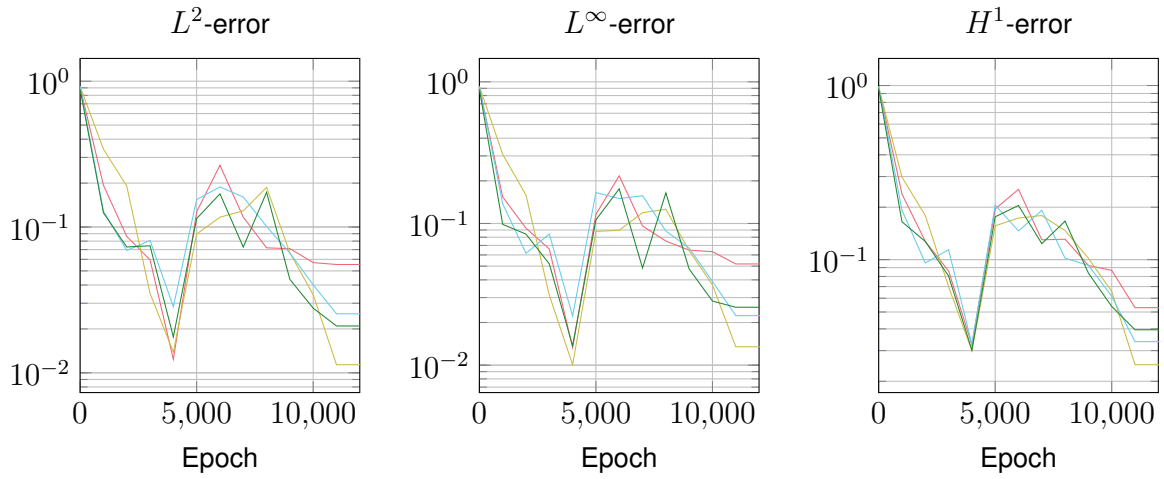


Figure 10: Training trajectories for Example 4. This example converges already at 4000 epochs; the warm restart phase (this occurs due to our choice of learning rate scheduler) from 4000 till the end does not greatly improve the error for all runs. Note the strong correlation between the two L^p and H^1 errors.

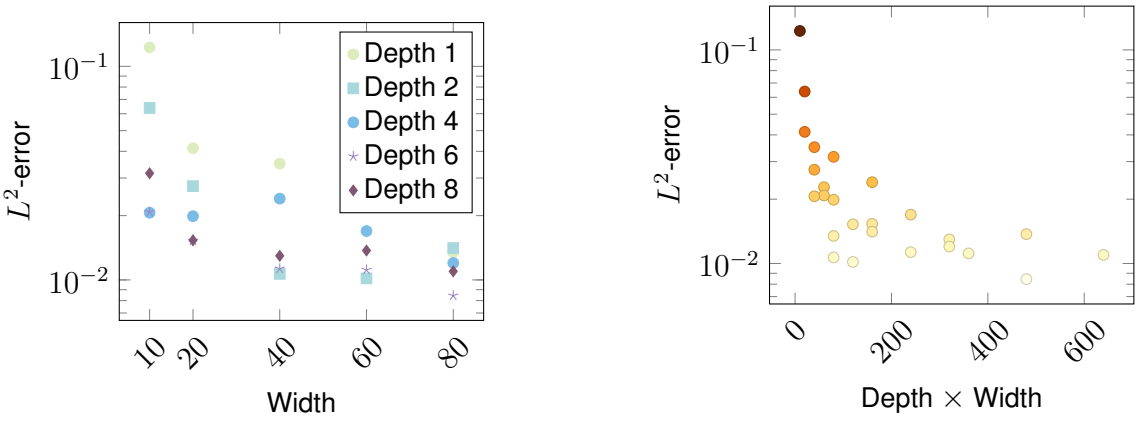


Figure 11: Effect of neural network size on the solution of Example 4 (1 seed is used).

4.2.2 Examples 5 and 6: biactive cases

Here we look at situations where biactivity (or lack of strict complementarity) is present. Let us briefly explain what this means. Recall from (5) that solutions of the VI satisfy the condition $(Au - f)(u - \psi) = 0$. This means that pointwise a.e., either $Au - f$ or $u - \psi$ or both must be equal to zero. When both are zero, we have biactivity, i.e., the multiplier $Au - f$ vanishes on the coincidence set $\{u = \psi\}$. It is well known that traditional numerical methods such as active set update strategies face great difficulty in handling problems with biactivity.

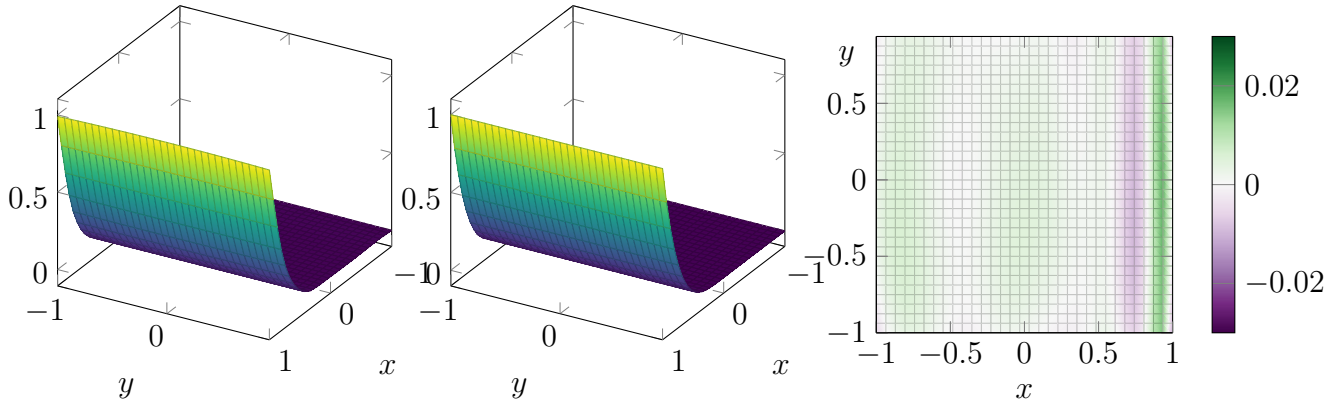
We consider first an example taken from [38, §4.8.1]. Set $\Omega = (-1, 1)^2$,

$$f(x, y) = \begin{cases} 0 & : x < 0, \\ -12x^2 & : \text{otherwise}, \end{cases}$$

$\psi \equiv 0$, and the exact solution

$$u(x, y) = \begin{cases} 0 & : x < 0, \\ x^4 & : \text{otherwise}, \end{cases}$$

of (1) with $Au = -\Delta u$. The boundary data h is determined by u. The results are shown in Figure 12.



(a) Learned solution u_{NN}. (b) Exact solution u_{exact}. (c) Difference $u_{\mathrm{NN}} - u_{\mathrm{exact}}$.

Figure 12: **Example 5**: biactive case from [38, §4.8.1].

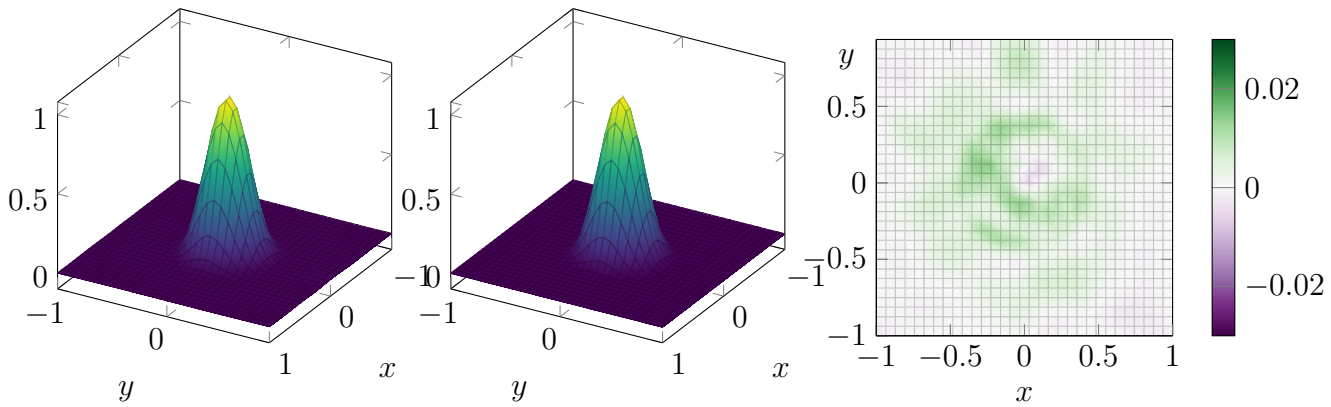
Example 6 also exhibits biactivity and displays a nonsmoothness in the multipler $Au - f$, and is taken from [38, §4.8.3]. Set $\Omega = (-1, 1)^2$, $\psi \equiv 0$, and the exact solution

$$u(x, y) = \begin{cases} (1 - 4x^2 - 4y^2)^4 & : x^2 + y^2 < \frac{1}{4}, \\ 0 & : \text{otherwise}, \end{cases}$$

informing the source term

$$f(x, y) = -\Delta u(x, y) - \begin{cases} 1 & : x^2 + y^2 > \frac{3}{4}, \\ 0 & : \text{otherwise}, \end{cases}$$

for (1) with $Au = -\Delta u$. The boundary data h again is taken from u. See Figure 13 for the results. In



(a) Learned solution u_{NN}. (b) Exact solution u_{exact}. (c) Difference $u_{\mathrm{NN}} - u_{\mathrm{exact}}$.

Figure 13: **Example 6**: second biactive case taken from [38, §4.8.3].

both cases, we see that our solution algorithm performs well (the results for Example 5 are comparable to the other examples; Example 6 is slightly more tricky, see Section 4.3) and does not encounter the biactivity as an issue due to our approach. Thus our work offers a potential advantage over active set methods.

Regularised gap weight experiment We investigate what effect changing the weight $1/(2\gamma)$ in front of the regularised gap term has on the solution for Example 6. See Figure 14 for a plot of the H^1 errors observed for different choices of gap weight. Observe that the error improves as the weight is increased until an optimum point is reached, after which the error increases exponentially. This tallies with our earlier remarks that γ should be sufficiently large (and hence $1/(2\gamma)$ should be sufficently small), but taking too small a value does not lead to a huge improvement, i.e., there is a trade-off.

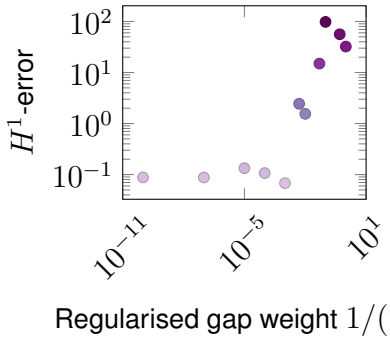


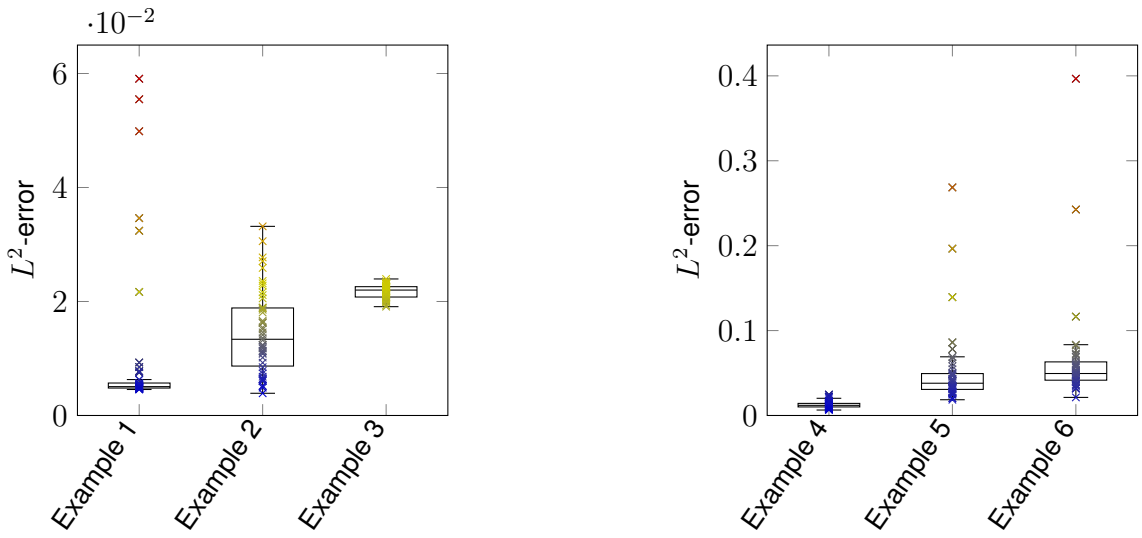
Figure 14: Mean over 10 seeds of the H^1-errors for Example 6 for different values of the regularised gap weight.

4.3 Replicability and randomness

It is well known that an improperly designed solution algorithm or loss function can lead to results that are radically different depending on the machine that the code is run on or depending on the choice of the random seed. In the context of our work, this means that it is worthwhile to check whether the hyperparameters we suggest actually produce quantifiably good results (i.e., low error of the learned solution in comparison to the exact solution) when run on different devices. As a statistical analysis of running our code on many different devices is impractical, we instead perform a sensitivity analysis where we vary the random seed and check whether the resulting model is robust with respect to the seed (the idea being that changing the random seed is akin to using a different device).

We present box plots of L^2 errors corresponding to different seeds for our examples in Figure 15a and Figure 15b. The bottom of each box corresponds to the 25th percentile and the top to the 75th percentile of 50 runs. For the 1D examples we see that the 75th percentile is smaller than 0.025; for the 2D examples the 75th percentile is smaller than 0.065.

We refer to [11] for more details and suggestions of good practices on this topic, and also to [47] for a study of model stability with respect to random seeds and techniques for improvement (see also references therein).



(a) 1D examples

(b) 2D examples

Figure 15: Box plots of L^2 errors corresponding to 50 different seeds for the 1D and 2D examples.

A Proofs

Proof of Proposition 2.2. We follow [52, §1:3, p. 4]. Let u solve (5) with the stated regularity. Taking $v \in K$, we have, making use of the Gelfand triple structure $V \subset L^2(\Omega) \subset V^*$ and that $Au \in L^2(\Omega)$,

$$\langle Au - f, u - v\rangle = \int_\Omega (Au - f)(u - \psi) + \int_\Omega (Au - f)(\psi - v) \leq 0.$$

This shows that u solves (1). Now for the reverse direction, suppose u solves (1). Choose the test function $v = u + \varphi$ where $\varphi \in C_c^\infty(\Omega)$ and $\varphi \geq 0$. Then since $Au \in L^2(\Omega)$, we get

$$\int_\Omega (Au - f)\varphi \geq 0$$

whence the arbitrariness of φ yields $Au - f \geq 0$ a.e.

Define the non-coincidence set $I := \{u > \psi\}$ which is an open set since u and ψ are both continuous. Take $\varphi \in C_c^\infty(I)$; it follows that there exists an ϵ_0 with $u \pm \epsilon\varphi \in K$ for all $\epsilon \leq \epsilon_0$. Using this as the choice of test function in (1), we obtain

$$\pm\epsilon\langle Au - f, \varphi\rangle = \pm\epsilon \int_\Omega (Au - f)\varphi \leq 0$$

whence $Au = f$ a.e. in I. From this one gets $(Au - f)(u - \psi) = 0$ a.e. in Ω. □

Proof of Lemma 3.1. Consider the $\mathcal{F}_{\mathrm{DRR}}$ case (the other case will follow by trivial adjustments) and write $u = M \circ \tilde{u}$. Firstly, observe that $\tilde{u}\colon \mathbb{R}^{\mathfrak{m}} \times \overline{\Omega} \to \mathbb{R}$ is continuous as it is the composition of continuous functions. Since $\mathfrak{B}_i$ is in fact continuously differentiable for all i (it is clear for $i = 0, \mathfrak{d}$, and for $i \in \{1, \ldots, \mathfrak{d} - 1\}$, this follows by the chain rule and the fact that $\sigma \in C^1(\mathbb{R})$), making use of the chain rule, $\nabla\tilde{u}\colon \mathbb{R}^{\mathfrak{m}} \times \overline{\Omega} \to \mathbb{R}^n$ is also continuous.

Now, to prove the desired continuity, we begin by taking a sequence $\theta_n \to \theta$. The set $K_1 := (\cup_{n\in\mathbb{N}}\{\theta_n\}) \cup \{\theta\}$ is a compact set and hence $\tilde{u}\colon \overline{\Omega} \times K_1 \to \mathbb{R}$ is uniformly continuous. For every $\delta > 0$, if n is sufficiently large, we get $|\theta_n - \theta| \leq \delta$. Fix $\epsilon > 0$. It follows by uniform continuity that there exists $N_0 \in \mathbb{N}$ such that if $n \geq N_0$, then

$$|\tilde{u}(x, \theta_n) - \tilde{u}(x, \theta)| \leq \epsilon,$$

uniformly in x. Here we can take the supremum over x and conclude that $\theta \mapsto \tilde{u}(\theta, \cdot)$ is continuous as a map from $\mathbb{R}^{\mathfrak{m}}$ to $C^0(\overline{\Omega})$. Using the continuity of $(\theta, x) \mapsto \nabla\tilde{u}(\theta, x)$, we obtain continuity into $C^1(\overline{\Omega})$ by a similar argument. The continuity of M finishes the argument. □

References

[1] R. A. Adams and J. J. F. Fournier. *Sobolev spaces*. Second. Vol. 140. Pure and Applied Mathematics (Amsterdam). Elsevier/Academic Press, Amsterdam, 2003, pp. xiv+305. ISBN: 0-12-044143-8.

[2] Y. Aizawa, M. Kimura, and K. Matsui. "Universal approximation properties for an ODENet and a ResNet: Mathematical analysis and numerical experiments". In: *Discrete and Continuous Dynamical Systems - B* 29.1 (2024), pp. 351–376. ISSN: 1531-3492. DOI: 10.3934/dcdsb.2023099. URL: https://www.aimsciences.org/article/id/646c64474a9fed1ce4f

[3] T. Akiba et al. "Optuna: A Next-generation Hyperparameter Optimization Framework". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2019.

[4] A. Alphonse, A. Kister, and C. H. Lun. *NNVI: Code for the paper*. https://github.com/amal-alphonse/NNVI. 2024. URL: https://github.com/amal-alphonse/NNVI.

[5] G. Auchmuty. "Variational principles for variational inequalities". In: *Numer. Funct. Anal. Optim.* 10.9-10 (1989), pp. 863–874. ISSN: 0163-0563,1532-2467. DOI: 10.1080/01630568908816335. URL: https://doi.org/10.1080/01630568908816335.

[6] A. Auslender. *Optimisation*. Méthodes numériques, Maîtrise de Mathématiques et Applications Fondamentales. Masson, Paris-New York-Barcelona, 1976, pp. vi+178.

[7] H. E. Bahja et al. *A physics-informed neural network framework for modeling obstacle-related equations*. 2023. arXiv: 2304.03552 [cs.LG]. URL: https://arxiv.org/abs/2304.03552.

[8] G. Bao et al. "Numerical solution of inverse problems by weak adversarial networks". In: *Inverse Problems* 36.11 (2020), pp. 115003, 31. ISSN: 0266-5611. DOI: 10.1088/1361-6420/abb447. URL: https://doi.org/10.1088/1361-6420/abb447.

[9] S. Bartels. *Numerical methods for nonlinear partial differential equations*. Vol. 47. Springer Series in Computational Mathematics. Springer, Cham, 2015, pp. x+393. DOI: 10.1007/978-3-319-13797-1. URL: https://doi.org/10.1007/978-3-319-13797-1.

[10] S. Bertoluzza, E. Burman, and C. He. *Best approximation results and essential boundary conditions for novel types of weak adversarial network discretizations for PDEs*. 2024. arXiv: 2307.05012 [math.NA]. URL: https://arxiv.org/abs/2307.05012.

[11] S. Bethard. "We need to talk about random seeds". In: *arXiv e-prints*, arXiv:2210.13393 (Oct. 2022), arXiv:2210.13393. DOI: 10.48550/arXiv.2210.13393. arXiv: 2210.13393 [cs.CL].

[12] I. Brevis, I. Muga, and K. G. van der Zee. "Neural control of discrete weak formulations: Galerkin, least squares & minimal-residual methods with quasi-optimal weights". In: *Computer Methods in Applied Mechanics and Engineering* 402 (2022). A Special Issue in Honor of the Lifetime Achievements of J. Tinsley Oden, p. 115716. ISSN: 0045-7825. DOI: https://doi.org/10.1016/j.cma.2022.115716. URL: https://www.sciencedirect.com/science/article/pii/S0045782522006715.

[13] X. Cheng et al. “A deep neural network-based method for solving obstacle problems”. In: *Nonlinear Anal. Real World Appl.* 72 (2023), Paper No. 103864, 16. ISSN: 1468-1218. DOI: 10.1016/j.nonrwa.2023.103864. URL: https://doi.org/10.1016/j.nonrwa.2023.103864.

[14] C. Christof and C. Meyer. “A note on a priori L^p-error estimates for the obstacle problem”. In: *Numer. Math.* 139.1 (2018), pp. 27–45. ISSN: 0029-599X,0945-3245. DOI: 10.1007/s00211-017-0931-5. URL: https://doi.org/10.1007/s00211-017-0931-5.

[15] P. Dondl, J. Müller, and M. Zeinhofer. “Uniform convergence guarantees for the deep Ritz method for nonlinear problems”. In: *Adv. Contin. Discrete Models* (2022), Paper No. 49, 19. ISSN: 2731-4235. DOI: 10.1186/s13662-022-03722-8. URL: https://doi.org/10.1186/s13662-022-03722-8.

[16] W. E and B. Yu. “The deep Ritz method: a deep learning-based numerical algorithm for solving variational problems”. In: *Commun. Math. Stat.* 6.1 (2018), pp. 1–12. ISSN: 2194-6701. DOI: 10.1007/s40304-018-0127-z. URL: https://doi.org/10.1007/s40304-018-0127-z.

[17] I. Ekeland and R. Témam. *Convex analysis and variational problems*. English. Vol. 28. Classics in Applied Mathematics. Translated from the French. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1999, pp. xiv+402. ISBN: 0-89871-450-8. DOI: 10.1137/1.9781611971088. URL: https://doi.org/10.1137/1.9781611971088.

[18] M. Fukushima. “Equivalent differentiable optimization problems and descent methods for asymmetric variational inequality problems”. In: *Math. Programming* 53.1 (1992), pp. 99–110. ISSN: 0025-5610,1436-4646. DOI: 10.1007/BF01585696. URL: https://doi.org/10.1007/BF01585696.

[19] R. Glowinski. *Numerical methods for nonlinear variational problems*. Springer Series in Computational Physics. Springer-Verlag, New York, 1984, pp. xv+493. ISBN: 0-387-12434-9. DOI: 10.1007/978-3-662-12613-4. URL: https://doi.org/10.1007/978-3-662-12613-4.

[20] R. Glowinski, J.-L. Lions, and R. Trémolières. *Numerical analysis of variational inequalities*. Vol. 8. Studies in Mathematics and its Applications. Translated from the French. North-Holland Publishing Co., Amsterdam-New York, 1981, pp. xxix+776. ISBN: 0-444-86199-8.

[21] I. Gühring and M. Raslan. “Approximation rates for neural networks with encodable weights in smoothness spaces”. In: *Neural Networks* 134 (2021), pp. 107–130. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2020.11.010. URL: https://www.sciencedirect.com/science/article/pii/S0893608020303956.

[22] K. He et al. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[23] K. He et al. “Identity Mappings in Deep Residual Networks”. In: *Computer Vision – ECCV 2016*. Ed. by B. Leibe et al. Cham: Springer International Publishing, 2016, pp. 630–645. ISBN: 978-3-319-46493-0.

[24] M. Hieber and I. Wood. “The Dirichlet problem in convex bounded domains for operators in non-divergence form with L^∞-coefficients”. In: *Differential Integral Equations* 20.7 (2007), pp. 721–734. ISSN: 0893-4983.

[25] M. Hintermüller, K. Ito, and K. Kunisch. “The primal-dual active set strategy as a semismooth Newton method”. In: *SIAM J. Optim.* 13.3 (2002), pp. 865–888. ISSN: 1052-6234,1095-7189. DOI: 10.1137/S1052623401383558. URL: https://doi.org/10.1137/S1052623401383558.

[26] M. Hintermüller and A. Laurain. “A shape and topology optimization technique for solving a class of linear complementarity problems in function space”. In: *Comput. Optim. Appl.* 46.3 (2010), pp. 535–569. ISSN: 0926-6003,1573-2894. DOI: 10.1007/s10589-008-9201-x. URL: https://doi.org/10.1007/s10589-008-9201-x.

[27] M. Hintermüller and A. Laurain. “Optimal shape design subject to elliptic variational inequalities”. In: *SIAM J. Control Optim.* 49.3 (2011), pp. 1015–1047. ISSN: 0363-0129,1095-7138. DOI: 10.1137/080745134. URL: https://doi.org/10.1137/080745134.

[28] M. Hintermüller and D. Korolev. *A hybrid physics-informed neural network based multiscale solver as a partial differential equation constrained optimization problem*. 2024. arXiv: 2309.04439 [math.OC]. URL: https://arxiv.org/abs/2309.04439.

[29] M. Hintermüller and K. Kunisch. “Path-following methods for a class of constrained minimization problems in function space”. In: *SIAM J. Optim.* 17.1 (2006), pp. 159–187. ISSN: 1052-6234,1095-7189. DOI: 10.1137/040611598. URL: https://doi.org/10.1137/040611598.

[30] R. H. W. Hoppe. “Multigrid algorithms for variational inequalities”. In: *SIAM J. Numer. Anal.* 24.5 (1987), pp. 1046–1065. ISSN: 0036-1429. DOI: 10.1137/0724069. URL: https://doi.org/10.1137/0724069.

[31] J. Huang, C. Wang, and H. Wang. “A deep learning method for elliptic hemivariational inequalities”. In: *East Asian J. Appl. Math.* 12.3 (2022), pp. 487–502. ISSN: 2079-7362. DOI: 10.4208/eajam.081121.161121. URL: https://doi.org/10.4208/eajam.081121.161121.

[32] N. V. Hung et al. “Gap functions and error bounds for variational-hemivariational inequalities”. In: *Acta Appl. Math.* 169 (2020), pp. 691–709. ISSN: 0167-8019. DOI: 10.1007/s10440-020-00319-9. URL: https://doi.org/10.1007/s10440-020-00319-9.

[33] K. Ito and K. Kunisch. “Semi-smooth Newton methods for variational inequalities of the first kind”. In: *M2AN Math. Model. Numer. Anal.* 37.1 (2003), pp. 41–62. ISSN: 0764-583X,1290-3841. DOI: 10.1051/m2an:2003021. URL: https://doi.org/10.1051/m2an:2003021.

[34] J. Jiang and X. Chen. “Optimality conditions for nonsmooth nonconvex-nonconcave min-max problems and generative adversarial networks”. In: *SIAM J. Math. Data Sci.* 5.3 (2023), pp. 693–722. ISSN: 2577-0187. DOI: 10.1137/22M1482238. URL: https://doi.org/10.1137/22M1482238.

[35] Y. Jiao et al. “Convergence Analysis of the Deep Galerkin Method for Weak Solutions”. In: *From Classical Analysis to Analysis on Fractals: A Tribute to Robert Strichartz, Volume 1*. Ed. by P. Alonso Ruiz et al. Cham: Springer International Publishing, 2023, pp. 53–82. ISBN: 978-3-031-37800-3. DOI: 10.1007/978-3-031-37800-3_4. URL: https://doi.org/10.1007/978-3-031-37800-3_4.

[36] C. Jin, P. Netrapalli, and M. Jordan. “What is Local Optimality in Nonconvex-Nonconcave Minimax Optimization?” In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 4880–4889. URL: https://proceedings.mlr.press/v119/jin20e.html.

[37] T. Kärkkäinen, K. Kunisch, and P. Tarvainen. “Augmented Lagrangian active set methods for obstacle problems”. In: *J. Optim. Theory Appl.* 119.3 (2003), pp. 499–533. ISSN: 0022-3239,1573-2878. DOI: 10.1023/B:JOTA.0000006687.57272.b6. URL: https://doi.org/10.1023/B:JOTA.0000006687.57272.b6.

[38] B. Keith and T. M. Surowiec. “Proximal Galerkin: A structure-preserving finite element method for pointwise bound constraints”. In: *arXiv e-prints*, arXiv:2307.12444 (July 2023), arXiv:2307.12444. DOI: 10.48550/arXiv.2307.12444. arXiv: 2307.12444 [math.NA].

[39] D. Kinderlehrer and G. Stampacchia. *An introduction to variational inequalities and their applications*. Vol. 88. Pure and Applied Mathematics. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, 1980, pp. xiv+313. ISBN: 0-12-407350-6.

[40] R. Kornhuber. “Monotone multigrid methods for elliptic variational inequalities. I”. In: *Numer. Math.* 69.2 (1994), pp. 167–184. ISSN: 0029-599X,0945-3245. DOI: 10.1007/BF03325426. URL: https://doi.org/10.1007/BF03325426.

[41] I. Lagaris, A. Likas, and D. Fotiadis. “Artificial neural networks for solving ordinary and partial differential equations”. In: *IEEE Transactions on Neural Networks* 9.5 (1998), pp. 987–1000. DOI: 10.1109/72.712178.

[42] T. Larsson and M. Patriksson. “A class of gap functions for variational inequalities”. In: *Math. Programming* 64.1 (1994), pp. 53–79. ISSN: 0025-5610,1436-4646. DOI: 10.1007/BF01582565. URL: https://doi.org/10.1007/BF01582565.

[43] Q. Li, T. Lin, and Z. Shen. “Deep learning via dynamical systems: An approximation perspective”. In: *Journal of the European Mathematical Society* 25.5 (2022), pp. 1671–1709.

[44] H. Lin and S. Jegelka. “ResNet with one-neuron hidden layers is a Universal Approximator”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018. URL: https://proceedings.neurips.cc/paper_files/paper/2018/file/03bfc1d4783966c69cc6aef8247e0103-Paper.pdf.

[45] C. LIU, E. Liang, and M. Chen. *Characterizing ResNet’s Universal Approximation Capability*. 2024. URL: https://openreview.net/forum?id=PCTqol2hvy.

[46] I. Loshchilov and F. Hutter. “Decoupled Weight Decay Regularization”. In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

[47] P. Madhyastha and R. Jain. “On Model Stability as a Function of Random Seed”. In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Ed. by M. Bansal and A. Villavicencio. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 929–939. DOI: 10.18653/v1/K19-1087. URL: https://aclanthology.org/K19-1087.

[48] C. Meyer and O. Thoma. “A priori finite element error analysis for optimal control of the obstacle problem”. In: *SIAM J. Numer. Anal.* 51.1 (2013), pp. 605–628. ISSN: 0036-1429,1095-7170. DOI: 10.1137/110836092. URL: https://doi.org/10.1137/110836092.

[49] A. Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[50] M. Raissi, P. Perdikaris, and G. Karniadakis. “Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations”. In: *Journal of Computational Physics* 378 (2019), pp. 686–707. ISSN: 0021-9991. DOI: https://doi.org/10.1016/j.jcp.2018.10.045. URL: https://www.sciencedirect.com/science/article/pii/S0021999118307125.

[51] M. Razaviyayn et al. “Nonconvex min-max optimization: Applications, challenges, and recent theoretical advances”. In: *IEEE Signal Processing Magazine* 37.5 (2020), pp. 55–66.

[52] J.-F. Rodrigues. *Obstacle problems in mathematical physics*. Vol. 134. North-Holland Mathematics Studies. Notas de Matemática [Mathematical Notes], 114. North-Holland Publishing Co., Amsterdam, 1987, pp. xvi+352. ISBN: 0-444-70187-7.

[53] C. Schwab and A. Stein. “Deep solution operators for variational inequalities via proximal neural networks”. In: *Res. Math. Sci.* 9.3 (2022), Paper No. 36, 35. ISSN: 2522-0144,2197-9847. DOI: 10.1007/s40687-022-00327-1. URL: https://doi.org/10.1007/s40687-022-00327-1.

[54] Y. Shin, Z. Zhang, and G. E. Karniadakis. “Error estimates of residual minimization using neural networks for linear PDEs”. In: *Journal of Machine Learning for Modeling and Computing* 4.4 (2023), pp. 73–101. ISSN: 2689-3967.

[55] M. Sofonea and A. Matei. *Variational inequalities with applications*. Vol. 18. Advances in Mechanics and Mathematics. A study of antiplane frictional contact problems. Springer, New York, 2009, pp. xx+230. ISBN: 978-0-387-87459-3.

[56] N. Sukumar and A. Srivastava. “Exact imposition of boundary conditions with distance functions in physics-informed deep neural networks”. In: *Computer Methods in Applied Mechanics and Engineering* 389 (2022), p. 114333. ISSN: 0045-7825. DOI: https://doi.org/10.1016/j.cma.2021.114333. URL: https://www.sciencedirect.com/science/article/pii/S0045782521006186.

[57] P. Valsecchi Oliva et al. “Towards fast weak adversarial training to solve high dimensional parabolic partial differential equations using XNODE-WAN”. In: *J. Comput. Phys.* 463 (2022), Paper No. 111233, 17. ISSN: 0021-9991,1090-2716. DOI: 10.1016/j.jcp.2022.111233. URL: https://doi.org/10.1016/j.jcp.2022.111233.

[58] F. Wang, W. Han, and X.-L. Cheng. “Discontinuous Galerkin methods for solving elliptic variational inequalities”. In: *SIAM J. Numer. Anal.* 48.2 (2010), pp. 708–733. ISSN: 0036-1429,1095-7170. DOI: 10.1137/09075891X. URL: https://doi.org/10.1137/09075891X.

[59] J. Wloka. *Partial differential equations*. Translated from the German by C. B. Thomas and M. J. Thomas. Cambridge University Press, Cambridge, 1987, pp. xii+518. DOI: 10.1017/CBO9781139171755. URL: https://doi.org/10.1017/CBO9781139171755.

[60] Z. Wu, C. Shen, and A. van den Hengel. “Wider or Deeper: Revisiting the ResNet Model for Visual Recognition”. In: *Pattern Recognition* 90 (2019), pp. 119–133. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2019.01.006. URL: https://www.sciencedirect.com/science/article/pii/S0031320319300135.

[61] N. Yamashita and M. Fukushima. “Equivalent unconstrained minimization and global error bounds for variational inequality problems”. In: *SIAM J. Control Optim.* 35.1 (1997), pp. 273–284. ISSN: 0363-0129. DOI: 10.1137/S0363012994277645. URL: https://doi.org/10.1137/S0363012994277645.

[62] Y. Zang et al. “Weak adversarial networks for high-dimensional partial differential equations”. In: *J. Comput. Phys.* 411 (2020), pp. 109409, 14. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2020.109409. URL: https://doi.org/10.1016/j.jcp.2020.109409.

[63] E. Zeidler. *Nonlinear functional analysis and its applications. I*. Fixed-point theorems, Translated from the German by Peter R. Wadsack. Springer-Verlag, New York, 1986, pp. xxi+897. ISBN: 0-387-90914-1. DOI: 10.1007/978-1-4612-4838-5. URL: https://doi.org/10.1007/978-1-4612-4838-5.

[64] X. E. Zhao, W. Hao, and B. Hu. “Two neural-network-based methods for solving elliptic obstacle problems”. In: *Chaos Solitons Fractals* 161 (2022), Paper No. 112313, 10. ISSN: 0960-0779. DOI: 10.1016/j.chaos.2022.112313. URL: https://doi.org/10.1016/j.chaos.2022.112313.