Available online at www.sciencedirect.com

ScienceDirect

Comput. Methods Appl. Mech. Engrg. 402 (2022) 115716

Computer methods in applied mechanics and engineering

www.elsevier.com/locate/cma

Neural control of discrete weak formulations: Galerkin, least squares & minimal-residual methods with quasi-optimal weights[☆]

Ignacio Brevis[a], Ignacio Muga[a,b], Kristoffer G. van der Zee[c,*]

[a] *Instituto de Matemáticas, Pontificia Universidad Católica de Valparaíso, Chile*
[b] *Basque Center for Applied Mathematics (BCAM), Spain*
[c] *School of Mathematical Sciences, University of Nottingham, UK*

Available online 12 November 2022

Abstract

There is tremendous potential in using neural networks to optimize numerical methods. In this paper, we introduce and analyze a framework for the *neural optimization of discrete weak formulations*, suitable for finite element methods. The main idea of the framework is to include a neural-network function acting as a *control* variable in the weak form. Finding the neural control that (quasi-) minimizes a suitable cost (or loss) functional, then yields a numerical approximation with desirable attributes. In particular, the framework allows in a natural way the incorporation of known data of the exact solution, or the incorporation of stabilization mechanisms (e.g., to remove spurious oscillations).

The main result of our analysis pertains to the well-posedness and convergence of the associated constrained-optimization problem. In particular, we prove under certain conditions, that the discrete weak forms are stable, and that quasi-minimizing neural controls exist, which converge quasi-optimally. We specialize the analysis results to Galerkin, least squares and minimal-residual formulations, where the neural-network dependence appears in the form of suitable weights. Elementary numerical experiments support our findings and demonstrate the potential of the framework.

Keywords: Optimal neural control; Artificial neural networks; Data-driven discretization; Weighted finite element methods; Quasi-minimization; Quasi-optimal convergence

1. Introduction

In recent years there has been tremendous interest in the merging of neural networks and machine-learning algorithms with traditional methods in scientific computing and computational science [1–4]. In this paper we demonstrate how neural networks can be utilized to optimize finite element methods. Let us first provide an elementary motivation as to why finite element methods would benefit from being optimized at all.

In one of its most familiar mathematical forms, the finite element method is a discretization technique for partial differential equations (PDEs) based on a weak formulation using discrete subspaces, i.e., the exact solution $u \in \mathbb{U}$ is approximated by $u_h \in \mathbb{U}_h$, which is the unique solution of the discrete problem:

Find $u_h \in \mathbb{U}_h$:

[☆] Dedicated to J. Tinsley Oden: An unprecedented *homo universalis*.

[*] Corresponding author.
E-mail address: kg.vanderzee@nottingham.ac.uk (K.G. van der Zee).

https://doi.org/10.1016/j.cma.2022.115716

$$b(u_h, v_h) = f(v_h)\,, \qquad \forall v_h \in \mathbb{V}_h\,, \tag{1}$$

where $\mathbb{U}_h$ is a discrete subspace of the infinite-dimensional Hilbert or Banach space $\mathbb{U}$ (typically a Sobolev space on a domain $\Omega \subset \mathbb{R}^d$), $\mathbb{V}_h$ is a subspace of a Hilbert or Banach space $\mathbb{V}$ with $\dim \mathbb{V}_h = \dim \mathbb{U}_h$, $b : \mathbb{U} \times \mathbb{V} \to \mathbb{R}$ is a continuous bilinear form, $f : \mathbb{V} \to \mathbb{R}$ a continuous linear form, and the exact solution $u \in \mathbb{U}$ satisfies $b(u, v) = f(v)$ for all $v \in \mathbb{V}$.[1]

It is well-known that the accuracy of u_h can be improved by enlarging $\mathbb{U}_h$ (e.g., by refining the underlying finite element mesh).[2] However, *for a fixed value of* h, the particular u_h defined by (1) may be very *unsatisfactory*. In fact, there is no reason why a certain quantity of interest of u_h is accurate at all,[3] or why the approximation inherits certain qualitative features of the exact solution.[4] Indeed, the discrete problem (1) is a *rigid* statement in the sense that it identifies a *single* element in $\mathbb{U}_h$, irrespective of desired attributes, whereas there could be many other elements in $\mathbb{U}_h$ that are far superior.

1.1. Neural optimization of discrete weak forms

The objective of this work is to propose and analyze a framework for the *neural optimization of discrete weak formulations* to significantly improve quantitative and qualitative attributes of discrete approximations. In particular, we consider *Galerkin*, *least squares*, and *minimal-residual* formulations.

The main idea of the framework is that it incorporates a neural-network function ξ as a control variable in the discrete test space $\mathbb{V}_h(\xi)$. That is, the approximation $u_h = u_{h,\xi}$ now depends on ξ and solves the discrete problem:

$$\begin{aligned}&\text{Find } u_h = u_{h,\xi} \in \mathbb{U}_h :\\ &\qquad b(u_{h,\xi}, v_h) = f(v_h)\,, \qquad \forall v_h \in \mathbb{V}_h(\xi)\,.\end{aligned} \tag{2}$$

Then, in order to obtain a desired approximation $u_{h,\bar{\xi}}$, we aim to find a neural-network function $\bar{\xi}$ that *quasi*-minimizes a desired cost (or loss) functional[5]:

$$J(u_{h,\bar{\xi}}) \longrightarrow \text{quasi-min}\,. \tag{3}$$

The notion of *quasi-minimization* is critical when aiming to minimize over a (non-closed) set of neural-network functions (i.e., the set of functions implemented by neural networks of a fixed architecture); see Section 2.2 for further details (in particular, Definitions 2.1 and 2.2).

The quasi-minimization problem (3) is essentially a *nonstandard* PDE-constrained optimization, with the nonstandard part being the dependence of the state problem (2) on ξ via the discrete test space $\mathbb{V}_h(\xi)$. Importantly, $\mathbb{V}_h(\xi)$ will be parameterized by ξ in such a way so as to ensure *stability* of the discrete problem (2), as well as imply existence and convergence of corresponding quasi-minimizers of (3). Moreover, as will become clear in the following sections, the basis functions in $\mathbb{V}_h(\xi)$ need not be computed explicitly, but equivalent formulations to (2) can be used, which instead incorporate ξ by means of suitable *weight functions*. These formulations essentially lead to a PDE-constrained optimization with a nonlinear control-to-state map.

1.2. Potential of the methodology

There are two main benefits of having neural control of discrete weak forms:

- *Incorporation of data*: Knowledge of quantities of the exact solution can be taken into account in a natural way by setting, for example,

$$J(u_{h,\xi}) = \frac{1}{2}\left|q(u_{h,\xi}) - \bar{q}\right|^2,$$

[1] When $\mathbb{U}_h = \mathbb{V}_h$, this is a *Galerkin* method, otherwise it is a *Petrov–Galerkin* method.

[2] Indeed, a priori error analysis reveals that $\|u - u_h\|_{\mathbb{U}} \le C \inf_{w_h \in \mathbb{U}_h} \|u - w_h\|_{\mathbb{U}}$, provided $b(\cdot,\cdot)$ satisfies a discrete inf–sup condition on $\mathbb{U}_h \times \mathbb{V}_h$; see e.g., [5,6].

[3] E.g., the value $u_h(x_0)$ for some point $x_0 \in \Omega$ is generally quite distinct from $u(x_0)$.

[4] E.g., u_h may exhibit spurious oscillations, while u is monotone.

[5] We also allow for the inclusion of a regularization term in the cost functional; see Section 2.1.

where $q : \mathbb{U} \to \mathbb{R}$ is a functional measuring the quantity of interest and $\bar{q} \in \mathbb{R}$ is known data.[6] Minimizing such a $J(\cdot)$ ensures that the discrete solution u_h to (2) is *data-driven* in the sense that u_h becomes constrained by the data.[7] We note that multiple quantities can be taken into account using, for example,

$$J(u_{h,\xi}) = \frac{1}{N_{\text{data}}} \sum_{i=1}^{N_{\text{data}}} \frac{1}{2} \left| q_i(u_{h,\xi}) - \bar{q}_i \right|^2,$$

or, more generally, using some operator $Q : \mathbb{U} \to \mathbb{Z}$; see Section 2.

- *Incorporation of stabilization mechanisms*: Qualitative attributes of the discrete solution can be enhanced by minimizing a suitably-chosen $J(\cdot)$. In this way, discrete solutions can be enforced to, e.g., satisfy an a priori known maximum principle, have monotone (or spurious oscillation free) behavior around discontinuities and layers, or have a certain discrete wave number (i.e., free from pollution). In the past decades, many different stabilized finite element methods have been proposed (and analyzed) that impose such attributes [8–13]. Within our framework, such a method is naturally obtained after (quasi-) minimization (i.e., method (2) with $\xi = \bar{\xi}$). As an example, Guermond [8] advocates the L^1-minimization of the residual; in other words, within our framework, one would choose:

$$J(u_{h,\xi}) = \left\| f - Bu_{h,\xi} \right\|_{L^1(\Omega)},$$

where $f - Bu_{h,\xi}$ is the strong form of the residual.

The idea of using neural networks to parameterize the test space was initially proposed in our earlier work [14], where it was restricted to minimal-residual formulations within a parametric PDE setting. The current work presents significantly more general settings and formulations as well as analyses of their well-posedness and convergence.

While the above shows examples of $J(\cdot)$ corresponding to *unsupervised* learning (i.e., there is no need to know the exact solution u), when the original problem is *parametric* itself (e.g., a parametric PDE), *supervised* learning becomes meaningful. Indeed, in that case, the data may be the exact solution u_{λ_i} for certain parameters λ_i, $i = 1, \dots, N_{\text{data}}$. This then allows for the training of finite element discretizations with superior accuracy in quantities of interest even on very coarse meshes. We refer to our earlier work [14] for the methodology and illustrative examples in that case.

1.3. Main contributions: Well-posedness, convergent quasi-minimizers, weighted conforming formulations

Let us briefly outline the main contributions of this work. The first main contribution is the analysis of an abstract constrained-optimization problem associated to (3); see Section 2. In particular, we consider an abstract state problem equivalent to (2), but in the form of a *mixed* system with a ξ-dependent bilinear form.[8] We prove, under suitable conditions, that the state problem is *well-posed* (uniformly with respect to ξ); see Proposition 2.9. Furthermore, we present differentiability conditions (on the ξ-dependence) that allow us to prove the *existence* of quasi-minimizers (within sets of neural-network functions, of some size n) to the associated constrained optimization (3), which converge *quasi-optimally* (upon $n \to \infty$); see Corollary 2.12 for details.

We note that our analysis is based on a fundamental result for the quasi-minimization of strongly-convex and differentiable functionals (see Theorem 2.A), which is of independent interest and applies, e.g., to the analysis of deep Ritz methods [17–19] and PINN methods [20–22].

The second main contribution of this work is the application of our framework to certain weak formulations used by conforming finite element methods; see Section 3. In these applications, the neural-network control variable ξ will appear by means of suitable weights in the bilinear forms. In particular, we will analyze weighted least squares, weighted Galerkin, and weighted minimal-residual formulations.

For weighted least squares and weighted minimal-residual formulations, suitable conditions on the weights imply (via the abstract result of the first main contribution) stability of the discrete problem (uniformly in ξ). Furthermore,

[6] The data $\bar{q}$ represents $q(u)$, and it could be obtained through experiments, high-fidelity computation, or otherwise.

[7] This is somewhat similar in spirit to *physics-informed neural networks* (PINN) [7], where however a single neural-network function minimizes a combination of the residual and data misfit.

[8] The mixed system is motivated by residual-minimization theory [15,16]: Minimal residual formulations are equivalent to mixed systems, which in turn are equivalent to Petrov–Galerkin formulations.

suitable differentiability conditions on the weights imply existence of (quasi-optimally) convergent quasi-minimizers of the associated constrained minimization.

On the other hand, for weighted Galerkin, it turns out that stability is *not* immediate, and may require constraints on ξ depending on the problem at hand.[9] Therefore, neural control is far more convenient for least squares and minimal-residual formulations, the fundamental reason being the inherent stability that comes with their underlying minimization principle.

We support our findings with numerical experiments in Section 4. While our theoretical results directly apply to any linear operator, we choose the advection–reaction PDE to illustrate various numerical aspects, viz., the incorporation of data (Section 4.1), the quasi-optimal convergence of quasi-minimizers (Section 4.2), and the incorporation of L^1-type stabilization (Section 4.3).

1.4. Related work

There are a number of works related to ours.

Optimizing numerical methods: Traditionally, the incorporation of known data or other desired attributes in numerical PDE approximations is achieved via the method of Lagrange multipliers, see e.g., Evans, Hughes & Sangalli [11], Kergrene, Prudhomme, Chamoin & Laforest [23], and references therein. The classical idea of optimizing Petrov–Galerkin methods by using special test spaces was originally motivated by the desire to obtain stable methods for nonsymmetric problems. It has a long history, with early work going back to, e.g., Brooks & Hughes [24], Barrett & Morton [25], and Oden [26], which have given rise to modern stabilized, minimal-residual and DPG methods.

Optimizing numerical methods using neural networks: Much more recent is the use of neural networks for the optimization of numerical methods, i.e., for the learning of parameters that define a numerical method; see Ray & Hesthaven [27], Mishra [28] and others [29–33]. Interestingly, a learning methodology for optimizing the anisotropy of the finite element mesh has been proposed by Fidkowski & Chen [34], while a learning methodology for adaptive mesh refinement that ensures optimal adaptive convergence has been analyzed by Bohn & Feischl [35]. Within the context of optimizing finite-element formulations, a minimal-residual framework that is guaranteed to be stable was proposed in our previous work [14]. Our current work contributes to these developments by providing the analysis of a general framework for neural network optimization of finite element methods.

Neural networks for PDEs: The use of neural networks for approximating directly the solution to PDEs has received wide-spread interest since the works by E & Yu [36], Sirignano & Spiliopoulos [37], Berg & Nyström [38] and Raissi, Perdikaris & Karniadakis [7], amongst others. Recently, there have been a number of ideas that propose an adaptive construction of neural-network approximations; see Ainsworth & Dong [39], Liu, Cai & Chen [40] and Uriarte, Pardo & Omella [41]. Neural networks can also be used to obtain the coefficients of the basis expansion used by a standard (linear) approximation [42,43].

Neural networks for inverse PDEs: In the context of inverse problems involving PDEs, the use of neural networks to represent unknown PDE coefficients (fields) and constitutive models has been explored by, e.g., Teichert, Natarajan, Van der Ven & Garikipati [44], Berg & Nyström [45], Xu & Darve [46] and You et al. [47]. These works are similar to the current work in the sense that standard (finite element) methods are used to solve the PDE, while a neural network is embedded within the discrete formulation. We note that the analysis provided by our current work can be extended to those inverse problems. Other works involve the use of a neural network to approximate the parameter-to-solution map (so-called neural operators); see e.g., [41,48,49] and references therein. These approximations are particularly useful for large-scale problems for which model reduction is essential.

Error analysis for neural-network approximations: There are a number of works containing a priori error analysis for neural-network based PDE approximations. For those related to the deep Ritz method; see Xu [17, Section 5], Pousin [18, Section 3], and Müller & Zeinhofer [19]. For those related to physics-informed neural networks (PINN) and least squares methods; see Sirignano & Spiliopoulos [37, Section 7], Mishra & Molinaro [21,50], Pousin [18, Section 4], Cai, Chen & Liu [22], and Berrone, Canuto & Pintore [51]. Recently, a posteriori error analysis has also been studied, in particular goal-oriented analysis using the dual-weighted residual (DWR) methodology; see, e.g., Roth, Schröder & Wick [52], and Minakowski & Richter [53]. We note that in our current work, while

[9] In essence, the reason for instability relates to a discrete inf–sup condition of a weighted bilinear form.

we have in mind the error analysis for neural-control approximations, the abstract analysis presented in Section 2 is essentially an extension of the above-mentioned a priori analysis to a certain class of problems involving a convex and differentiable cost functional.

2. Abstract framework

In this section we present the analysis of the abstract state equation (in the form of a mixed system) and the associated optimization problem. We essentially follow the classical theory of optimal control (PDE-constrained optimization) by Lions [54]; see also, [55–57]. Our resulting optimization problem bears similarity to that of parameter identification of PDE coefficients; see Rannacher & Vexler [58] and references therein for its error analysis. While we present our abstract framework within Hilbert spaces (and using a quadratic cost), we note that extensions to Banach spaces are feasible, but not within the scope of the current work.

2.1. Discrete state problem and associated cost functional

Let $\mathbb{X}$ be a Hilbert space for the control variable. We shall think of a control variable $\xi \in \mathbb{X}$ as being a function in an infinite-dimensional function space $\mathbb{X}$ (for example, $\mathbb{X} = L^2(\Omega)$).[10] Let $\mathbb{U}$ and $\mathbb{V}$ be Hilbert spaces for trial and test functions, respectively, $\mathbb{U}_h \subset \mathbb{U}$ be a discrete (finite element) subspace, and $\hat{\mathbb{V}} \subseteq \mathbb{V}$.[11] In all that follows, we think of h (hence $\mathbb{U}_h$) as being fixed. Given $\xi \in \mathbb{X}$ and $f \in \mathbb{V}^*$ (the dual of $\mathbb{V}$), we consider the discrete state problem given by:

$$\begin{cases} \text{Find } (r, u_h) \in \hat{\mathbb{V}} \times \mathbb{U}_h : & & \\ \quad a(\xi; r, v) + b(u_h, v) = f(v), & \forall v \in \hat{\mathbb{V}}, & \text{(a)} \\ \quad b(w_h, r) \qquad\qquad\quad = 0, & \forall w_h \in \mathbb{U}_h\,, & \text{(b)} \end{cases} \tag{4}$$

where $b(\cdot,\cdot)$ is a continuous bilinear form on $\mathbb{U} \times \mathbb{V}$, and for each $\xi \in \mathbb{X}$, $a(\xi;\cdot,\cdot)$ is a continuous bilinear form on $\mathbb{V} \times \mathbb{V}$. To explicitly indicate the dependence of r and u_h on ξ, we use the notation:

$$(r_\xi, u_{h,\xi}) = \text{solution of (4)(a)–(4)(b) for a given } \xi\,.$$

In Section 2.4, we demonstrate that (4)(a)–(4)(b) is equivalent to (2) for a particular choice of $\mathbb{V}_h(\xi)$; see Proposition 2.10. The discrete problem in (4)(a)–(4)(b) is essentially a general formulation, which for a specific choice of $a(\cdot\,;\cdot,\cdot)$ and $\hat{\mathbb{V}}$ reduces to a (weighted) Galerkin, least-squares or minimal residual method; see Section 3.

Next, let $\mathbb{Z}$ be a Hilbert space, and let $Q : \mathbb{U} \to \mathbb{Z}$ be a linear continuous (observation) operator. Then, given an observation $z_o \in \mathbb{Z}$ and regularization parameter $\alpha \geq 0$, we consider the cost (or loss) functional $J : \mathbb{U}_h \times \mathbb{X} \to \mathbb{R}$ defined by:

$$J(w_h, \xi) := J_1(w_h) + \alpha\, j_2(\xi)\,, \tag{5}$$

where

$$J_1(w_h) := \frac{1}{2}\left\| Q(w_h) - z_o \right\|^2_{\mathbb{Z}}\,,$$
$$j_2(\xi) := \frac{1}{2}\|\xi\|^2_{\mathbb{X}}\,.$$

The associated *reduced* cost functional $j : \mathbb{X} \to \mathbb{R}$ is then given by:

$$j(\xi) := j_1(\xi) + \alpha\, j_2(\xi)\,, \tag{6}$$

where $j_1 : \mathbb{X} \to \mathbb{R}$ is defined by:

$$j_1(\xi) := J_1(u_{h,\xi}) = \frac{1}{2}\left\| Q(u_{h,\xi}) - z_o \right\|^2_{\mathbb{Z}}\,,$$

While ideally we would like to minimize $j(\cdot)$ over (the infinite-dimensional) $\mathbb{X}$, we proceed by considering neural-network approximations.

[10] In Section 2.2, we let ξ be a neural network function.

[11] Later on, when considering minimal residual formulations, $\hat{\mathbb{V}}$ will be a discrete (finite element) subspace of $\mathbb{V}$, but for the other formulations $\hat{\mathbb{V}} = \mathbb{V}$.

2.2. Neural quasi-minimization

To accommodate neural optimization, we consider the subset $\mathcal{M}_n \subset \mathbb{X}$ consisting of all functions implemented by neural networks of a fixed architecture parameterized by n.[12] We shall simply refer to $\mathcal{M}_n$ as a set of *neural-network functions*, and we think of n as a measure of the size of the architecture (e.g., the total number of neurons, or total number of parameters).

When aiming to minimize $j(\cdot)$, a significant complication is that the set $\mathcal{M}_n$ is not a linear subspace and it *may not be closed* (topologically) in $\mathbb{X}$.[13] Hence, even though $j(\cdot)$ may have an infimum on $\mathcal{M}_n$, there may not be a minimizer in $\mathcal{M}_n$. Therefore, one should not aim to *completely* minimize $j(\cdot)$, but instead use a relaxed notion of *quasi-minimization* as used by Shin, Zhang & Karniadakis [20][14] (for which the existence of an infimum implies the existence of a quasi-minimizer):

Definition 2.1 (*Quasi-Minimizers and Quasi-Minimizing Sequences*)**.** Let $j : \mathbb{X} \to \mathbb{R}$ be a cost functional.

(i) Let $\delta_n > 0$ and $\mathcal{M}_n \subset \mathbb{X}$ be a subset of $\mathbb{X}$ (not necessarily closed in $\mathbb{X}$). A function $\bar{\xi}_n \in \mathcal{M}_n$ is said to be a *quasi-minimizer* of $j(\cdot)$ if the following holds true[15]:

$$j(\bar{\xi}_n) \leq \inf_{\xi_n \in \mathcal{M}_n} j(\xi_n) + \frac{\delta_n}{2}. \tag{7}$$

(ii) Consider a sequence of subsets $(\mathcal{M}_n)_{n\in\mathcal{N}}$ of $\mathbb{X}$, with $\mathcal{N}$ being a strictly-increasing sequence of natural numbers. A sequence $(\bar{\xi}_n)_n$, with $\bar{\xi}_n \in \mathcal{M}_n$, is said to be a *quasi-minimizing sequence* if (7) holds true for all $n \in \mathcal{N}$ with $\delta_n > 0$ such that:

$$\delta_n \to 0 \quad \text{as} \quad n \to \infty. \quad \square$$

In summary, the neural optimization problem that we consider is the following:

Definition 2.2 (*The Quasi-Minimizing Control Problem*)**.** The following statements are equivalent.
Reduced quasi-minimizing control problem: For $j(\cdot)$ given by (6), we aim to quasi-minimize $j(\cdot)$, i.e., given $\delta_n > 0$,

$$\begin{cases} \text{Find } \bar{\xi}_n \in \mathcal{M}_n : \\ \quad j(\bar{\xi}_n) \leq \inf_{\xi_n \in \mathcal{M}_n} j(\xi_n) + \dfrac{\delta_n}{2}. \end{cases} \tag{8}$$

Constrained quasi-minimizing control problem: For $J(\cdot,\cdot)$ given by (5), we aim to quasi-minimize $J(u_h, \xi)$ subject to (4)(a)–(4)(b), i.e., given $\delta_n > 0$,

$$\begin{cases} \text{Find } \bar{\xi}_n \in \mathcal{M}_n : \\ \quad J(u_{h,\bar{\xi}_n}, \xi_n) \leq \inf_{\xi_n \in \mathcal{M}_n} J(u_{h,\xi_n}, \eta_n) + \dfrac{\delta_n}{2}. \end{cases} \quad \square \tag{9}$$

Example 2.3 (*Need for Quasi-Minimizers*)**.** Let us discuss a simple example illustrating the non-existence of minimizers, hence the need for quasi-minimizers.[16]

[12] In the terminology of Petersen, Raslan and Voigt [59], the set $\mathcal{M}_n$ consists of the *realizations* of all possible neural networks of some fixed architecture (and some given activation function). While a neural network is identified with the set of weight and bias parameters, its realization is the *function* implemented by the network.

[13] For example, [59, Theorem 3.1] shows that, under mild conditions on the architecture and activation function, $\mathcal{M}_n$ is not a closed subset of $L^2(\Omega)$ (or, more generally, $L^p(\Omega)$, with $0 < p < \infty$), unless, e.g., an upper bound is imposed on the weight parameters [59, Proposition 3.7].

[14] Quasi-minimization can also be thought of as solving the minimization problem up to some optimization accuracy, cf. [19].

[15] Observe that if $j(\cdot)$ has an infimum on $\mathcal{M}_n$, then immediately a quasi-minimizer exists (in $\mathcal{M}_n$). This is true simply by the definition of the infimum.

[16] This is essentially an example of a PINN problem, i.e., minimizing a strong residual and boundary condition in least-squares sense. It is not difficult to construct a similar example for a neural control problem.

Let $x = (x_1, x_2) \in \Omega = (0,1)^2 \subset \mathbb{R}^2$. Given $z \in (0,1)$, let $\chi_{[z,1]}$ denote the characteristic function of the subset $[z, 1]$.[17] Consider the following cost functional:

$$j(\xi) = \frac{1}{2}\int_0^1 \int_0^1 \left(\frac{\partial \xi}{\partial x_2}\right)^2 \mathrm{d}x_1\, \mathrm{d}x_2 + \int_0^1 \Big(\xi(x_1, 0) - \chi_{[z,1]}(x_1)\Big)^2 \mathrm{d}x_1$$

for $\xi \in \mathbb{X} = \Big\{\eta \in L^2(\Omega) \,\Big|\, \frac{\partial\eta}{\partial x_2} \in L^2(\Omega)\Big\}$. Minimizing $j(\cdot)$ over $\mathbb{X}$ solves a first-order PDE (constant advection in the direction of the x_2-axis) with discontinuous data given by $\chi_{[z,1]}$, which is a well-posed problem [60, Section 1 and 6].

Let $\mathcal{M}_n$ be the set of two-layer neural-network functions $\Omega \mapsto \mathbb{R}^2 \mapsto \mathbb{R}$ using two neurons and the *Rectified Linear Unit* $\mathrm{ReLU}(\cdot) := \max\{0, \cdot\}$ activation function in the hidden layer, i.e.,

$$\mathcal{M}_n = \left\{\xi_n : \Omega \to \mathbb{R} \,\middle|\, \xi_n(x) = \sum_{i=1}^{2} a_i \,\mathrm{ReLU}(w_i \cdot x - b_i)\,,\ a_i, b_i \in \mathbb{R},\, w_i \in \mathbb{R}^2\right\}.$$

Note that an infimizing sequence of $j(\cdot)$ in $\mathcal{M}_n$ is given by:

$$\xi_m(x) = \begin{cases} 0 & 0 \le x_1 < z_m := (1 - \frac{1}{m})z\,, \\ \dfrac{x_1 - z_m}{z - z_m} & z_m \le x_1 < z\,, \\ 1 & z \le x_1 \le 1\,, \end{cases}$$

for $m = 1, 2, 3, \ldots$, but whose limit $\xi_m \to \bar{\xi}$ in $\mathbb{X}$ as $m \to \infty$ is a *discontinuous* function (with $j(\bar{\xi}) = 0$). Therefore the infimizer $\bar{\xi}$ does *not* exist in $\mathcal{M}_n \subset C(\overline{\Omega})$.

On the other hand, quasi-minimizers $\bar{\xi}_n$ do exist in $\mathcal{M}_n$, in particular, ξ_m as defined above is a quasi-minimizer for m large enough.[18] □

2.3. Analysis of reduced control problem

We first proceed with the analysis of the reduced control problem (8). Let the state operators $R_h : \mathbb{X} \to \hat{\mathbb{V}}$ and $S_h : \mathbb{X} \to \mathbb{U}_h$ be defined by:

$$R_h(\xi) := r_{h,\xi}\,, \qquad \forall \xi \in \mathbb{X}\,, \tag{10a}$$

$$S_h(\xi) := u_{h,\xi}\,, \qquad \forall \xi \in \mathbb{X}\,, \tag{10b}$$

where $r_{h,\xi}$ and $u_{h,\xi}$ are the first and second component, respectively, of the solution to the mixed system (4). Then, the reduced cost $j(\cdot)$ given in (6) can be written as follows:

$$\begin{aligned} j(\xi) &= j_1(\xi) + \alpha\, j_2(\xi) = J_1\big(S_h(\xi), \xi\big) + \alpha j_2(\xi) \\ &= \frac{1}{2}\big\| Q \circ S_h(\xi) - z_o \big\|_{\mathbb{Z}}^2 + \frac{\alpha}{2}\|\xi\|_{\mathbb{X}}^2\,. \end{aligned} \tag{11}$$

Our main result depends on the following fundamental theorem, which is of independent interest:

Theorem 2.A (*Differentiable, Strongly-Convex Quasi-Minimization*). *Let $j : \mathbb{X} \to \mathbb{R}$ be a cost functional. Assume that $j(\cdot)$ is Gâteaux differentiable with derivative $j' : \mathbb{X} \to \mathbb{X}^*$ being Lipschitz continuous, i.e., there is a constant $L > 0$ such that*

$$\big\| j'(\xi) - j'(\eta)\big\|_{\mathbb{X}^*} \le L \big\|\xi - \eta\big\|_{\mathbb{X}}\,, \qquad \forall \xi, \eta \in \mathbb{X}\,,$$

Furthermore, assume that $j(\cdot)$ is strongly convex, i.e., there is a constant $\gamma > 0$ such that

$$\Big\langle j'(\xi) - j'(\eta)\,,\, \xi - \eta\Big\rangle_{\mathbb{X}^*,\mathbb{X}} \ge \gamma \big\|\xi - \eta\big\|_{\mathbb{X}}^2\,, \qquad \forall \xi, \eta \in \mathbb{X}\,. \tag{12}$$

[17] That is, $\chi_{[z,1]}(x_1) = 1$ if $x_1 \in [z, 1]$ and $= 0$ otherwise.

[18] Indeed, one can verify by direct calculation that m must be such that $\frac{1}{3}(z - z_m) \le \frac{\delta_n}{2}$, i.e., $m \ge \frac{2}{3} z \delta_n^{-1}$.

Then, the following hold true:

(i) $j(\cdot)$ *has a unique minimizer* $\bar{\xi} \in \mathbb{X}$, *which satisfies:*

$$j'(\bar{\xi}) = 0 \qquad \text{in } \mathbb{X}^* .$$

(ii) For any subset $\mathcal{M}_n \subset \mathbb{X}$, $j(\cdot)$ *has a quasi-minimizer* $\bar{\xi}_n \in \mathcal{M}_n$ *that satisfies* (7).

(iii) Any quasi-minimizer $\bar{\xi}_n$ *in* $\mathcal{M}_n$ *satisfies the following quasi-optimal error estimate:*

$$\|\bar{\xi} - \bar{\xi}_n\|_{\mathbb{X}} \leq \left(\frac{L}{\gamma} \inf_{\xi_n \in \mathcal{M}_n} \|\bar{\xi} - \xi_n\|^2_{\mathbb{X}} + \frac{\delta_n}{\gamma} \right)^{1/2} . \quad \square \tag{13}$$

Proof. See Appendix A.1. ■

We now analyze when our $j(\cdot)$ satisfies the assumptions of Theorem 2.A.

Theorem 2.B (*Reduced Control Problem: Differentiability & Strong Convexity*)**.** *Let* $\alpha > 0$ *and* $j(\cdot) = j_1(\cdot) + \alpha\, j_2(\cdot)$ *be as in* (11). *Let* $Q \in \mathcal{L}(\mathbb{U}, \mathbb{Z})$. *Assume* $S_h : \mathbb{X} \to \mathbb{U}_h$ *is differentiable,* $S_h(\cdot)$ *and* $S'_h(\cdot)$ *are uniformly bounded on* $\mathbb{X}$, *and* $S'_h(\cdot)$ *is Lipschitz continuous. Then:*

(i) $j_1, j_2, j : \mathbb{X} \to \mathbb{R}$ *are Gâteaux differentiable with* $j'_1, j'_2, j' : \mathbb{X} \to \mathbb{X}^*$ *Lipschitz continuous.*

Additionally, assume α *is sufficiently large. Then:*

(ii) $j : \mathbb{X} \to \mathbb{R}$ *is strongly convex, i.e., there is a constant* $\gamma > 0$ *such that* (12) *holds true.*[19] □

Proof. See Appendix A.2. ■

Corollary 2.4 (*Reduced Control Problem: (quasi-)Minimizers & Quasi-Optimality*)**.** *Under the conditions of Theorem* 2.B, *the statements (i), (ii) and (iii) of Theorem* 2.A *hold true.* □

Proof. The results of Theorem 2.B are the assumptions of Theorem 2.A. ■

Remark 2.5 (*Quasi-Optimal Rates*)**.** The first part on the right-hand side of the quasi-optimality result (13) can be estimated in terms of n using results from neural-network approximation theory; see, e.g., Yarotsky [61], Gühring, Kutyniok and Petersen [62], and references therein. Such a result may be useful in finding a proper balance of δ_n as $n \to \infty$. Alternatively, the choice of δ_n may be found through a proper *a posteriori* estimator, which seems to be an open problem.[20] □

Remark 2.6 (*Condition on* α)**.** The proof of Theorem 2.B reveals that the condition that α is sufficiently large may be weakened if j_1 has additional structure (e.g., convexity). Indeed, convexity of j_1 guarantees that j will be strongly convex, with strongly convexity constant equal to $\alpha > 0$. If the case, there is no need of Lipschitzness of j'_1 in order to prove only $\alpha > 0$ will be enough. Furthermore, becomes:

$$\|\bar{\xi} - \bar{\xi}_n\|_{\mathbb{X}} < \left(\frac{\alpha + L_1}{\alpha} \inf_{\eta_n \in \mathcal{M}_n} \|\bar{\xi} - \eta_n\|^2_{\mathbb{X}} + \frac{\delta_n}{\alpha} \right)^{1/2} . \quad \square$$

Remark 2.7 (*Physics-Informed Neural Networks (PINN)*)**.** Theorem 2.A can be applied to PINN [7] (for neural-network approximations to PDEs). Indeed, consider

$$j(\xi) = \frac{1}{2} \|f - B\xi\|^2_{\mathbb{L}},$$

where $B : \mathbb{X} \to \mathbb{L}$, and $f - B\xi$ is an abstract residual in some abstract Hilbert space $\mathbb{L}$ (which may include the PDE residual, initial condition and boundary conditions, as in [21], as well as a data residual, as in [50]). Note

[19] In particular, when $\alpha > L_1$, where L_1 is the Lipschitz constant of $j'_1(\cdot)$, then $\gamma = \alpha - L_1$.

[20] There are some works on a posteriori error analysis for neural networks approximations; see, e.g., [63].

that ξ is an approximation to the PDE solution, and *not* the underlying trainable parameters of a neural network. If $B : \mathbb{X} \to \mathbb{L}$ is a linear operator, then the assumptions of Theorem 2.A (Lipschitz continuity and strong convexity) hold true.[21] □

Remark 2.8 (*Deep Ritz Method*). Theorem 2.A can also be applied to the Deep Ritz method [36]. Indeed, consider

$$j(\xi) = \frac{1}{2} b(\xi, \xi) - f(\xi),$$

where $b : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is a continuous coercive and symmetric bilinear form, and $f \in \mathbb{X}^*$. For such a $j(\cdot)$, the assumptions of Theorem 2.A (Lipschitz continuity and strong convexity) hold true. □

2.4. Analysis of constrained control problem

We now proceed with the analysis of the *constrained* control problem (9). We begin by providing conditions that guarantee the well-posedness of the state problem.

Proposition 2.9 (*Stability of the State Problem*). *For each $\xi \in \mathbb{X}$, let $a(\xi; \cdot, \cdot) : \mathbb{V} \times \mathbb{V} \to \mathbb{R}$ and $b : \mathbb{U} \times \mathbb{V} \to \mathbb{R}$ be continuous bilinear forms. For $\mathbb{U}_h \subset \mathbb{U}$ and $\hat{\mathbb{V}} \subseteq \mathbb{V}$, consider the kernel subspace $\hat{\mathbb{K}} := \{v \in \hat{\mathbb{V}} : b(w_h, v) = 0, \forall w_h \in \mathbb{U}_h\}$. Then, the following statements hold true:*

(i) For each $\xi \in \mathbb{X}$, problem (4) is well-posed (for any $f \in \mathbb{V}^*$) if and only if there exist constants $\alpha_h \equiv \alpha_h(\xi) > 0$ and $\beta_h > 0$ such that:[22]

$$\left.\begin{aligned} &\inf_{v_1 \in \hat{\mathbb{K}}} \sup_{v_2 \in \hat{\mathbb{K}}} \frac{a(\xi; v_1, v_2)}{\|v_1\|_{\mathbb{V}} \|v_2\|_{\mathbb{V}}} \geq \alpha_h \,, \\ &\left\{v_2 \in \hat{\mathbb{K}} : a(\xi; v_1, v_2) = 0, \ \forall v_1 \in \hat{\mathbb{K}}\right\} = \{0\} \,, \end{aligned}\right\} \tag{14a}$$

$$\inf_{w_h \in \mathbb{U}_h} \sup_{v \in \hat{\mathbb{V}}} \frac{b(w_h, v)}{\|w_h\|_{\mathbb{U}} \|v\|_{\mathbb{V}}} \geq \beta_h \,. \tag{14b}$$

(ii) If (14) is satisfied, then the following a priori bound holds true for the solution $u_h \in \mathbb{U}_h$ of problem (4):

$$\|u_h\|_{\mathbb{U}} \leq \frac{1}{\beta_h} \left(1 + \frac{\|a(\xi; \cdot, \cdot)\|_{\mathcal{L}(\mathbb{V};\mathbb{V}^*)}}{\alpha_h}\right) \|f\|_{\mathbb{V}^*} \,.$$

(iii) Furthermore, if $a(\xi, \cdot, \cdot)$ is an equivalent inner-product on $\mathbb{V}$, with associated norm $\|\cdot\|_{\mathbb{V},\xi} := \sqrt{a(\xi; \cdot, \cdot)}$, i.e., for some $C_{1,\xi}, C_{2,\xi} > 0$,

$$C_{1,\xi} \|v\|_{\mathbb{V}} \leq \|v\|_{\mathbb{V},\xi} \leq C_{2,\xi} \|v\|_{\mathbb{V}}, \quad \forall v \in \mathbb{V}, \tag{15}$$

then $\alpha_h = (C_{1,\xi})^2$ in (14a), and additionally, the following improved a priori bound holds true:

$$\|u_h\|_{\mathbb{U}} \leq \frac{C_{2,\xi}}{C_{1,\xi}} \frac{1}{\beta_h} \|f\|_{\mathbb{V}^*} \,. \quad \square \tag{16}$$

Proof. See Appendix A.3. ■

To establish the equivalence between the mixed system (4) and the Petrov–Galerkin statement (2), let us define the operators $A : \mathbb{X} \to \mathcal{L}(\hat{\mathbb{V}}; \hat{\mathbb{V}}^*)$ and $B \equiv B_h \in \mathcal{L}(\mathbb{U}_h; \hat{\mathbb{V}}^*)$ by:

$$A(\xi)\hat{v} := a(\xi; \hat{v}, \cdot) \in \hat{\mathbb{V}}^*, \quad \forall \xi \in \mathbb{X}, \ \forall \hat{v} \in \hat{\mathbb{V}}; \tag{17a}$$

$$B w_h := b(w_h, \cdot) \in \hat{\mathbb{V}}^*, \qquad \forall w_h \in \mathbb{U}_h. \tag{17b}$$

[21] To avoid confusion, let us stress that Lipschitz continuity and strong convexity are required with respect to the Hilbert space $\mathbb{X}$, and *not* with respect to the trainable parameters defining a neural network function $\xi_n \in \mathcal{M}_n \subset \mathbb{X}$. The same applies for Remark 2.8.

[22] Only when $\hat{\mathbb{V}}$ is infinite-dimensional, one needs the extra hypothesis in $(14a)_2$. Whenever $a(\xi, \cdot, \cdot)$ is an equivalent inner product on $\mathbb{V}$, then this condition is actually automatically satisfied. Indeed, zero is the only element in $\mathbb{V}$ which is orthogonal to itself.

Note that the state Eqs. (4)(a)–(4)(b) can then be written as follows:

$$A(\xi)r + Bu_h = f \qquad \text{in } \hat{\mathbb{V}}^*\,, \tag{18a}$$
$$B^*r \qquad = 0 \qquad \text{in } (\mathbb{U}_h)^*\,. \tag{18b}$$

Proposition 2.10 (*Equivalent Petrov–Galerkin Problem*). *For each $\xi \in \mathbb{X}$, let $a(\xi;\cdot,\cdot) : \mathbb{V}\times\mathbb{V} \to \mathbb{R}$ and $b : \mathbb{U}\times\mathbb{V} \to \mathbb{R}$ be continuous bilinear forms. Given $\hat{\mathbb{V}} \subseteq \mathbb{V}$ and a discrete trial space $\mathbb{U}_h \subset \mathbb{U}$, let the discrete test space be given by*

$$\mathbb{V}_h(\xi) := \left\{ v \in \hat{\mathbb{V}} \,\middle|\, A(\xi)^* v \in B\mathbb{U}_h \right\}, \tag{19}$$

where $A(\xi)$ and B are defined in (17a) and (17b) (respectively). Assume the existence of $\alpha(\xi) > 0$ such that

$$\inf_{v_1\in\hat{\mathbb{V}}} \sup_{v_2\in\hat{\mathbb{V}}} \frac{a(\xi; v_1, v_2)}{\|v_1\|_{\mathbb{V}}\|v_2\|_{\mathbb{V}}} \geq \alpha(\xi)\,. \tag{20}$$

Then the state problem (18) is equivalent to the Petrov–Galerkin problem (2). □

Proof. See Appendix A.4. ■

Finally, we now present (differentiability) conditions on $\xi \mapsto A(\xi)$ that guarantee the (differentiability) requirements on $\xi \mapsto S_h(\xi)$ in Theorem 2.B and Corollary 2.4. Once in place, existence of (quasi)-minimizers and quasi-optimal convergence follow immediately for the constrained control problem.

To anticipate the connection between derivatives A' and S_h' (as well as R_h'),[23] note that a formal differentiation of (18) (with $r = R_h(\xi)$ and $u_h = S_h(\xi)$) with respect to ξ in the direction $\eta \in \mathbb{X}$ yields:

$$\begin{aligned} A(\xi)R_h'(\xi)\eta + BS_h'(\xi)\eta &= -A'(\xi)\eta\, R_h(\xi) && \text{in } \hat{\mathbb{V}}^*\,,\\ B^*R_h'(\xi)\eta &= 0 && \text{in } (\mathbb{U}_h)^*\,. \end{aligned}$$

One may therefore expect that suitable conditions on $A(\cdot)$ will imply desired conditions on $S_h(\cdot)$ (and $R_h(\cdot)$):

Proposition 2.11 (*State Differentiability*). *Let $R_h(\cdot)$ and $S_h(\cdot)$ be the state operators as defined in (10), and let $A(\cdot)$ be as defined in (17a). Assume the conditions of Proposition 2.9, including the well-posedness conditions (14). Then, the following statements hold true:*

(i) If $A(\cdot)$ has a Gâteaux derivative at $\xi \in \mathbb{X}$ in the direction $\eta \in \mathbb{X}$, then $R_h(\cdot)$ and $S_h(\cdot)$ have a Gâteaux derivative at ξ in the direction η.

(ii) If $A(\cdot)$ is Gâteaux-differentiable at ξ, then so are $R_h(\cdot)$ and $S_h(\cdot)$.

(iii) If $A(\cdot)$, $A'(\cdot)$ and $\alpha_h^{-1}(\cdot)$ are uniformly bounded on $\mathbb{X}$, then $R_h'(\cdot)$ and $S_h'(\cdot)$ are also uniformly bounded on $\mathbb{X}$.

(iv) Additionally, if $A'(\cdot)$ is Lipschitz continuous, then $R_h'(\cdot)$ and $S_h'(\cdot)$ are Lipschitz continuous as well. □

Proof. See Appendix A.5. ■

Corollary 2.12 (*Constrained Problem: (quasi-)Minimizers & Quasi-Optimality*). *Let $J(w_h, \xi) = J_1(w_h) + \alpha\, j_2(\xi)$ as in (5) with $Q \in \mathcal{L}(\mathbb{U};\mathbb{Z})$. Let the associated $j(\cdot)$ be as in (11). Under the conditions of Propositions 2.9 and 2.11, and assuming α is sufficiently large, the statements (i), (ii) and (iii) of Theorem 2.A hold true.*

In other words, the constrained control problem (9) has a quasi-minimizer in $\mathcal{M}_n$ that converges quasi-optimally to the unique minimizer in $\mathbb{X}$. □

Proof. The results of Propositions 2.9 and 2.11, together with α sufficiently large, are the assumptions of Theorem 2.B, whose results are the assumptions of Theorem 2.A. ■

[23] Recall that the Gâteaux derivative of, e.g., A at $\xi \in \mathbb{X}$ in the direction $\eta \in \mathbb{X}$ is given by $A'(\xi)\eta = \lim_{t\to 0} \frac{A(\xi+t\eta)-A(\xi)}{t}$, provided the limit exists in $\mathcal{L}(\hat{\mathbb{V}};\hat{\mathbb{V}}^*)$. If the map $\eta \mapsto A'(\xi)\eta$ is linear and continuous from $\mathbb{X}$ to $\mathcal{L}(\hat{\mathbb{V}};\hat{\mathbb{V}}^*)$, then A is Gâteaux differentiable at $\xi \in \mathbb{X}$.

3. Conforming weak formulations with suitable control

In this section, we study various weighted versions of conforming weak formulations, viz., least squares, Galerkin and minimal-residual formulations, and we illustrate these with PDE examples. The aim is to propose suitable ξ-dependent weighting within the weak forms, in order to be able to prove the assumptions of Propositions 2.9 and 2.11. By Corollary 2.12, we can then conclude that the corresponding constrained neural-control problem has desired properties (existence of quasi-minimizers and quasi-optimal convergence).

In what follows, we often consider a (positive) weight function $\omega(\xi) : \Omega \to \mathbb{R}$. E.g., we consider a mapping $\omega : L^2(\Omega) \to L^\infty(\Omega)$, which takes a control $\xi \in \mathbb{X} = L^2(\Omega)$ and generates a function $\omega(\xi)$, which is positive a.e. in Ω.[24] Explicit examples of such mappings are discussed in Remarks 3.6 and 3.7. We shall use the notation $\varpi(\xi) := 1/\omega(\xi)$ to indicate the (multiplicative) inverse of $\omega(\xi)$.

3.1. Weighted least squares formulations

Let $d \in \mathbb{N}$ and $\Omega \subset \mathbb{R}^d$ be an open bounded Lipschitz domain. Let $B : \mathbb{H}_B \to L^2(\Omega)$ be a linear differential operator in strong form, where $\mathbb{U} = \mathbb{H}_B$ denotes a general graph space:

$$\mathbb{H}_B := \Big\{ w \in L^2(\Omega) \,\Big|\, Bw \in L^2(\Omega) \text{ and suitable homogeneous boundary conditions} \Big\}.$$

Examples of $\mathbb{H}_B$ are presented below. We assume that $\mathbb{H}_B$ is a Hilbert space when endowed with the inner product

$$\big(w_1, w_2\big)_{\mathbb{H}_B} := \big(w_1, w_2\big)_{L^2(\Omega)} + \big(Bw_1, Bw_2\big)_{L^2(\Omega)}, \qquad \forall w_1, w_2 \in \mathbb{H}_B,$$

and that B is boundedly invertible from $\mathbb{H}_B$ onto $\mathbb{V}^* := L^2(\Omega) =: \mathbb{V} = \hat{\mathbb{V}}$.

Let $f \in L^2(\Omega)$, and let $\mathbb{U}_h \subset \mathbb{H}_B$ be a conforming discrete (finite element) space. Given a mapping $\omega : L^2(\Omega) \to L^\infty(\Omega)$, which generates a positive weight function $\omega(\xi)$, we aim to find $u_h \equiv S_h(\xi) \in \mathbb{U}_h$, which is the solution of the *weighted* least squares problem:

$$u_h = \underset{w_h \in \mathbb{U}_h}{\arg\min}\ \frac{1}{2} \Big\| \sqrt{\omega(\xi)}\big(f - Bw_h\big) \Big\|^2_{L^2(\Omega)}.$$

The optimality condition of such a minimizer is given by the weighted least squares (LSQ) method:

$$\Big(\omega(\xi)(f - Bu_h), Bw_h\Big)_{L^2(\Omega)} = 0, \quad \forall w_h \in \mathbb{U}_h\,. \tag{21}$$

In particular, notice that we can directly identify the test space in (2) as

$$\mathbb{V}_h(\xi) = \omega(\xi) B \mathbb{U}_h = \big\{ v \in L^2(\Omega) \,\big|\, v = \omega(\xi) Bw_h \text{ for some } w_h \in \mathbb{U}_h \big\}\,.$$

Example 3.1 (*Weighted LSQ for Advection–Reaction*). Let $\beta \in (L^\infty(\Omega))^d$ be an advection field, and let $c \in L^\infty(\Omega)$ be a reaction coefficient. The inflow boundary is $\partial\Omega_- := \{x \in \partial\Omega : \beta(x) \cdot n(x) < 0\}$, where $n(x)$ corresponds to the unit outward normal. Next, define $Bw := \beta \cdot \nabla w + cw$ and $\mathbb{H}_B := \big\{ w \in L^2(\Omega) : \beta \cdot \nabla w + cw \in L^2(\Omega) \text{ and } w|_{\partial\Omega_-} = 0 \big\}$. Therefore, finding $u \in \mathbb{H}_B$ such that $Bu = f$, corresponds to the strong form of the advection–reaction PDE with homogeneous inflow data, which is well-posed under suitable conditions on β and c (see, e.g., [64]).

The weighted LSQ method (21) translates into finding $u_h \in \mathbb{U}_h$ such that

$$\int_\Omega \omega(\xi)\big(\beta \cdot \nabla u_h + cu_h\big)\big(\beta \cdot \nabla w_h + cw_h\big) = \int_\Omega \omega(\xi)\, f\, \big(\beta \cdot \nabla w_h + cw_h\big)\,, \qquad \forall w_h \in \mathbb{U}_h\,. \quad \square$$

Example 3.2 (*Weighted LSQ for the Strong Laplacian*). Set $\mathbb{H}_B := \big\{ w \in H^1_0(\Omega) : \Delta w \in L^2(\Omega) \big\}$, where $Bw := -\Delta w$. Finding $u \in \mathbb{U}_B$ such that $Bu = f$, corresponds to the strong form of the Poisson equation with homogeneous Dirichlet boundary data. The weighted LSQ problem (21) translates into finding $u_h \in \mathbb{U}_h$ such that

$$\int_\Omega \omega(\xi)\, \Delta u_h\, \Delta w_h = -\int_\Omega \omega(\xi)\, f\, \Delta w_h\,, \qquad \forall w_h \in \mathbb{U}_h\,. \quad \square$$

[24] The rationale behind this mapping is that it generates a desired weight function $\omega(\xi)$ in $L^\infty(\Omega)$, while keeping an unconstrained Hilbert setting for the control variable $\xi \in \mathbb{X} = L^2(\Omega)$.

Remark 3.3 (*Weighted LSQ for Laplacian in Mixed Form*)**.** The above can be extended to least squares of mixed problems, e.g., the Laplacian as a first-order system. Let $\mathbb{V} := L^2(\Omega) \times \big(L^2(\Omega)\big)^d$ and $\mathbb{H}_B := H_0^1(\Omega) \times H(\text{div}; \Omega)$, where $B : \mathbb{H}_B \to \mathbb{V}$ is defined by $B(u, \vec{q}\,) := (\text{div}\,\vec{q}, \vec{q} - \nabla u)$. Given $g \in L^2(\Omega)$, the problem $B(u, \vec{q}\,) = (g, \vec{0}\,)$ corresponds to the Poisson equation in mixed form with homogeneous Dirichlet boundary conditions.

Given a pair of conforming discrete subspaces $\mathbb{U}_h \subset \mathbb{H}_B$, a pair of controls $\xi = (\xi_1, \xi_2) \in L^2(\Omega)^2 =: \mathbb{X}$, and two mappings $\omega_1, \omega_2 : L^2(\Omega) \to L^\infty(\Omega)$, a possible weighted LSQ method is to find $(u_h, \vec{q}_h) \in \mathbb{U}_h$ such that

$$\int_\Omega \omega_1(\xi_1)\,\text{div}\,\vec{q}_h\ \text{div}\,\vec{p}_h + \int_\Omega \omega_2(\xi_2)\,(\vec{q}_h - \nabla u_h)\cdot(\vec{p}_h - \nabla w_h) = \int_\Omega \omega_1(\xi_1)\,f\,\text{div}\,\vec{p}_h\,,$$

for all $(w_h, \vec{p}_h) \in \mathbb{U}_h$. □

To establish the connection with the general mixed system (4), we set $r = \omega(\xi)(f - Bu_h)$ so that (21) is equivalent to:

$$\big(\varpi(\xi)\,r, v\big)_{L^2(\Omega)} + \big(Bu_h, v\big)_{L^2(\Omega)} = (f, v)_{L^2(\Omega)}, \quad \forall v \in \mathbb{V}, \tag{22a}$$

$$\big(Bw_h, r\big)_{L^2(\Omega)} = 0, \quad \forall w_h \in \mathbb{U}_h\,. \tag{22b}$$

Thus, in this case, the continuous bilinear forms $a(\xi; \cdot, \cdot) : \mathbb{V} \times \mathbb{V} \to \mathbb{R}$ and $b : \mathbb{H}_B \times \mathbb{V} \to \mathbb{R}$ in (4) are given by

$$a(\xi; v_1, v_2) := \big(\varpi(\xi)\,v_1, v_2\big)_{L^2(\Omega)}, \quad \forall v_1, v_2 \in \mathbb{V} = L^2(\Omega), \tag{23a}$$

$$b(w, v) := (Bw, v)_{L^2(\Omega)}, \quad \forall w \in \mathbb{H}_B, \forall v \in \mathbb{V}. \tag{23b}$$

Proposition 3.4 (*Weighted Least Squares*)**.** *Let $\varpi : L^2(\Omega) \to L^\infty(\Omega)$ be a differentiable mapping, such that for some positive constants $\varpi_{\min}$, $\varpi_{\max}$, ϖ'_∞, and ϖ_L, the application $\varpi(\cdot)$ satisfies*

- $\varpi_{\min} \le \varpi(\xi) \le \varpi_{\max}$, *for all $\xi \in L^2(\Omega)$;*
- $\|\varpi'(\xi)\|_{\mathcal{L}(L^2(\Omega);L^\infty(\Omega))} \le \varpi'_\infty$, *for all $\xi \in L^2(\Omega)$;*
- $\|\varpi'(\xi_1) - \varpi'(\xi_2)\|_{\mathcal{L}(L^2(\Omega);L^\infty(\Omega))} \le \varpi_L \|\xi_1 - \xi_2\|_{L^2(\Omega)}$, *for all $\xi_1, \xi_2 \in L^2(\Omega)$.*

Then, the following statements hold true:

(i) The bilinear forms in (23) satisfy the inf − sup *conditions (14), and thus the weighted least squares problem (22) is well-posed.*
(ii) The state operator $S_h(\cdot)$ $(= u_h)$ of the problem (22) is uniformly bounded on $\mathbb{X} = L^2(\Omega)$ and differentiable.
(iii) The derivative $S'_h(\cdot)$ is uniformly bounded on $\mathbb{X} = L^2(\Omega)$ and Lipschitz continuous. □

Proof. See Appendix A.6 ■

Remark 3.5 (*Neural Control of Weighted Least Squares*)**.** Proposition 3.4 guarantees that the conditions of Propositions 2.9 and 2.11 are satisfied, hence Corollary 2.12 applies to the neural optimization of the above weighted least squares formulation. In particular, this means that it can be applied to the PDEs in Examples 3.1 and 3.2 (and with minor modifications also to the setting in Remark 3.3), provided the weight $\varpi(\xi)$ satisfies the three nontrivial conditions in Proposition 3.4. The next two remarks discuss this in further detail.

Remark 3.6 (*Suitable Weight Functions: Integral Operators*)**.** A general way to obtain a weight function $\varpi(\xi)$ $(= 1/\omega(\xi))$ is when the mapping $\varpi : L^2(\Omega) \to L^\infty(\Omega)$ is an *integral operator*. Indeed, given a kernel function $k : \Omega \times \Omega \to \mathbb{R}$, and differentiable real functions $f, g : \mathbb{R} \to \mathbb{R}$, we can define

$$[\varpi(\xi)](\cdot) = f\left(\int_\Omega k(\cdot, y)\,g\big(\xi(y)\big)\,\mathrm{d}y\right), \quad \forall \xi \in L^2(\Omega). \tag{24}$$

There are several options to obtain a well-defined expression (24). For instance, we can ask f to be bounded and $k(x, \cdot)g(\xi(\cdot)) \in L^1(\Omega)$, for all $x \in \Omega$. Moreover, if we want to obtain Gâteaux differentiability of ϖ, then the following expression has to be well-defined for any $\xi, \eta \in L^2(\Omega)$:

$$\big[\varpi'(\xi)\big]\eta = f'\left(\int_\Omega k(\cdot, y)\,g\big(\xi(y)\big)\,\mathrm{d}y\right)\int_\Omega k(\cdot, y)\,g'(\xi(y))\,\eta(y)\,\mathrm{d}y,$$

which additionally requires $k(x, \cdot)\,g'\big(\xi(\cdot)\big) \in L^2(\Omega)$, for each $x \in \Omega$.

Let us describe a fundamental example for $\varpi(\xi)$ that satisfies the assumptions in Proposition 3.4. Consider a positive constant $M > 0$ and the sigmoid function $\sigma(s) = 1/(1+e^{-s})$. Let χ_{B_r} be the characteristic function of a ball $B_r \subset \Omega$ of radius $r > 0$ centered at the origin. Then define

$$[\varpi(\xi)](x) = 1 + M\sigma\left(\frac{1}{|B_r|}\int_\Omega \chi_{B_r}(y-x)\,\xi(y)\,\mathrm{d}y\right), \tag{25}$$

which has the form of expression (24) for $f(s) = 1 + M\sigma(s)$, $g(s) = s$, and $k(x,y) = \frac{1}{|B_r|}\chi_{B_r}(y-x)$. To verify the assumptions in Proposition 3.4, first observe that $1 \le \varpi(\xi) \le 1 + M$, for all $\xi \in L^2(\Omega)$. Second,

$$\left\|\left[\varpi'(\xi)\right]\eta\right\|_{L^\infty(\Omega)} \le \frac{M}{4}|B_r|^{-\frac{1}{2}}\|\eta\|_{L^2(\Omega)}, \quad \forall\xi,\eta \in L^2(\Omega),$$

and thus, $\|\varpi(\xi)'\|_{\mathcal{L}(L^2(\Omega);L^\infty(\Omega))} \le \frac{M}{4}|B_r|^{-\frac{1}{2}}$. And third, denoting the Lipschitz constant of $\sigma'(\cdot)$ by $L_{\sigma'} > 0$,[25] we get

$$\left\|\varpi'(\xi_1)\eta - \varpi'(\xi_2)\eta\right\|_{L^\infty(\Omega)} \le M|B_r|^{-1}L_{\sigma'}\|\xi_1-\xi_2\|_{L^2(\Omega)}\|\eta\|_{L^2(\Omega)},$$

for all $\xi_1, \xi_2, \eta \in L^2(\Omega)$, which implies that $\varpi'(\cdot)$ is Lipschitz continuous. □

Remark 3.7 (*Practical Weight Functions: Approximation of Integral Operator*)**.** For simplicity, let us consider (25) when $\Omega \subset \mathbb{R}$. One can approximate the integral in (25) by (Gaussian) quadrature using points and weights $\{(x_q, w_q)\}$, with $x_q \in (-1,1)$, for $q = 1, 2, \ldots, N$. We then have

$$[\varpi(\xi)](x) = 1 + M\sigma\left(\frac{1}{2r}\int_\Omega \chi_{(-r,r)}(y-x)\,\xi(y)\,\mathrm{d}y\right) \tag{26}$$

$$\approx [\varpi_N(\xi)](x) := 1 + M\sigma\left(\frac{1}{2}\sum_{q=1}^N w_q\,\xi(x+rx_q)\right). \tag{27}$$

In particular, for a single quadrature point (mid-point rule), we obtain the approximation

$$[\varpi(\xi)](x) \approx [\varpi_1(\xi)](x) = 1 + M\sigma(\xi(x)), \tag{28}$$

which is just a composition of functions, $(1+M\sigma)\circ\xi(\cdot)$, hence attractive in practical implementations.

We note however that $\varpi_1(\cdot)$ does *not* satisfy the assumptions of Proposition 3.4,[26] hence Corollary 2.12 cannot be applied. Nevertheless, numerical experiments in Section 4 indicate that the use of $\varpi_1(\xi)$ does not deteriorate performance. Therefore, we expect the result of Corollary 2.12 to be valid for a larger class of mappings $\varpi(\cdot)$. □

3.2. Weighted Galerkin formulations

Consider a Hilbert space $\mathbb{U} = \mathbb{V}$ on $\Omega \subset \mathbb{R}^d$ and a continuous bilinear form $b : \mathbb{V}\times\mathbb{V} \to \mathbb{R}$ satisfying (for some constant $\beta > 0$) the following conditions

$$\sup_{v\in\mathbb{V}} \frac{b(w,v)}{\|v\|_\mathbb{V}} \ge \beta\|w\|_\mathbb{V}, \quad \forall w \in \mathbb{V}, \tag{29a}$$

$$\left\{v \in \mathbb{V} : b(w,v) = 0, \forall w \in \mathbb{V}\right\} = \{0\}. \tag{29b}$$

Given $f \in \mathbb{V}^*$, Babuška–Brezzi theory (see, e.g., [6]) ensures the existence of a unique $u \in \mathbb{V}$ such that

$$b(u,v) = f(v), \quad \forall v \in \mathbb{V}. \tag{30}$$

Now, given a mapping $\omega : \mathbb{X} \to \mathbb{W}$ (requirements on the space $\mathbb{W}$ are clarified below), a control $\xi \in \mathbb{X}$, and a conforming discrete subspace $\mathbb{U}_h \subset \mathbb{V}$, consider the following *weighted Galerkin* discretization of (30):

$$\begin{cases}\text{Find } u_h \equiv S_h(\xi) \in \mathbb{U}_h : \\ \qquad b\big(u_h, \omega(\xi)w_h\big) = f\big(\omega(\xi)w_h\big), \quad \forall w_h \in \mathbb{U}_h.\end{cases} \tag{31}$$

[25] Indeed, we can use the uniform bound of $\sigma''(\cdot)$.

[26] Indeed, one can verify that $\varpi_1'(\xi)(\eta) = \big(1+M\sigma'(\xi)\big)\eta$, which is in general not in $L^\infty(\Omega)$ for $\xi,\eta \in L^2(\Omega)$. Thus, the second assumption in Proposition 3.4 is violated.

Notice that one can directly connect (31) to the form stated in (2) by identifying

$$\mathbb{V}_h(\xi) = \omega(\xi)\mathbb{U}_h = \left\{ v_h \in \mathbb{V} \,\middle|\, v_h = \omega(\xi) w_h \text{ for some } w_h \in \mathbb{U}_h \right\}.$$

Example 3.8 (*Weighted Galerkin for Laplacian*)**.** Let $\mathbb{V} = H_0^1(\varOmega)$ and $f \in \mathbb{V}^* = H^{-1}(\varOmega)$. Consider a weight function $\omega(\xi)$, where $\omega : L^2(\varOmega) \to W^{1,\infty}(\varOmega)$ (hence $\mathbb{W} = W^{1,\infty}(\varOmega)$). Then, the weighted Galerkin formulation for the Poisson equation with homogeneous Dirichlet boundary conditions is to find $u_h \in \mathbb{U}_h \subset H_0^1(\varOmega)$ such that

$$\int_\varOmega \nabla u_h \cdot \big(v_h \nabla \omega(\xi) + \omega(\xi) \nabla v_h \big) = \Big\langle f \,,\, \omega(\xi) v_h \Big\rangle, \qquad \forall v_h \in \mathbb{U}_h. \quad \square \tag{32}$$

The above example illustrates that $\mathbb{W}$ should be such that the product $\omega(\xi)\, v_h$ is well-defined in $\mathbb{V}$, whenever $\omega(\xi) \in \mathbb{W}$ and $v_h \in \mathbb{U}_h \subset \mathbb{V}$. In fact, we shall assume that $\mathbb{W}$ is such that, for any $\mathrm{w} \in \mathbb{W}$, the multiplication operator $M_\mathrm{w} : \mathbb{V} \to \mathbb{V}$, defined by

$$M_\mathrm{w} v := \mathrm{w} v, \quad \forall v \in \mathbb{V},$$

is a linear and continuous map.

Remark 3.9 (*Multiplication in H^1*)**.** Let $\mathbb{V} = H^1(\varOmega)$ and $\mathbb{W} = W^{1,\infty}(\varOmega)$. Then, it is easy to see that $M_\mathrm{w} : H^1(\varOmega) \to H^1(\varOmega)$ is a linear and continuous map, for all $\mathrm{w} \in W^{1,\infty}(\varOmega)$.[27] □

To establish the equivalence with the mixed formulation (4) (and thereby fit the weighted Galerkin formulation within the abstract setting of Section 2), we furthermore let $\mathbb{W} \equiv \mathbb{W}(\varOmega)$ consist of measurable functions on $\varOmega$, and introduce a particular subset of interest:

$$\mathbb{W}_+ := \left\{ \mathrm{w} \in \mathbb{W} \,\middle|\, \exists \mathrm{w}_{\min} > 0 \text{ for which } \mathrm{w}_{\min} \le \mathrm{w}(x) \le \tfrac{1}{\mathrm{w}_{\min}}, \forall x \in \varOmega \right\} \subset \mathbb{W}.$$

Notice that $\frac{1}{\mathrm{w}} \in \mathbb{W}_+$ iff $\mathrm{w} \in \mathbb{W}_+$. We can then define the inverse of multiplication $M_\mathrm{w}^{-1} := M_{\frac{1}{\mathrm{w}}}$, which is justified by the fact that

$$M_\mathrm{w}^{-1}(M_\mathrm{w} v) = v = M_\mathrm{w}(M_\mathrm{w}^{-1} v), \quad \forall v \in \mathbb{V}. \tag{33}$$

The adjoint operators of M_w and M_w^{-1} will be denoted by M_w^* and M_w^{-*} respectively. Using the relations (33), it is straightforward to see that the adjoint operators satisfy

$$M_\mathrm{w}^{-*}(M_\mathrm{w}^* \ell) = \ell = M_\mathrm{w}^*(M_\mathrm{w}^{-*} \ell), \quad \forall \ell \in \mathbb{V}^*. \tag{34}$$

Next, we translate problem (31) into operator notation by means of the operator $B \in \mathcal{L}(\mathbb{V}; \mathbb{V}^*)$ such that $\mathbb{V} \ni w \mapsto Bw := b(w, \cdot) \in \mathbb{V}^*$. Notice that such an operator is invertible thanks to conditions (29). Problem (31) translates into finding $u_h \equiv S_h(\xi) \in \mathbb{U}_h$ such that

$$\Big\langle B u_h, M_{\varpi(\xi)}^{-1} v_h \Big\rangle = \Big\langle f, M_{\varpi(\xi)}^{-1} v_h \Big\rangle, \quad \forall v_h \in \mathbb{U}_h.$$

Hence, by means of the adjoint relation we get

$$\Big\langle M_{\varpi(\xi)}^{-*}(f - B u_h), v_h \Big\rangle = 0, \quad \forall v_h \in \mathbb{U}_h. \tag{35}$$

Since B is invertible, so is $B^* : \mathbb{V} \to \mathbb{V}^*$ defined by $\mathbb{V} \ni v \mapsto b(\cdot, v) \in \mathbb{V}^*$. Therefore, there exists a unique $r \in \mathbb{V}$ such that $B^* r = M_{\varpi(\xi)}^{-*}(f - B u_h)$ in $\mathbb{V}^*$. Thus, multiplying this last equation by $M_{\varpi(\xi)}^*$, using (34), (35), and the definition of $r \in \mathbb{V}$, we arrive to the mixed form

$$\left\{ \begin{array}{lll} \big\langle B^* r, M_{\varpi(\xi)} v \big\rangle + b(u_h, v) & = f(v), \quad \forall v \in \mathbb{V}, & \text{(a)} \\ \qquad\qquad b(v_h, r) & = 0, \qquad \forall v_h \in \mathbb{U}_h. & \text{(b)} \end{array} \right. \tag{36}$$

[27] Indeed, $\|\mathrm{w} v\|_{H^1(\varOmega)} \le C \|\mathrm{w}\|_{W^{1,\infty}(\varOmega)} \|v\|_{H^1(\varOmega)}$. It is also true for Hilbert spaces $\mathbb{V} \subset L^2(\varOmega)$ containing at most first-order (weak) derivatives in $L^2(\varOmega)$ (e.g., first-order graph spaces).

Observe that (36) has the structure of (4) for $\hat{\mathbb{V}} := \mathbb{V} = \mathbb{U}$, and

$$a(\xi; r, v) := \langle B^* r, M_{\varpi(\xi)} v\rangle = b\big(\varpi(\xi)v, r\big).$$

The next proposition establishes the well-posedness of the weighted Galerkin formulation (31).

Proposition 3.10 (*Weighted Galerkin Formulation*). *Let $\omega : \mathbb{X} \to \mathbb{W}$. Assume that for some positive function $\alpha_h : \mathbb{X} \to \mathbb{R}$, the continuous bilinear form $b : \mathbb{V} \times \mathbb{V} \to \mathbb{R}$ satisfies the discrete* inf − sup *condition*

$$\sup_{v_h \in \mathbb{U}_h} \frac{b(w_h, \omega(\xi)v_h)}{\|v_h\|_{\mathbb{V}}} \geq \alpha_h(\xi)\|w_h\|_{\mathbb{V}}\,, \qquad \forall w_h \in \mathbb{U}_h\,,\ \forall \xi \in \mathbb{X}\,. \tag{37}$$

Then, the following statements hold true:

(i) For any $f \in \mathbb{V}^*$ and $\xi \in \mathbb{X}$, problem (31) is well-posed.
(ii) If there exist uniform constants $\alpha > 0$ and $\omega_\infty > 0$ such that $\alpha_h(\xi) \geq \alpha$ and $\|\omega(\xi)\|_{\mathbb{W}} \leq \omega_\infty$ for all $\xi \in \mathbb{X}$, then the solution $u_h \equiv S_h(\cdot)$ to problem (31) is uniformly bounded on $\mathbb{X}$.
(iii) Additionally, if $\omega(\cdot)$ is differentiable, then $S_h(\cdot)$ is also differentiable. Moreover, if $\omega'(\cdot)$ is uniformly bounded and Lipschitz-continuous, then $S_h'(\cdot)$ is also uniformly bounded and Lipschitz-continuous. □

Proof. See Appendix A.7. ■

Remark 3.11 (*Neural Control of Weighted Galerkin*). Proposition 3.10 guarantees that the conditions of Propositions 2.9 and 2.11 are satisfied, hence Corollary 2.12 applies to the neural optimization of the above weighted Galerkin formulation. □

Remark 3.12 (*Inconvenient Condition for Weighted Galerkin*). While for the weighted least squares method the conditions on the weight are explicit (recall Proposition 3.4), for weighted Galerkin the condition (37) is problem-dependent. This is even true when (37) is replaced by the stronger condition of *coercivity*:

$$b\big(v_h, \omega(\xi)v_h\big) \geq \alpha_h(\xi)\|v_h\|^2_{\mathbb{V}}\,, \qquad \forall v_h \in \mathbb{U}_h\,. \tag{38}$$

A more detailed analysis of when (37) or (38) is satisfied in general requires further study and is outside of the scope of this work. Indeed, for the examples in Remarks 3.13 and 3.14, satisfying (38) may require inconvenient constraints on ξ. It is therefore much more convenient to have neural control of least squares formulations. When the continuous setting of the PDE at hand does not fit a least squares formulation (as in the examples in Remarks 3.13 and 3.14), instead of weighted Galerkin, we then recommend the use of (weighted) dual minimal-residual formulations; see Section 3.3. □

Remark 3.13 (*Weighted Galerkin for Laplacian: Nontrivial Stability*). Let us illustrate the difficulty of satisfying coercivity (38) with an elementary example for the Laplacian. Recall Example 3.8, and consider a 1-D setting, taking $\Omega = (0, 1)$. Assume a Neumann boundary condition at $x = 0$ and a Dirichlet condition at $x = 1$. In that case

$$\begin{aligned} &\mathbb{V} = H^1_{0)}(\Omega) := \big\{v \in H^1(\Omega) \mid v(1) = 0\big\}\,, \\ &b(u, \mathrm{w}v) = \int_0^1 u'\big(v\,\mathrm{w}' + \mathrm{w}\,v'\big)\,\mathrm{d}x\,. \end{aligned} \tag{39}$$

Assume $\mathbb{U}_h \subset \mathbb{V}$ is a standard linear finite element space.

Take the weight function as $\mathrm{w}(x) = cx$ for any $c > 0$. Note that $\mathrm{w}(x) > 0$ for all $x \in (0, 1)$, hence this weight seems harmless. However, for any $v_h \in \mathbb{U}_h \subset \mathbb{V}$,

$$\begin{aligned} b(v_h, \mathrm{w}v_h) &= c\int_0^1 x\,(v_h')^2\,\mathrm{d}x + c\int_0^1 v_h\,v_h'\,\mathrm{d}x \\ &= c\left(\int_0^1 x(v_h')^2\,\mathrm{d}x - \frac{1}{2}v_h(0)^2\right) \qquad (v_h(1) = 0) \end{aligned}$$

Surprisingly, the right-hand side *vanishes* for, for example, the left-most half-hat function:

$$v_h(x) = \begin{cases} 1 - \frac{x}{h}\,, & x \in (0,h), \quad \text{with } (h \leq 1)\,, \\ 0\,, & \text{otherwise}\,. \end{cases}$$

This shows that coercivity (38) can *not* be satisfied in general without additional conditions on w.[28]

On the other hand, a condition on w can be found that ensures coercivity. We consider the general case for the Laplacian, starting from the bilinear form in (32). First notice that, for any $\mathrm{w} \in W^{1,\infty}(\Omega)$ such that $\mathrm{w}(x) \geq \mathrm{w}_{\min}$ for all $x \in \Omega$,

$$\begin{aligned} b(v, \mathrm{w}v) &\geq \mathrm{w}_{\min} \|\nabla v\|^2_{L^2(\Omega)} - \|\nabla \mathrm{w}\|_{L^\infty(\Omega)} \|v\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} \\ &\geq \big(\mathrm{w}_{\min} - C_\Omega \|\nabla \mathrm{w}\|_{L^\infty(\Omega)}\big) \|\nabla v\|^2_{L^2(\Omega)}\,, \end{aligned}$$

where a Poincaré inequality was used. Therefore, the constraint $C_\Omega \|\nabla \mathrm{w}\|_{L^2(\Omega)} < \mathrm{w}_{\min}$ is sufficient to guarantee (38) (and thereby (37)). Unfortunately, since $\mathrm{w} = \omega(\xi)$, such a condition translates into a constraint on $\nabla\xi$, which may be very inconvenient to impose in practice. □

Remark 3.14 (*Weighted Galerkin for Advection: Surprising Stability*). While *weighted* Galerkin may *de*stabilize the standard Galerkin method, as illustrated in Remark 3.13 for the Laplacian, for other PDEs, the addition of a weight may also *stabilize* an otherwise unstable method. In both situations, stability does require a nontrivial condition on the weight function.

Let us illustrate the stabilizing effect of weighted Galerkin for advection in 1-D for simplicity, i.e., the differential equation $u' = f$ in $\Omega = (0,1)$, and $u(0) = 0$, and the following weak form:

$$\text{Find } u_h \in \mathbb{U}_h : \qquad b(u_h, \mathrm{w}v_h) := -\int_\Omega u_h\,(\mathrm{w}\,v_h)' = \big\langle f\,,\, \mathrm{w}\,v_h \big\rangle \qquad \forall v_h \in \mathbb{U}_h\,,$$

where $\mathbb{U}_h \subset H^1_{0)}(0,1) := \big\{v \in H^1(0,1) : v(1) = 0\big\}$ and f is allowed to be rough, i.e., in $[H^1_{0)}(0,1)]^*$.[29] Note that standard Galerkin has $\mathrm{w}(x) = 1$ for $x \in \Omega$, and fails to be coercive on $H^1_{0)}(0,1)$.

For weighted Galerkin, the left-hand side of (38) becomes in this case:

$$b(v_h, \mathrm{w}v_h) = -\int_\Omega \mathrm{w}'\,v_h^2 - \int_\Omega v_h\,\mathrm{w}\,v_h' = -\int_\Omega \mathrm{w}'\,v_h^2 + \int_\Omega v_h\,(\mathrm{w}\,v_h)' + \mathrm{w}(0)\,v_h^2(0)$$

Therefore,

$$b(v_h, \mathrm{w}v_h) = -\frac{1}{2}\int_\Omega \mathrm{w}'\,v_h^2 + \frac{1}{2}\mathrm{w}(0)v_h^2(0)\,,$$

which motivates to assume a (global) inverse inequality[30]:

$$\|v_h\|^2_{L^2(\Omega)} \geq C_{\mathrm{inv}} h^2 \|v_h'\|^2_{L^2(\Omega)} \qquad \forall v_h \in \mathbb{U}_h\,,$$

and the following conditions on w:

$$\mathrm{w}'(x) \leq -\frac{2\hat{C}}{h^2} < 0\,, \quad \forall x \in \Omega\,, \qquad \text{and} \qquad \mathrm{w}(0) \geq 0\,, \tag{40}$$

for some $\hat{C} > 0$. Indeed, we then have coercivity (38) (and thereby (37)):

$$b(v_h, \mathrm{w}v_h) \geq \frac{\hat{C}}{h^2}\int_\Omega v_h^2 \geq C_{\mathrm{inv}}\hat{C}\int_\Omega (v_h')^2\,.$$

The conclusion of Remark 3.13 applies here as well: Since $\mathrm{w} = \omega(\xi)$, condition (40) translates into a constraint on derivatives of ξ, which may be very inconvenient to impose in practice. □

[28] We note that this counterexample to coercivity also works for strictly positive weights on [0, 1], in the case of, for example, quadratic finite elements: Let $\mathrm{w}_{\min} > 0$, and consider $\mathrm{w}(x) = \mathrm{w}_{\min} + cx$, with $c > 0$ to be specified. Taking $v_h(x) = \frac{1}{2}(x-1)^2$ yields $b(v_h, \mathrm{w}v_h) = \mathrm{w}_{\min}\int_0^1 (v_h')^2\,\mathrm{d}x + c\big(\int_0^1 x\,(v_h')^2\,\mathrm{d}x - \frac{1}{2}v_h(0)^2\big) = \frac{1}{24}(8\mathrm{w}_{\min} - c)$, which is zero for $c = \frac{1}{8}\mathrm{w}_{\min}$.

[29] A similar analysis also applies to the stronger setting having bilinear form $b(u_h, \mathrm{w}\,v_h) = \int_\Omega u_h'\,\mathrm{w}\,v_h$, $f \in L^2(\Omega)$, and $\mathbb{U}_h \subset H^1_{0)}(0,1) := \big\{w \in H^1(0,1) : w(0) = 0\big\}$.

[30] This holds for example when $\mathbb{U}_h$ is a quasi-uniform FE space. If quasi-uniformity does not hold, one can extend the analysis by assuming element-wise inverse inequalities.

3.3. Weighted dual minimal residual formulations

In this section, we consider weighted minimal residual (MinRes) formulations. These are particularly useful if the continuous setting of the PDE does not fit a weighted least squares formulation, that is, when the residual is not in $L^2(\Omega)$. As examples, we consider the Laplacian in $H_0^1(\Omega)$, and *weak* advection–reaction with solution in $L^2(\Omega)$.

Let $\mathbb{U}_h \subset \mathbb{U}$ and $\mathbb{V}_h \subset \mathbb{V}$ be discrete subspaces, and assume:

$$\begin{cases} \dim(\mathbb{V}_h) > \dim(\mathbb{U}_h), & \text{(a)} \\ \exists\, \beta_h > 0 \ : \ \inf_{w_h \in \mathbb{U}_h} \sup_{v_h \in \mathbb{V}_h} \dfrac{b(w_h, v_h)}{\|w_h\|_{\mathbb{U}} \|v_h\|_{\mathbb{V}}} \geq \beta_h\,. & \text{(b)} \end{cases} \tag{41}$$

For each $\xi \in \mathbb{X}$, we consider an equivalent (weighted) inner product $(\cdot, \cdot)_{\mathbb{V},\xi}$ on $\mathbb{V}$, i.e., such that its induced norm

$$\mathbb{V} \ni v \mapsto \|v\|_{\mathbb{V},\xi} := \sqrt{(v, v)_{\mathbb{V},\xi}}$$

satisfies (15). The minimal-residual method that we consider is then: Given $\xi \in \mathbb{X}$, find $r_h \in \mathbb{V}_h$ and $u_h \equiv S_h(\xi) \in \mathbb{U}_h$ such that

$$\begin{cases} \big(r_h, v_h\big)_{\mathbb{V},\xi} + b(u_h, v_h) & = f(v_h)\,, \quad \forall v_h \in \mathbb{V}_h\,, \quad \text{(a)} \\ b(w_h, r_h) & = 0\,, \qquad \forall w_h \in \mathbb{U}_h\,. \quad \text{(b)} \end{cases} \tag{42}$$

This has the structure of (4) for $\hat{\mathbb{V}} := \mathbb{V}_h$ and $a(\xi; r, v) := (r, v)_{\mathbb{V},\xi}$. Because $(\cdot, \cdot)_{\mathbb{V},\xi}$ and $\|\cdot\|_{\mathbb{V}_{h,\xi}}$ depend on ξ, we refer to the above as a *weighted* discrete-dual MinRes formulation.

Example 3.15 (*Weighted Minres for Weak Advection–Reaction*)**.** Recall the advection–reaction PDE $\beta\cdot\nabla u + cu = f$ and inflow boundary condition $u|_{\partial\Omega_-} = 0$ of Example 3.1. Under suitable conditions on β and c, this admits the following well-posed weak formulation (see, e.g., [64,65]):

$$\text{Find } u \in \mathbb{U} := L^2(\Omega) : \qquad \int_\Omega \Big(-u \operatorname{div}(\beta\, v) + c\, u\, v\Big) = \big\langle f\,,\ v\big\rangle\,, \qquad \forall v \in \mathbb{V}\,,$$

where $\mathbb{V} := \{v \in L^2(\Omega) : \beta \cdot \nabla v \in L^2(\Omega) \text{ and } v|_{\partial\Omega_+} = 0\}$, endowed with the norm $\|\beta \cdot \nabla(\cdot)\|_{L^2(\Omega)}$, $f \in \mathbb{V}^*$, and the outflow boundary is defined by

$$\partial\Omega_+ := \big\{x \in \partial\Omega : \beta(x) \cdot n(x) > 0\big\}\,.$$

Consider now a discrete trial/test pairing $\mathbb{U}_h/\mathbb{V}_h$ satisfying (41), and a mapping $\omega : L^2(\Omega) \to L^\infty(\Omega)$ such that $\omega(\xi)$ is a positive weight function for all $\xi \in L^2(\Omega)$. A weighted discrete-dual MinRes formulation is then to find $(r_h, u_h) \in \mathbb{V}_h \times \mathbb{U}_h$ such that:

$$\begin{aligned} &\int_\Omega \omega(\xi)(\beta \cdot \nabla r_h)(\beta \cdot \nabla v_h) + \int_\Omega \Big(-u_h \operatorname{div}(\beta v_h) + c\, u\, v\Big) = \big\langle f\,,\ v_h\big\rangle\,, \quad \forall v_h \in \mathbb{V}_h\,, \\ &-\int_\Omega w_h \operatorname{div}(\beta r_h) \qquad\qquad\qquad\qquad\qquad\qquad\qquad = 0\,, \qquad \forall w_h \in \mathbb{U}_h\,. \quad \square \end{aligned}$$

Example 3.16 (*Weighted Minres for Poisson Equation*)**.** Consider $f \in H^{-1}(\Omega)$, a mapping $\omega : L^2(\Omega) \to L^\infty(\Omega)$ as in Example 3.15, and discrete subspaces $\mathbb{U}_h \subset \mathbb{V}_h \subset H_0^1(\Omega)$. A weighted discrete-dual MinRes formulation for the Poisson equation with homogeneous Dirichlet boundary conditions is to find $u_h \in \mathbb{U}_h$ and $r_h \in \mathbb{V}_h$ such that

$$\begin{aligned} &\int_\Omega \omega(\xi)\nabla r_h \cdot \nabla v_h + \int_\Omega \nabla u_h \cdot \nabla v_h = \big\langle f\,,\ v_h\big\rangle\,, \quad \forall v_h \in \mathbb{V}_h\,, \\ &\int_\Omega \nabla w_h \cdot \nabla r_h \qquad\qquad\qquad\quad = 0\,, \qquad \forall w_h \in \mathbb{U}_h\,. \quad \square \end{aligned}$$

As shown in [16, Theorem 4.1], the mixed formulation (42) is equivalent to minimizing the residual as measured by a discrete-dual norm:

$$u_h = \arg\min_{w_h \in \mathbb{U}_h} \left(\sup_{v_h \in \mathbb{V}_h} \frac{|f(v_h) - b(w_h, v_h)|}{\|v_h\|_{\mathbb{V}_h,\xi}} \right). \tag{43}$$

Proposition 3.17 (*Weighted MinRes*)**.** *Let the continuous bilinear form $b : \mathbb{U} \times \mathbb{V} \to \mathbb{R}$ and the pairing $(\mathbb{U}_h, \mathbb{V}_h)$ satisfy* (41). *Consider a parameterized set of equivalent inner-products $\{(\cdot,\cdot)_{\mathbb{V},\xi} : \xi \in \mathbb{X}\}$, whose induced norms $\|\cdot\|_{\mathbb{V},\xi}$ satisfy* (15) *for some equivalence constants $C_{1,\xi} > 0$ and $C_{2,\xi} > 0$. Let $A : \mathbb{X} \to \mathcal{L}(\mathbb{V}; \mathbb{V}^*)$ be defined by $A(\xi)v := (v,\cdot)_{\mathbb{V},\xi} \in \mathbb{V}^*$, for all $\xi \in \mathbb{X}$ and $v \in \mathbb{V}$. Then, the following statements hold true:*

(i) The mixed discrete formulation (42) *is well-posed.*
(ii) If there exist uniform constants $\tilde{C}_1 > 0$ and $\tilde{C}_2 > 0$ such that $C_{1,\xi} \geq \tilde{C}_1$ and $C_{2,\xi} \leq \tilde{C}_2$ for all $\xi \in \mathbb{X}$, then the solution $u_h \equiv S_h(\cdot)$ to problems (42) and (43) *is uniformly bounded on $\mathbb{X}$.*
(iii) Additionally, if $A(\cdot)$ is differentiable, then $S_h(\cdot)$ is also differentiable. Moreover, if $A'(\cdot)$ is uniformly bounded and Lipschitz-continuous, then also $S_h'(\cdot)$ is uniformly bounded and Lipschitz continuous. □

Proof. See Appendix A.8. ■

Remark 3.18 (*Neural Control of Weighted MinRes*)**.** Proposition 3.17 guarantees that the conditions of Propositions 2.9 and 2.11 are satisfied, hence Corollary 2.12 applies to the neural optimization of the above weighted minimal-residual formulation. In particular, this means that it can be applied to the PDEs in Examples 3.15 and 3.16, provided the weight $\omega(\xi)$ is such that the induced norm $\|\cdot\|_{\mathbb{V},\xi}$ and operator $A(\xi)$ (defined in Proposition 3.17) satisfy the stated nontrivial conditions. It turns out that these conditions hold true when $\omega(\cdot)$ satisfies the same three assumptions as for weighted *least squares*; recall Proposition 3.4 (and the subsequent Remarks 3.5, 3.6 and 3.7). The next remark demonstrates this in further detail. □

Remark 3.19 (*Weighted $H^1(\Omega)$ Inner-Product*)**.** In this remark we show that the conditions in Proposition 3.17 hold when $(\cdot,\cdot)_{\mathbb{V},\xi}$ is a suitably-weighted H^1 inner-product.

Consider a differentiable mapping $\omega : L^2(\Omega) \to L^\infty(\Omega)$, such that, for $\omega_{\max} > \omega_{\min} > 0$ and $\omega'_\infty, \omega_L > 0$,

- $\omega_{\min} \leq \omega(\xi) \leq \omega_{\max}$, for all $\xi \in L^2(\Omega)$;
- $\|\omega'(\xi)\|_{\mathcal{L}(L^2(\Omega);L^\infty(\Omega))} \leq \omega'_\infty$, for all $\xi \in L^2(\Omega)$;
- $\|\omega'(\xi_1) - \omega'(\xi_2)\|_{\mathcal{L}(L^2(\Omega);L^\infty(\Omega))} \leq \omega_L \|\xi_1 - \xi_2\|_{L^2(\Omega)}$, for all $\xi_1, \xi_2 \in L^2(\Omega)$.

The construction of such mappings was discussed in Remarks 3.6 and 3.7.

Given $\xi \in L^2(\Omega)$, consider the weighted $H^1(\Omega)$ inner-product

$$(v_1, v_2)_{H^1,\xi} := \int_\Omega \omega(\xi)\nabla v_1 \cdot \nabla v_2 + \int_\Omega v_1 v_2.$$

Observe that

$$\min\{1,\omega_{\min}\}\|v\|^2_{H^1} \leq (v,v)_{H^1,\xi} \leq \max\{1,\omega_{\max}\}\|v\|^2_{H^1}, \quad \forall v \in H^1(\Omega).$$

Hence, statement (ii) of Proposition 3.17 is satisfied with $\tilde{C}_1 = \sqrt{\min\{1,\omega_{\min}\}}$ and $\tilde{C}_2 = \sqrt{\max\{1,\omega_{\max}\}}$.

On the other hand, given $\xi \in L^2(\Omega)$, the operator $A(\xi)$ is defined by the following action:

$$A(\xi)v = \big(\omega(\xi)\nabla v, \nabla(\cdot)\big)_{L^2(\Omega)} + (v,\cdot)_{L^2(\Omega)}, \quad \forall v \in H^1(\Omega).$$

Therefore, is easy to see that $A(\cdot)$ satisfies the statement (iii) of Proposition 3.17. Indeed, observe that $A(\cdot)$ is differentiable and $[A'(\xi)\eta]v = \big([\omega'(\xi)\eta]\nabla v, \nabla(\cdot)\big)_{L^2(\Omega)}$ for any direction $\eta \in L^2(\Omega)$. Moreover, $A'(\cdot)$ is uniformly bounded and Lipschitz-continuous, since $\omega'(\cdot)$ is uniformly bounded and Lipschitz continuous.

Of course, for any $v_1, v_2 \in H^1(\Omega)$, we could have chosen the following equivalent inner-products, for which one can prove similar results:

$$(v_1, v_2)_{H^1,\xi} := \big(\nabla v_1, \nabla v_2\big)_{L^2(\Omega)} + \big(\omega(\xi)v_1, v_2\big)_{L^2(\Omega)},$$
$$(v_1, v_2)_{H^1,\xi} := \big(\omega(\xi)\nabla v_1, \nabla v_2\big)_{L^2(\Omega)} + \big(\omega(\xi)v_1, v_2\big)_{L^2(\Omega)}.$$

Also, for $H^1_0(\Omega)$, one can consider $\big(\omega(\xi)\nabla v_1, \nabla v_2\big)_{L^2(\Omega)}$, as in Example 3.16. Finally, for the graph space $\mathbb{V}$ defined in Example 3.15, one can consider the weighted inner product $\big(\omega(\xi)\,\beta\cdot\nabla v_1\,,\ \beta\cdot\nabla v_2\big)_{L^2(\Omega)}$ provided $\|\beta\cdot\nabla(\cdot)\|_{L^2(\Omega)}$ is a norm on $\mathbb{V}$. □

4. Numerical results

In this section, we consider numerical examples for the advection–reaction PDE in 1-D and 2-D. We consider both weighted least squares (Example 3.1) and weighted residual minimization (Example 3.15).[31] We construct mappings $\varpi, \omega : L^2(\Omega) \to L^\infty(\Omega)$ as explained in Remarks 3.6 and 3.7.

4.1. Quantities of interest (point values)

The examples in this section aim to incorporate knowledge of data, in particular, the imposition of the exact value of the solution at some point in the domain.

4.1.1. Weighted least squares

Let $\Omega = (0, 1) \subset \mathbb{R}$ and $c > 0$. Consider the advection–reaction problem

$$\begin{cases} u' + c\,u &= c \quad \text{in } \Omega\,, \\ \quad u(0) &= 0\,. \end{cases} \tag{44}$$

Since the exact solution to (44) is $u(x) = 1 - \exp(-cx)$, we observe that $u(x) \to 1$ when $r \to +\infty$, for all $x > 0$. Hence, for $c > 0$ sufficiently large, the exact solution has a boundary layer in the neighborhood of $x = 0$.

Let $\mathbb{U}_h \subset H^1_{(0}(\Omega) = \{w \in H^1(\Omega) : w(0) = 0\}$ be the lowest-order conforming subspace of continuous piecewise linear functions on the uniform mesh of N elements of size $h = 1/N$. We use the weighted least squares method from (21), with practical weight function

$$\omega\big(\xi(x)\big) := 1 + \frac{M}{1 + \exp(-\xi(x))}\,, \qquad M > 0, \tag{45}$$

which fits (28) in Remark 3.7 with

$$\sigma(\cdot) = \frac{1}{1 + \exp(-(\cdot))}\,. \tag{46}$$

It is well known that the standard least squares solution (i.e., the one with $\omega(\xi) \equiv 1$) will exhibit overshoots around the boundary layer. Aiming to remedy this situation, and assuming prior knowledge of the value that the exact solution takes at $x = h$, we choose a cost functional that measures the distance to the exact solution at the point value $x = h$. In fact, we consider

$$j(\xi) := \frac{1}{2}\Big(u(h) - u_{h,\xi}(h)\Big)^2 + \frac{\alpha}{2}\|\xi\|^2_{L^2}\,, \qquad \alpha \geq 0.$$

Let $\mathcal{M}_8$ be the set of neural network functions with one hidden layer, 8-neurons, and ReLU activation, i.e.,

$$\mathcal{M}_8 := \left\{ \eta_8(x) = \sum_{j=1}^{8} c_j \mathrm{ReLU}(W_j x + b_j) \,\middle|\, c_j, W_j, b_j \in \mathbb{R} \right\}. \tag{47}$$

We then consider the neural optimization of $j(\cdot)$; see Definition 2.2.

For our first experiment, we choose a finite element space $\mathbb{U}_h$ consisting of $N = 16$ elements of size $h = 1/16$. We set $c = 160$ and $\alpha = 0$. We compute least squares approximations (1-D version of formulation in Example 3.1) for several configurations of the weight function (45), varying the M constant. Fig. 1 (left) shows that the weight needs to have enough room for variability ($M = 100$) in order to pull down the cost functional to zero (see also the associated weight functions at the right panel). Fig. 1 (middle) shows that our strategy is effective in reducing the overshoots of the finite element solution.

For the second experiment of this section, we fix $M = 100$ and we investigate variations of the α-parameter. Fig. 2 (left) suggest that the L^2-norm of ξ has to be able to reach high values (case when $\alpha = 0$) in order to pull down to zero the cost functional. This is also related to allowing the weight to have more variability. Fig. 2 (middle) shows the impact of α reducing the overshoots of the finite element solution (the smaller α, the better). The associated weight functions are depicted in Fig. 2 (right).

[31] Weighted Galerkin is not considered in view of Remark 3.12.

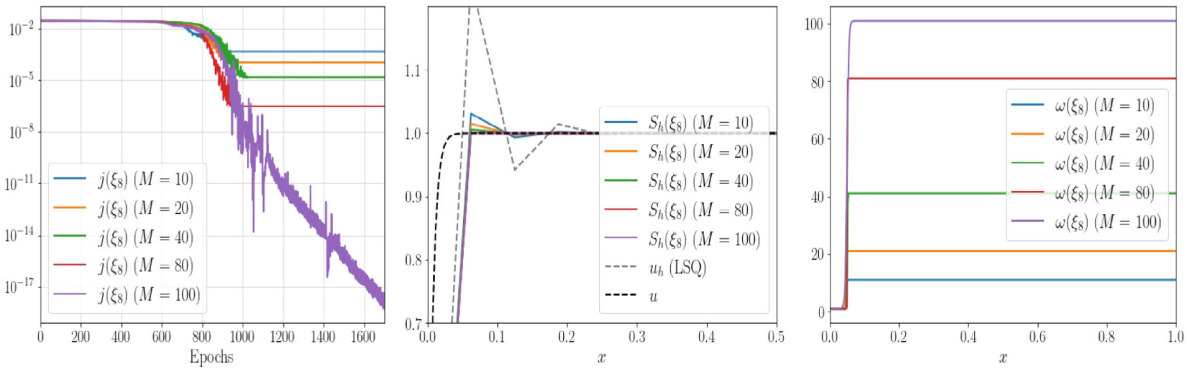

Fig. 1. Point value control for weighted least squares. Minimization of the cost functional for several values of M (left). Overshoot control of the discrete solutions (middle). Associated weight functions (right).

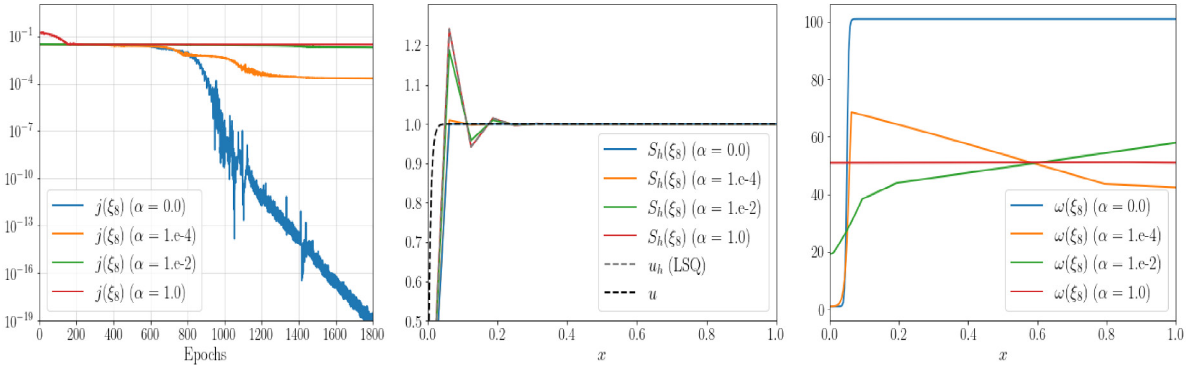

Fig. 2. Point value control for weighted least squares. Minimization of the cost functional for several values of α (left). Overshoot control of the discrete solutions (middle). Associated weight functions (right).

For the third experiment, we investigate the use of the integral operator (26) for the mapping $\omega : L^2(\Omega) \to L^\infty(\Omega)$, with $\sigma(\cdot)$ given by (46). We use a sufficiently large number of quadrature points (i.e., (27) with $N = 4$), when computing the integral in $\omega(\xi)$. We vary the kernel width r, and are particularly interested in $r \to 0$, upon which $\omega(\xi)$ converges to the practical weight function (45) as used above. Fig. 3 shows a very minor effect of r on the results (left and middle of Fig. 3). There is only a minor deviation visible for the result of $\omega(\xi_8)$ with $r = 10^{-1}$, which is attributed to the use of a tolerance in the optimizer. Furthermore, the results for the neural networks themselves (right of Fig. 3) seem to converge upon $r \to 0$ (note that the large variations in ξ are not at all noticed in the discrete solutions, because $\omega(\xi_8)$ enters the method).

Finally, we study the convergence upon varying the number of quadrature points, i.e., $N = 1, 2, 3$, and 4 in (27). Recall that $N = 1$ coincides with the practical weight function (45) as used above. We fix the kernel width at $r = 10^{-3}$ (similar results (not shown) are obtained for other values of r). Fig. 4 shows again a very minor dependence on N. The results for the discrete solutions (left) and for $\omega(\xi_8)$ (middle) are all nearly the same, while there is very quick converge for ξ_8 itself as $N \to 1$.

These latter numerics show that practical weight functions perform equally well compared to the integral operators as covered by theory. This supports our conjecture stated at the end of Remark 3.7, that the results of the theory may apply to practical weight functions. (The remainder of the numerical experiments are performed with practical weight functions.)

4.1.2. Weighted MinRes

This experiment has exactly the same configuration of the previous experiment in Section 4.1.1, except that $S_h(\xi)$ is computed with the discrete-dual minimal residual methodology. First, the approximation (*trial*) space $\mathbb{U}_h \subset L^2(\Omega)$

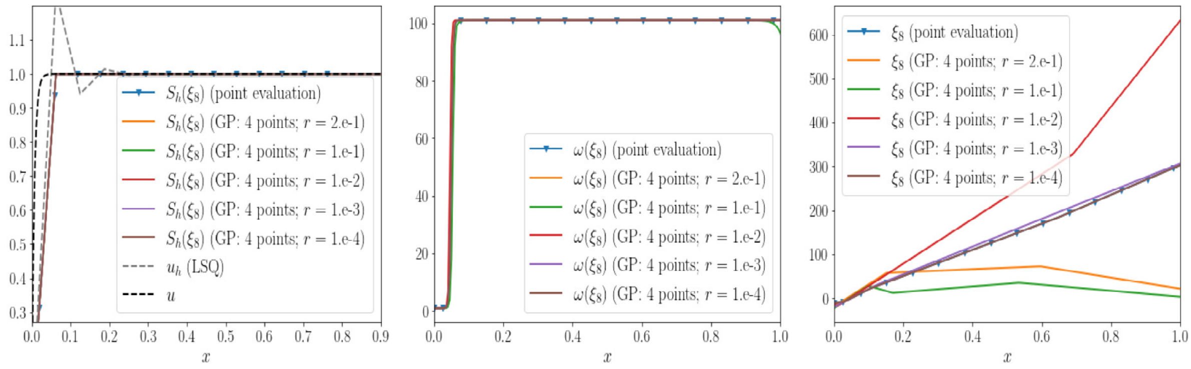

Fig. 3. Point value control for weighted least squares using an integral operator for the weight function (approximated with 4 Gaussian quadrature points (GP)). Convergence study as kernel $r \to 0$. Discrete solutions (left). Associated weight functions (middle). Associated neural network functions (right).

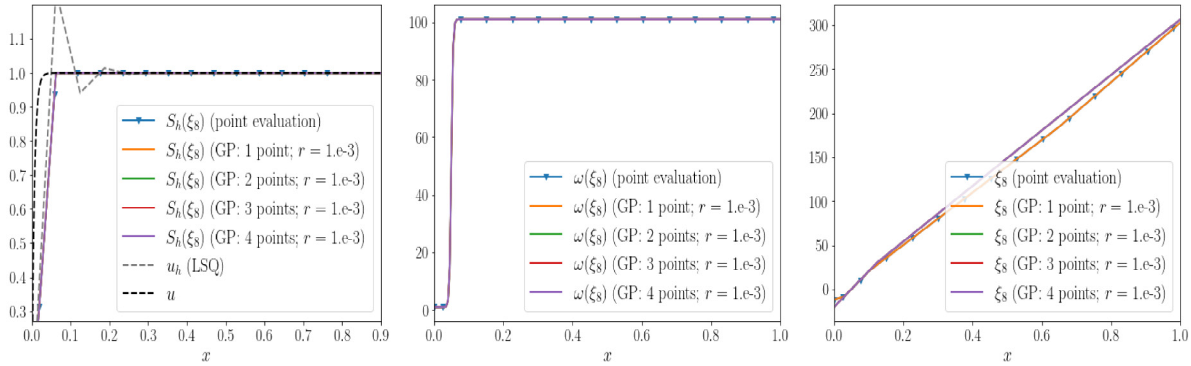

Fig. 4. Point value control for weighted least squares using an integral operator for the weight function (with fixed kernel width $r = 10^{-3}$). Convergence study for number of Gaussian quadrature points (GP). Discrete solutions (left). Associated weight functions (middle). Associated neural network functions (right).

corresponds to the lowest-order space of piecewise constants functions over the mesh. Additionally, we make use of a discrete *test* space $\mathbb{V}_h \subset H^1_{0)}(\Omega) := \{v \in H^1(\Omega) : v(1) = 0\}$ consisting in conforming piecewise linear functions over the refined uniform mesh of $2N = 32$ elements. The weighted discrete-dual residual minimization formulation that computes $S_h(\xi)$ is as follows (1-D version of formulation in Example 3.15): Find $r_h \in \mathbb{V}_h$ and $u_h \equiv S_h(\xi) \in \mathbb{U}_h$ such that

$$\begin{cases} \displaystyle\int_0^1 \omega(\xi) r_h' v_h' - \int_0^1 u_h (v_h' - c\, v_h) & = c \displaystyle\int_0^1 v_h\,, \quad \forall v_h \in \mathbb{V}_h\,, \\ \displaystyle -\int_0^1 w_h (r_h' - c\, r_h) & = 0\,, \qquad\quad \forall w_h \in \mathbb{U}_h\,. \end{cases} \tag{48}$$

As in the previous Section 4.1.1, the computation of S_h is carried out for several configurations of the weight function $\omega(\xi)$ (see (45)), varying its M constant. Fig. 5 (left) shows that larger values of M allow to pull down faster the cost functional in the training procedure (see also the associated weight functions at the right panel). Fig. 5 (middle) shows how the overshoots of the finite element solutions are controlled.

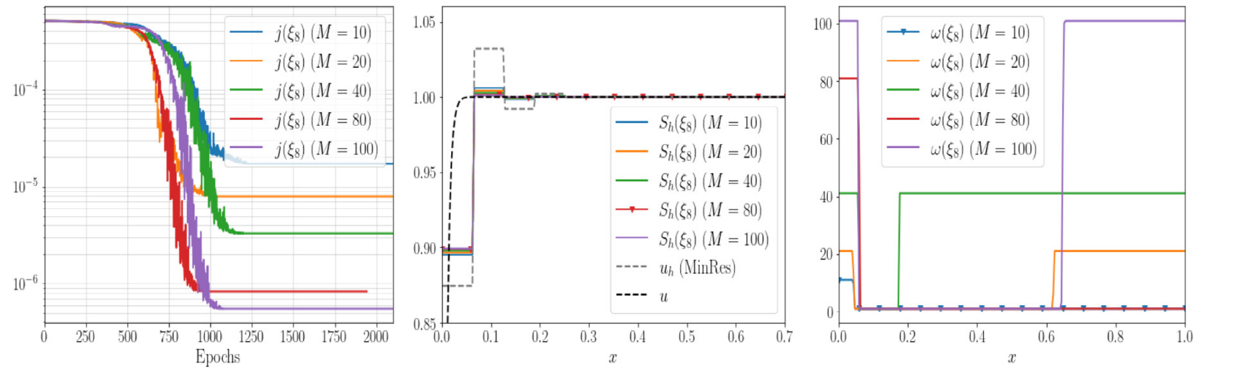

Fig. 5. Point value control for weighted discrete-dual residual minimization. Optimization of the cost functional for several values of M (left). Overshoot control of the discrete solutions (middle). Associated weight functions (right).

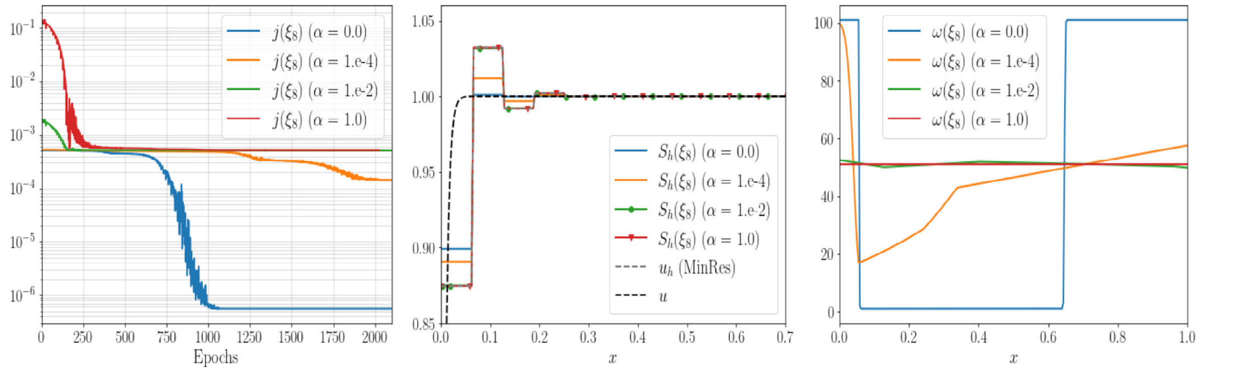

Fig. 6. Point value control for weighted discrete-dual residual minimization. Optimization of the cost functional for several values of α (left). Overshoot control of the discrete solutions (middle). Associated weight functions (right).

The second experiment investigates variations of the α-parameter. Fig. 6 (left) suggests that the smaller α, the better for faster minimization of $j(\cdot)$. Fig. 6 (middle) shows the impact of α reducing the overshoots of the finite element solution; while Fig. 6 (right) exposes the associated weight functions.

4.2. Convergence of artificial neural networks

In this section, we study the quasi-optimal convergence behavior expected from theory as neural networks become larger.

Let $\Omega := (0, 1) \subset \mathbb{R}$ and consider

$$\begin{cases} u' = & f \quad \text{in } \Omega\,, \\ u(0) = & 0\,, \end{cases} \tag{49}$$

with $f(x) := \pi \sin(\pi x)$. Notice the exact solution to (49) is $u(x) = 1 - \cos(\pi x)$.

Let $H^1_{(0}(\Omega) = \{w \in H^1(\Omega) : w(0) = 0\}$ and let $\mathbb{U}_h \subset H^1_{(0}(\Omega)$ be the finite element subspace of continuous piecewise linear functions on a uniform mesh consisting of N elements of size $h = 1/N$. We consider the weighted

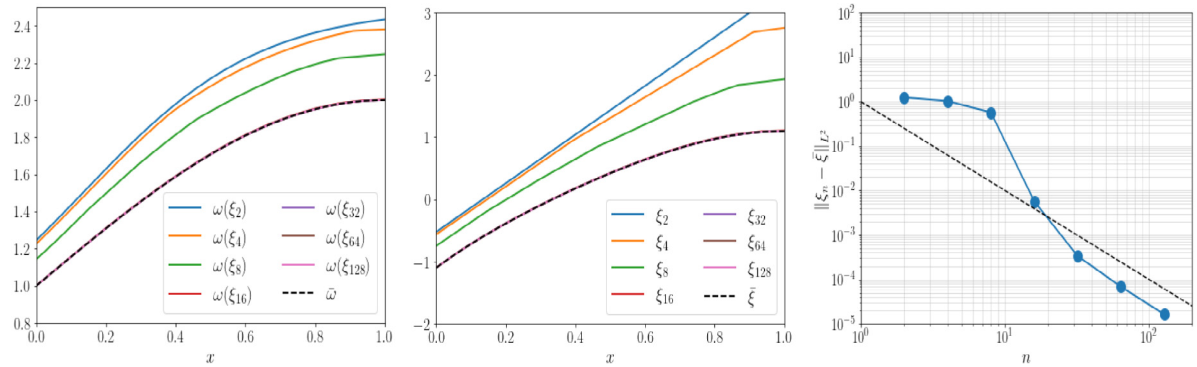

Fig. 7. As $n \to +\infty$, convergence of: $\omega(\xi_n) \to \bar{\omega}$ (left); $\xi_n \to \bar{\xi}$ (middle); and $\|\xi_n - \bar{\xi}\|_{L^2} \to 0$ (right).

least squares formulation (1-D version of Example 3.1):

$$\begin{cases} \text{Find } u_h \equiv S_h(\xi) \in \mathbb{U}_h : \\ \displaystyle\int_0^1 \omega(\xi)\left(f - u_h'\right) w_h' = 0 \,, \quad \forall w_h \in \mathbb{U}_h \,, \end{cases} \tag{50}$$

where the weight function is similar as used in Section 4.1.1:

$$\omega\big(\xi(x)\big) := \frac{1}{2} + \frac{2}{1 + \exp(-\xi(x))} \,. \tag{51}$$

Let $\mathcal{M}_n$ be the set of neural network functions with one hidden layer, n-neurons, and ReLU activation, i.e.,

$$\mathcal{M}_n := \left\{ \eta_n(x) = \sum_{j=1}^{n} c_j \mathrm{ReLU}(W_j x + b_j) \,\middle|\, c_j, W_j, b_j \in \mathbb{R} \right\}.$$

Consider the cost functional

$$j(\xi) := \frac{1}{2} \int_0^1 \bar{\omega}(x) \Big(f(x) - u_{h,\xi}'(x) \Big)^2 \,\mathrm{d}x \,, \tag{52}$$

where we choose $\bar{\omega}(x)$ as smooth function, i.e.,

$$\bar{\omega}(x) = 1 + \sin(\pi x/2) \,,$$

to allow for optimal convergence behavior as $n \to \infty$. Indeed, since the minimization of the cost functional and the discrete problem (50) are both weighted least squares formulations of the same problem (49), we expect that $\omega(\xi_n(\cdot)) \to \bar{\omega}(\cdot)$ as $n \to \infty$, which is confirmed in Fig. 7 (left). Additionally, solving for ξ_n we get (see Fig. 7 (middle))

$$\xi_n(x) \longrightarrow \bar{\xi}(x) = -\ln\left(\frac{2}{\sin(\pi x/2) + 1/2} - 1 \right), \qquad \text{as } n \to +\infty.$$

To initialize the minimization algorithm, we have chosen $\xi_n^{(0)} \in \mathcal{M}_n$ as the neural network function that (linearly) interpolates $\bar{\xi}$ on a uniform mesh of $n-1$ subintervals of Ω (i.e., having n uniformly distributed nodal points). The space $\mathbb{U}_h$ has been fixed to $N = 16$ uniform elements.

The error $\|\bar{\xi} - \xi_n\|_{L^2}$ is depicted in Fig. 7 (right), which confirms quasi-optimal convergence behavior; indeed the asymptotic rate is $O(n^{-1/2})$, which is expected for our single-hidden-layer ReLU neural network approximations (continuous piecewise-linear polynomials).

4.3. L^1-based controls

We now consider numerical experiments that incorporate a stabilization mechanism. We note that the employed cost functionals use an L^1-type norm, and hence do not fit within the currently presented theory. However, our numerics show that desirable quasi-minimizers have been computed.

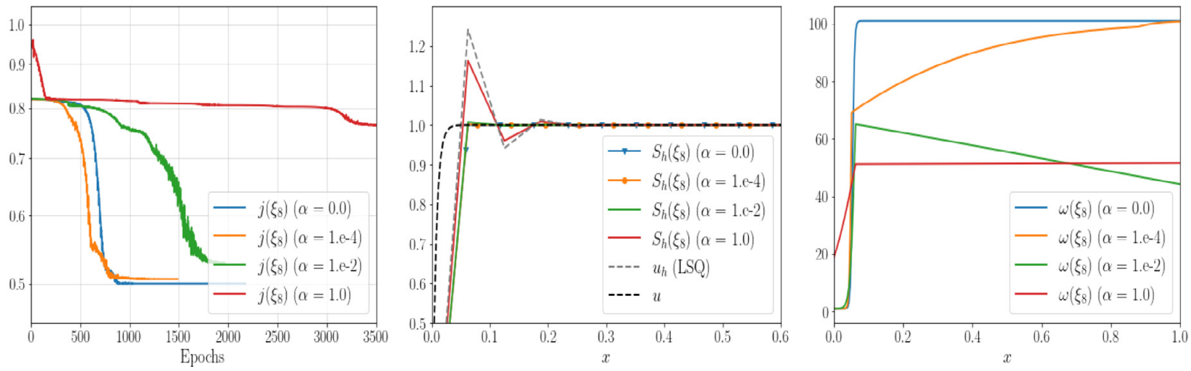

Fig. 8. Total variation control. Minimization of the cost functional for several values of α (left). Overshoot control of the discrete solutions (middle). Associated weight functions (right).

4.3.1. *Minimizing the total variation*

In this section we work exactly with the same problem of Section 4.1.1, but we introduce a modification in the cost functional. Instead of minimizing the distance to the exact solution of a particular point value (supervised training), we take an unsupervised approach by minimizing the total variation of u_h (i.e., the L^1-norm of u_h').[32] Hence, we consider the cost functional:

$$j(\xi) := \left\| u_{h,\xi}' \right\|_{L^1} + \frac{\alpha}{2}\|\xi\|_{L^2}^2 , \qquad \alpha \geq 0.$$

For a fixed value of $M = 100$, Fig. 8 (left) shows the behavior of the cost functional for different values of α, indicating that this value has to be chosen small enough to speed up the minimization process. Fig. 8 (middle) shows the quality of overshoot reduction for several values of α; while Fig. 8 (right) exposes the associated weight functions.

4.3.2. *Minimizing the L^1 residual (1D domain)*

This experiment is inspired by the example of Guermond [8, Section 4.6.2]. As usual $\Omega = (0, 1) \subset \mathbb{R}$. The idea is to interpret the following overconstrained problem:

$$\begin{cases} \quad u' + u & = 1 \quad \text{in } \Omega\,, \\ u(0) = u(1) & = 0\,, \end{cases} \tag{53}$$

as the limiting case of a *vanishing viscosity regime* (i.e., an equivalent problem having an extra $-\varepsilon u''$ term that vanishes as $\varepsilon \to 0^+$). Of course, the exact solution that we want to approach ($u(x) = 1 - e^{-x}$) only satisfies one of the boundary conditions. However, any discrete solution in a $H_0^1(\Omega)$-conforming space must satisfy both constraints. In this case, it is well-known that the standard least squares solution to this problem does not deliver satisfactory results. To remedy this drawback, we propose a cost functional that mimics the L^1 residual minimization as proposed in [8]. Thus, our (*unsupervised*) cost functional will be

$$j(\xi) := \Big\| \underbrace{1 - u_{h,\xi} - u_{h,\xi}'}_{\text{residual}} \Big\|_{L^1} + \frac{\alpha}{2}\|\xi\|_{L^2}^2 , \qquad \alpha \geq 0.$$

We consider the weighted least squares formulation for $u_{h,\xi}$, solved on a uniform mesh of $N = 8$ elements. For a fixed $M = 1000$ constant in the weight function (45), we compute the discrete solution for several values of the α-parameter. Large values of α allow for small values of $\|\xi\|_{L^2}$, and thus the weight becomes almost constant (close to the standard least squares approach). On the other hand, small values of α allow for more variability of the weight (see Fig. 9, right), and thus, we observe that we can recover a discrete solution mimicking the vanishing viscosity case (see Fig. 9, left).

[32] It is well-known that minimizing the total variation translates into a reduction of overshoots; see, e.g., [66].

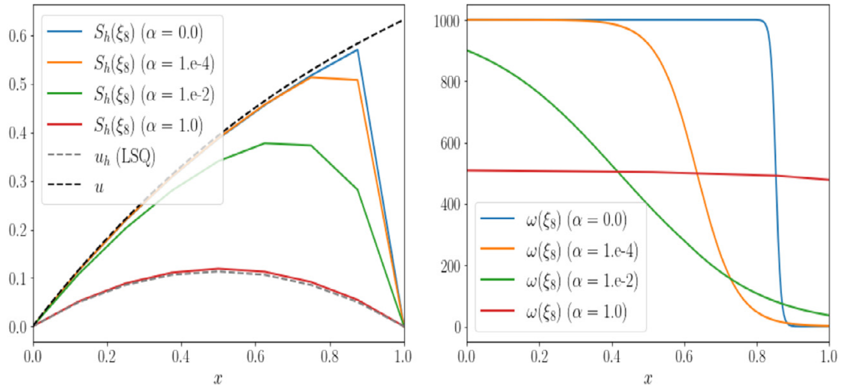

Fig. 9. Discrete weighted least squares solutions (left) and associated weight functions (right), with L^1 residual minimization control, for several values of the α-parameter.

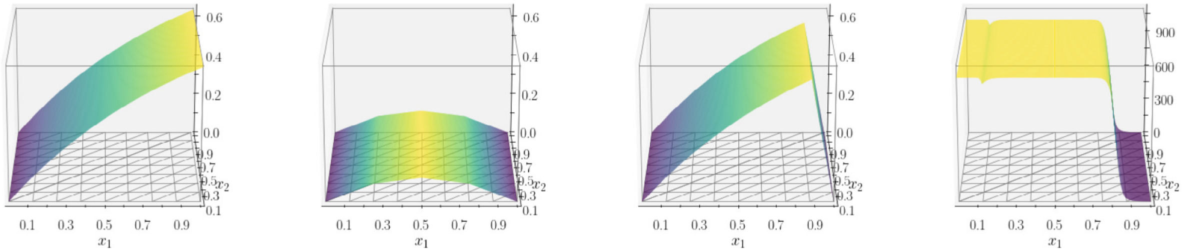

Fig. 10. Overconstrained weighted least squares for advection–reaction, with L^1 residual minimization control. From left to right: exact solution; standard overconstrained least-squares, controlled weighted least-squares, and associated weight function.

4.3.3. Minimizing L^1 residual (2D domain)

This is the two-dimensional version of the previous example in Section 4.3.2. Let $\Omega = (0,1)^2 \subset \mathbb{R}^2$. For an advection field $\vec{\beta} = (1,0)$, we consider the over-constrained problem:

$$\begin{cases} \vec{\beta}\cdot\nabla u + u &= 1 \quad \text{in} \quad \Omega\,, \\ \qquad\qquad u &= 0 \quad \text{on} \quad \{(x_1,x_2)\in\partial\Omega : x_1 = 0 \text{ or } x_1 = 1\}\,. \end{cases} \tag{54}$$

We approach (54) using a coarse (and over-constrained) finite element space of piecewise linear functions of the form

$$\mathbb{U}_h \subset \{w \in H_0^1(\Omega) : w(0,x_2) = w(1,x_2) = 0,\ \forall x_2 \in [0,1]\}.$$

We use the weighted least squares method given in Example 3.1 using the weight (45) with $M = 1000$. On the other hand, the cost functional $j(\cdot)$ for this case is defined as

$$j(\xi) := \left\| 1 - u_{h,\xi} - \beta\cdot\nabla u_{h,\xi}\right\|_{L^1} + \frac{\alpha}{2}\|\xi\|_{L^2}^2\,, \quad \alpha \geq 0.$$

The discrete neural network space where we minimize $j(\cdot)$ will be $\mathcal{M}_8$ (see (47)). Results for the $\alpha = 0$ case are depicted in Fig. 10. We observe a strong correlation with the results in [8, Figure 9].

5. Concluding remarks

The objective of this work was to introduce and analyze the neural optimization of finite element methods. We proposed a notion of quasi-minimization to enable the consideration of neural network functions as control variables, and proved a general theorem on the existence and convergence of quasi-minimizers. We applied our theory to the optimization of least squares, Galerkin, and minimal residual finite element methods, where the neural network function entered as a suitable weight within the discrete weak forms.

The notion of quasi-minimization is critical, since sets of neural-network functions are generally not closed. If instead, one aims to minimize over a (closed) linear space of classical approximations,[33] the standard notion of minimization is adequate. We have been motivated to explore the use of neural networks as these generate a new class of functions that have shown recent success in situations where classical approximations may not work (e.g., high-dimensional problems [2]). Although there are many open questions (e.g., those related to robust and efficient optimization algorithms, hence training of the neural network), there is currently a growing literature providing deeper mathematical understanding, proposing new algorithms and developing accessible relevant software.

While this paper has explored the neural optimization of finite element methods from a mostly theoretical perspective, there remain many avenues that require further work to enable a practical methodology. For example, derivatives with respect to the trainable parameters would need to be computed to utilize a gradient-based algorithm, and further work would be required to better understand the role of parameters used within the methodology (such as α in (6), M in (45) and r in (25)), and the approximation of integral-operator weight functions by practical alternatives.

Furthermore, the idea of a weighted Galerkin formulation, while straightforward, exposes itself to instability, unless the involved weight function is suitable constrained. Optimizing methods to control their stability is an interesting avenue for further research. On the other hand, weighted least squares and weighted residual minimization are without such cumbersome constraints, and seem to be the only (conforming) weighted formulations for which stability is guaranteed.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors are grateful to the reviewers for their helpful suggestions. KvdZ acknowledges helpful discussions with Dante Kalise, Hamd Alsobhi, Anne Boschman and Andrew Stuart. This research has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 777778 (MATHROCKS). IM acknowledges support from the project DI Investigación Innovadora Interdisciplinaria PUCV 2021 Nº 039.409/2021: *Nanoiónica: Un enfoque interdisciplinario*. The work by IB was partially supported by ANID FONDECYT/Postdoctorado No 3200827. The research by KvdZ was supported by the Engineering and Physical Sciences Research Council (EPSRC), UK under Grant EP/T005157/1 and EP/W010011/1.

Appendix. Proofs

A.1. Proof of Theorem 2.A

(i) Strong convexity of j implies coercivity, i.e., $j(\xi) \to +\infty$ when $\|\xi\|_{\mathbb{X}} \to +\infty$. Moreover, j is continuous in the strong topology since it is differentiable. Additionally, we know that convexity plus continuity implies that j is weakly lower semicontinuous (see, e.g. [67, Corollary 3.9]). We thus satisfy all the hypothesis of the *theorem of existence of minimizers for coercive and sequentially weakly lower semicontinuous functionals* [68, Theorem 9.3-1]. Moreover, strong convexity ensures that such a (global) minimizer $\bar{\xi} \in \mathbb{X}$ is unique. Besides, global differentiability of j implies the first-order necessary optimality condition $j'(\bar{\xi}) = 0$.

[33] That is, those spanned by a basis such as a finite element space, b-spline approximations, etc.

(ii) We know that j has a global lower bound. Thus, by the infimum property, for any $\delta_n > 0$ there must exist $\bar{\xi}_n \in \mathcal{M}_n$ such that

$$j(\bar{\xi}_n) < \inf_{\eta_n \in \mathcal{M}_n} j(\eta_n) + \frac{\delta_n}{2}. \tag{55}$$

(iii) Let $\bar{\xi} \in \mathbb{X}$ be the global minimizer and let $\bar{\xi}_n \in \mathcal{M}_n$ satisfy (7). By characterization of strong convexity, we have for all $t \in (0,1)$

$$j\big(\bar{\xi}\big) \leq j\big(t\bar{\xi} + (1-t)\bar{\xi}_n\big) \leq tj\big(\bar{\xi}\big) + (1-t)j\big(\bar{\xi}_n\big) - \frac{\gamma}{2}t(1-t)\|\bar{\xi} - \bar{\xi}_n\|^2_{\mathbb{X}}.$$

Thus, for all $t \in (0,1)$ and $\eta_n \in \mathcal{M}_n$ we get

$$\frac{\gamma}{2}t\|\bar{\xi} - \bar{\xi}_n\|^2_{\mathbb{X}} \leq j\big(\bar{\xi}_n\big) - j\big(\bar{\xi}\big) < j\big(\eta_n\big) - j\big(\bar{\xi}\big) + \frac{\delta_n}{2}. \tag{56}$$

On the other hand, using the facts that j' is L-Lipschitz and $j'\big(\bar{\xi}\big) = 0$, we deduce [68, cf. proof of Thm. 7.7-3, page 488]

$$\begin{aligned} j\big(\eta_n\big) - j\big(\bar{\xi}\big) &= \int_0^1 \big\langle j'\big(s\eta_n + (1-s)\bar{\xi}\big), \eta_n - \bar{\xi}\big\rangle ds \\ &= \int_0^1 \big\langle j'\big(s\eta_n + (1-s)\bar{\xi}\big) - j'\big(\bar{\xi}\big), \eta_n - \bar{\xi}\big\rangle ds \\ &\leq L\|\eta_n - \bar{\xi}\|^2_{\mathbb{X}} \int_0^1 s = \frac{L}{2}\|\eta_n - \bar{\xi}\|^2_{\mathbb{X}}. \end{aligned} \tag{57}$$

Hence, combining (56) with (57), taking the limit when $t \to 1$ and the infimum over all $\eta_n \in \mathcal{M}_n$, we get the estimate

$$\gamma\,\|\bar{\xi} - \bar{\xi}_n\|^2_{\mathbb{X}} < L \inf_{\eta_n \in \mathcal{M}_n} \big\|\bar{\xi} - \eta_n\big\|^2_{\mathbb{X}} + \delta_n,$$

from which (13) is deducted.

A.2. *Proof of Theorem 2.B*

We proceed to prove each one of the statements.

(i) Since $\mathbb{Z}$ and $\mathbb{X}$ are a Hilbert spaces, the quadratic maps $\mathbb{Z} \ni z \mapsto \frac{1}{2}\|z\|^2_{\mathbb{Z}}$ and $\mathbb{X} \ni \xi \mapsto \frac{1}{2}\|\xi\|^2_{\mathbb{X}}$ are differentiable. On the other hand, S_h and Q are also differentiable (Q is linear), and thus j_1 is differentiable by means of the chain rule (see, e.g. [56, Theorem 2.20]). Moreover,

$$j_1'(\eta)(\cdot) = \big(QS_h(\eta)\,,\; QS_h'(\eta)(\cdot)\big)_{\mathbb{Z}} = \big(S_h'(\eta)^\star Q^\star QS_h(\eta)\,,\;\cdot\big)_{\mathbb{X}}.$$

Thus, we conclude that j_1 is Lipschitz since

$$\begin{aligned} \big\|j_1'(\eta) - j_1'(\zeta)\big\|_{\mathbb{X}^*} &= \big\|S_h'(\eta)^\star Q^\star QS_h(\eta) - S_h'(\zeta)^\star Q^\star QS_h(\zeta)\big\|_{\mathbb{X}} \\ &\leq \big\|S_h'(\eta)^\star Q^\star Q\big(S_h(\eta) - S_h(\zeta)\big)\big\|_{\mathbb{X}} \\ &\quad + \big\|\big(S_h'(\eta) - S_h'(\zeta)\big)^\star Q^\star QS_h(\zeta)\big\|_{\mathbb{X}} \\ &\leq \|Q\|^2_{\mathcal{L}(\mathbb{U};\mathbb{Z})}\big(M^2_{S'} + L_{S'}M_S\big)\|\eta - \zeta\|_{\mathbb{X}}, \end{aligned}$$

where we have used the mean value theorem together with

- the boundedness of S_h', with bounding constant $M_{S'}$;
- the Lipschitzness of S_h', with Lipschitz constant $L_{S'}$;
- the boundedness of S_h, with bounding constant M_S.

Finally, by making $L_1 := \|Q\|^2_{\mathcal{L}(\mathbb{U};\mathbb{Z})}\big(M^2_{S'} + L_{S'}M_S\big)$, it is straightforward to see that $L_1 + \alpha$ will be a Lipschitz constant for j'.

(ii) Just observe that

$$\begin{aligned}\left\langle j'(\eta) - j'(\zeta), \eta - \zeta\right\rangle_{\mathbb{X}^*,\mathbb{X}} &= \left\langle j_1'(\eta) - j_1'(\zeta), \eta - \zeta\right\rangle_{\mathbb{X}^*,\mathbb{X}} + \alpha\|\eta - \zeta\|_{\mathbb{X}}^2 \\ &\geq (-L_1 + \alpha)\|\eta - \zeta\|_{\mathbb{X}}^2 .\end{aligned}$$

Thus, j is strongly convex whenever $\alpha > 0$ is sufficiently large.

A.3. *Proof of Proposition 2.9*

The statements (i) and (ii) are classical from Babuška–Brezzi theory (see, e.g., Ern & Guermond [6, Theorem 49.13]). To prove statement (iii), first observe that

$$\sup_{v_2 \in \hat{\mathbb{K}}} \frac{a(\xi; v_1, v_2)}{\|v_1\|_{\mathbb{V}}\|v_2\|_{\mathbb{V}}} \geq \frac{a(\xi; v_1, v_1)}{\|v_1\|_{\mathbb{V}}\|v_1\|_{\mathbb{V}}} \geq \frac{a(\xi; v_1, v_1)}{\|v_1\|_{\mathbb{V},\xi}\|v_1\|_{\mathbb{V},\xi}}(C_{1,\xi})^2 = (C_{1,\xi})^2,$$

which confirms $\alpha_h = (C_{1,\xi})^2$ in (14a). For the a priori bound, since $a(\xi; \cdot, \cdot)$ is an equivalent inner-product on $\hat{\mathbb{V}}$, consider $\hat{z} \in \hat{\mathbb{V}}$ such that

$$a(\xi; \hat{z}, \hat{v}) = b(u_h, \hat{v}), \quad \forall \hat{v} \in \hat{\mathbb{V}}.$$

Hence,

$$\sup_{\hat{v} \in \hat{\mathbb{V}}} \frac{b(u_h, \hat{v})}{\|\hat{v}\|_{\mathbb{V},\xi}} = \sup_{\hat{v} \in \hat{\mathbb{V}}} \frac{a(\xi; \hat{z}, \hat{v})}{\|\hat{v}\|_{\mathbb{V},\xi}} = \frac{a(\xi; \hat{z}, \hat{z})}{\|\hat{z}\|_{\mathbb{V},\xi}} = \frac{b(u_h, \hat{z})}{\|\hat{z}\|_{\mathbb{V},\xi}}. \tag{58}$$

Moreover,

$$a(\xi, \hat{r}, \hat{z}) = a(\xi, \hat{z}, \hat{r}) = b(u_h, \hat{r}) = 0. \tag{59}$$

Next, observe that

$$\begin{aligned}\|u_h\|_{\mathbb{U}} &\leq \frac{1}{\beta_h} \sup_{\hat{v} \in \hat{\mathbb{V}}} \frac{b(u_h, \hat{v})}{\|\hat{v}\|_{\mathbb{V}}} \leq \frac{C_{2,\xi}}{\beta_h} \sup_{\hat{v} \in \hat{\mathbb{V}}} \frac{b(u_h, \hat{v})}{\|\hat{v}\|_{\mathbb{V},\xi}} && \text{(by (14) and (15))} \\ &= \frac{C_{2,\xi}}{\beta_h} \frac{b(u_h, \hat{z})}{\|\hat{z}\|_{\mathbb{V},\xi}} = \frac{C_{2,\xi}}{\beta_h} \frac{\big(f(\hat{z}) - a(\xi, \hat{r}, \hat{z})\big)}{\|\hat{z}\|_{\mathbb{V},\xi}} && \text{(by (58) and (4))} \\ &\leq \frac{C_{2,\xi}}{C_{1,\xi}} \frac{1}{\beta_h} \frac{f(\hat{z})}{\|\hat{z}\|_{\mathbb{V}}}, && \text{(by (15) and (59))}\end{aligned}$$

from which (16) can be easily deducted.

A.4. *Proof of Proposition 2.10*

Assumption (20) implies two important facts: $A(\xi)^*$ is surjective, and the range of $A(\xi)$ is closed (see, e.g., [67, Theorem 2.21]). Therefore, by Banach closed range theorem, the range of $A(\xi)$ must be characterized by $\big(\ker A(\xi)^*\big)^{\perp}$.

Assume that Eq. (2) is satisfied. Hence, $f - Bu_h \in \mathbb{V}_h(\xi)^{\perp}$. On another hand, observe that $\ker A(\xi)^* \subset \mathbb{V}_h(\xi)$. In particular, $f - Bu_h \in (\ker A(\xi)^*)^{\perp}$, which means that $f - Bu_h$ is in the range of $A(\xi)$. Thus, there exists an $r \in \hat{\mathbb{V}}$ such that $A(\xi)r = f - Bu_h$, which is Eq. (18a). Next, given any $w_h \in \mathbb{U}_h$, there must be $v_h \in \hat{\mathbb{V}}$ such that $A(\xi)^* v_h = Bw_h$ (by surjectivity of $A(\xi)^*$). Hence, $v_h \in \mathbb{V}_h(\xi)$ by definition of this last space (see (19)). Besides,

$$\left\langle r, Bw_h\right\rangle = \left\langle r, A(\xi)^* v_h\right\rangle = \left\langle A(\xi)r, v_h\right\rangle = \left\langle f - Bu_h, v_h\right\rangle = 0,$$

which proves Eq. (18b).

Conversely, let $(r, u_h) \in \hat{\mathbb{V}} \times \mathbb{U}_h$ solve the state problem (4), or equivalently (18) in operator form. Testing with elements in $v_h \in \mathbb{V}_h(\xi)$, we get

$$\begin{aligned}\langle f, v_h\rangle &= \langle A(\xi)r, v_h\rangle + \langle Bu_h, v_h\rangle && \text{(by (18a))} \\ &= \left\langle r, A(\xi)^* v_h\right\rangle + \langle Bu_h, v_h\rangle && \text{(using the adjoint property)}\end{aligned}$$

$$= \langle r, Bw_h\rangle + \langle Bu_h, v_h\rangle \qquad \text{(by definition of } \mathbb{V}_h(\xi))$$
$$= \langle Bu_h, v_h\rangle\ . \qquad \text{(by (18b))}$$

Thus, (2) is satisfied.

A.5. Proof of Proposition 2.11

Let us start proving statements (i), (ii), and (iii) at the same time.

Recall the definition of the kernel space $\hat{\mathbb{K}} := \ker B^* \subset \hat{\mathbb{V}}$. For any $\xi \in \mathbb{X}$, consider the restricted operator $A(\xi)\big|_{\hat{\mathbb{K}}} : \hat{\mathbb{K}} \to \hat{\mathbb{K}}^*$, as well as the restriction $f\big|_{\hat{\mathbb{K}}} \in \hat{\mathbb{K}}^*$. Observe that the inf − sup condition (14) ensures that $A(\xi)\big|_{\hat{\mathbb{K}}}$ is a boundedly invertible linear operator. Thus, given a direction $\eta \in \mathbb{X}$ and $t \in \mathbb{R}$, from the first equation of the mixed system (18) (restricted to $\hat{\mathbb{K}}$) we obtain that

$$A(\xi + t\eta)\big|_{\hat{\mathbb{K}}} R_h(\xi + t\eta) = f\big|_{\hat{\mathbb{K}}} \tag{60a}$$
$$A(\xi)\big|_{\hat{\mathbb{K}}} R_h(\xi) = f\big|_{\hat{\mathbb{K}}}\ . \tag{60b}$$

In particular, continuity of $A(\cdot)$ implies continuity of $R_h(\cdot)$. Moreover, using the inf − sup condition (14), it is clear that

$$\|R_h(\cdot)\|_{\mathbb{V}} \le \frac{\|f\|_{\hat{\mathbb{V}}^*}}{\alpha_h(\cdot)}\ . \tag{61}$$

Next, adding the term $A(\xi)\big|_{\hat{\mathbb{K}}} R_h(\xi + t\eta)$ on both sides of Eq. (60a), rearranging it, and subtracting Eq. (60b) we get

$$R_h(\xi + t\eta) - R_h(\xi) = \left[A(\xi)\big|_{\hat{\mathbb{K}}}\right]^{-1} \left(A(\xi)\big|_{\hat{\mathbb{K}}} - A(\xi + t\eta)\big|_{\hat{\mathbb{K}}}\right) R_h(\xi + t\eta),$$

from which, if $A'(\xi)\eta$ exists, it implies that $R_h(\cdot)$ has a Gâteaux derivative and

$$R_h'(\xi)\eta = -\left[A(\xi)\big|_{\hat{\mathbb{K}}}\right]^{-1} A'(\xi)\eta\Big|_{\hat{\mathbb{K}}} R_h(\xi). \tag{62}$$

Finally, if $A(\cdot)$ is Gâteaux-differentiable at ξ, then using the inf − sup condition (14), the boundedness of the linear operator $A'(\xi)$, and the estimate (61), we obtain

$$\|R_h'(\xi)\eta\|_{\mathbb{V}} \le \frac{\|A'(\xi)\eta\|_{\mathcal{L}(\hat{\mathbb{V}};\hat{\mathbb{V}}^*)}\|R_h(\xi)\|_{\mathbb{V}}}{\alpha_h} \le \frac{\|A'(\xi)\|\,\|f\|_{\mathbb{V}^*}}{\alpha_h^2}\|\eta\|_{\mathbb{X}}\,, \tag{63}$$

which proves that $R_h(\cdot)$ is Gâteaux-differentiable at ξ. Besides, if $A'(\cdot)$ and $\alpha_h^{-1}(\cdot)$ are uniformly bounded on $\mathbb{X}$, then $R_h'(\cdot)$ is uniformly bounded on $\mathbb{X}$.

Now is the turn of S_h. From the mixed system (18), we deduce

$$BS_h(\xi + t\eta) = \quad f - A(\xi + t\eta)R_h(\xi + t\eta)$$
$$BS_h(\xi) = \quad f - A(\xi)R_h(\xi).$$

Since B is boundedly invertible onto its closed range, we get

$$S_h(\xi + t\eta) - S_h(\xi) = B^{-1}\Big([A(\xi) - A(\xi + t\eta)]R_h(\xi + t\eta) + A(\xi)[R_h(\xi) - R_h(\xi + t\eta)]\Big).$$

Therefore, if $A'(\xi)\eta$ exists, then we already know that $R_h'(\xi)\eta$ exists, and thus

$$S_h'(\xi)\eta = B^{-1}\Big(-[A'(\xi)\eta]R_h(\xi) - A(\xi)R_h'(\xi)\eta\Big). \tag{64}$$

Moreover, if $A(\cdot)$ is Gâteaux-differentiable, then using the inf − sup condition (14) and the estimate (63), we get

$$\begin{aligned}
\|S_h'(\xi)\eta\|_{\mathbb{U}} &\le \frac{1}{\beta_h}\|B[S_h'(\xi)\eta]\|_{\hat{\mathbb{V}}^*} \\
&\le \frac{\|A'(\xi)\|\,\|R_h(\xi)\|_{\mathbb{V}} + \|A(\xi)\|_{\mathcal{L}(\hat{\mathbb{V}};\hat{\mathbb{V}}^*)}\|R_h'(\xi)\|_{\mathcal{L}(\mathbb{X};\hat{\mathbb{V}})}}{\beta_h}\|\eta\|_{\mathbb{X}} \qquad (65)\\
&\le \frac{\|A'(\xi)\|\,\|f\|_{\mathbb{V}^*}}{\alpha_h\beta_h}\left(1 + \frac{\|A(\xi)\|_{\mathcal{L}(\hat{\mathbb{V}};\hat{\mathbb{V}}^*)}}{\alpha_h}\right)\|\eta\|_{\mathbb{X}}\,,
\end{aligned}$$

which proves that $S_h(\cdot)$ is Gâteaux-differentiable. Besides, it is clear from (65) that $\|S_h'(\cdot)\|_{\mathcal{L}(\mathbb{X};\mathbb{U})}$ will be uniformly bounded on $\mathbb{X}$ whenever $\|A(\cdot)\|_{\mathcal{L}(\hat{\mathbb{V}};\hat{\mathbb{V}}^*)}$ and $\|A'(\cdot)\|$ are uniformly bounded on $\mathbb{X}$, as well as $\alpha_h^{-1}(\cdot)$.

(iv) Let us prove Lipschitzness. Using (62), observe that for any $\xi_1, \xi_2, \eta \in \mathbb{X}$ we have

$$A(\xi_2)\big|_{\hat{\mathbb{K}}}\big(R_h'(\xi_1) - R_h'(\xi_2)\big)\eta = \big[A'(\xi_2) - A'(\xi_1)\big]\eta\Big|_{\hat{\mathbb{K}}} R_h(\xi_2) + \big[A(\xi_2) - A(\xi_1)\big]\Big|_{\hat{\mathbb{K}}} R_h'(\xi_1)\eta \\ + A'(\xi_1)\eta\Big|_{\hat{\mathbb{K}}}\big[R_h(\xi_2) - R_h(\xi_1)\big].$$

Hence,

$$\|R_h'(\xi_1) - R_h'(\xi_2)\|_{\mathcal{L}(\mathbb{X};\hat{\mathbb{V}})} \le \frac{\|R_h(\xi_2)\|_{\mathbb{V}}}{\alpha_h(\xi_2)}\|A'(\xi_1) - A'(\xi_2)\| \tag{66a}$$

$$+ \frac{\|R_h'(\xi_1)\|_{\mathcal{L}(\mathbb{X};\hat{\mathbb{V}})}}{\alpha_h(\xi_2)}\|A(\xi_1) - A(\xi_2)\|_{\mathcal{L}(\hat{\mathbb{V}};\hat{\mathbb{V}}^*)} \tag{66b}$$

$$+ \frac{\|A'(\xi_1)\|}{\alpha_h(\xi_2)}\|R_h(\xi_1) - R_h(\xi_2)\|_{\mathbb{V}}\,. \tag{66c}$$

Recall that under our hypothesis, $\alpha_h^{-1}(\cdot)$, $R_h(\cdot)$, $R_h'(\cdot)$, and $A'(\cdot)$ are all uniformly bounded on $\mathbb{X}$. Therefore, the first term on the right hand side (expression (66a)) is Lipschitz by the Lipschitz assumption on $A'(\cdot)$; the second term (expression (66b)) is Lipschitz as a consequence of the mean value theorem on $A(\cdot)$ and the uniform boundedness of $A'(\cdot)$; while the last term (expression (66c)) is Lipschitz by the mean value theorem on $R_h(\cdot)$ and the uniform boundedness of $R_h'(\cdot)$.

Finally, to prove the Lipschitzness of $S_h'(\cdot)$, we use (64) to write

$$B\big(S_h'(\xi_1)\eta - S_h'(\xi_2)\eta\big) = [A'(\xi_2)\eta]\big(R_h(\xi_2) - R_h(\xi_1)\big) + A(\xi_2)\big[R_h'(\xi_2)\eta - R_h'(\xi_1)\eta\big] \\ + \big[(A'(\xi_2) - A'(\xi_1))\eta\big]R_h(\xi_1) + \big[A(\xi_2) - A(\xi_1)\big]R_h'(\xi_1)\eta\,.$$

Hence,

$$\big\|S_h'(\xi_1) - S_h'(\xi_2)\big\|_{\mathcal{L}(\mathbb{X};\mathbb{U})} \le \frac{\|A'(\xi_2)\|}{\beta_h}\|R_h(\xi_1) - R_h(\xi_2)\|_{\mathbb{V}} \tag{67a}$$

$$+ \frac{\|A(\xi_2)\|_{\mathcal{L}(\hat{\mathbb{V}};\hat{\mathbb{V}}^*)}}{\beta_h}\|R_h'(\xi_1) - R_h'(\xi_2)\|_{\mathcal{L}(\mathbb{X};\hat{\mathbb{V}})} \tag{67b}$$

$$+ \frac{\|R_h(\xi_1)\|_{\mathbb{V}}}{\beta_h}\|A'(\xi_1) - A'(\xi_2)\| \tag{67c}$$

$$+ \frac{\|R_h'(\xi_1)\|_{\mathcal{L}(\mathbb{X};\hat{\mathbb{V}})}}{\beta_h}\|A(\xi_1) - A(\xi_2)\|_{\mathcal{L}(\hat{\mathbb{V}};\hat{\mathbb{V}}^*)}\,. \tag{67d}$$

We recall again that $R_h(\cdot)$, $R_h'(\cdot)$, $A(\cdot)$, and $A'(\cdot)$ are all uniformly bounded on $\mathbb{X}$. Therefore, the Lipschitzness of $S_h'(\cdot)$ is implied by the following facts: the Lipschitzness of the first term on right hand side (expression (67a)) is a consequence of the mean value theorem applied to $R_h(\cdot)$ and the uniform boundedness of $R_h'(\cdot)$; the Lipschitzness of the second term (expression (67b)) is due to the previously proved Lipschitzness of $R_h'(\cdot)$; the Lipschitzness of the third term (expression (67c)) is implied by the assumed Lipschitzness of $A'(\cdot)$; and the Lipschitzness of the last term (expression (67d)) is consequence of the mean value theorem applied to A and the uniform boundedness of $A'(\cdot)$.

A.6. Proof of Proposition 3.4

Let us prove item by item.

(i) Observe that in this case, the bilinear form $a(\xi,\cdot,\cdot)$ defines a weighted inner product in $L^2(\Omega)$, for which its induced norm $\|v\|_{\mathbb{V},\xi} := \sqrt{(\varpi(\xi)v, v)_{L^2}}$ satisfies

$$\sqrt{\varpi_{\min}}\,\|v\|_{L^2} \le \|v\|_{\mathbb{V},\xi} \le \sqrt{\varpi_{\max}}\,\|v\|_{L^2}\,, \quad \forall v \in \mathbb{V} = L^2(\Omega).$$

Hence, the first inf − sup condition in (14) is satisfied with $\alpha_h = \varpi_{\min}$; see Proposition 2.9(iii) and Footnote 22.

On the other hand, we are under the assumption that the operator $B : \mathbb{H}_B \to \mathbb{V}^*$ is boundedly invertible. Hence, there must be a uniform constant $\beta > 0$ such that

$$\sup_{v\in\mathbb{V}} \frac{b(w_h, v)}{\|v\|_{\mathbb{V}}} = \|Bw_h\|_{\mathbb{V}^*} \geq \beta\|w_h\|_{\mathbb{H}_B}, \quad \forall w_h \in \mathbb{U}_h,$$

which implies the second inf − sup condition in (14).

(ii) Uniform boundedness of $S_h(\cdot)$ is a consequence of Proposition 2.9(iii). Indeed, in our particular case we get

$$\|S_h(\xi)\|_{\mathbb{H}_B} \leq \frac{\varpi_{\max}}{\varpi_{\min}} \frac{1}{\beta} \|f\|_{L^2}, \quad \forall \xi \in L^2(\Omega).$$

To show differentiability of $S_h(\cdot)$, let us recall the operator $A : \mathbb{X} \to \mathcal{L}(\mathbb{V}; \mathbb{V}^*)$ defined in Section 2.4, which in this particular case takes the form

$$A(\xi)v := (\varpi(\xi)v, \cdot)_{L^2}, \quad \forall v \in L^2(\Omega).$$

Furthermore, we have the uniform bound

$$\|A(\xi)\| = \sup_{v\in L^2(\Omega)} \frac{\|\varpi(\xi)v\|_{L^2}}{\|v\|_{L^2}} = \|\varpi(\xi)\|_{L^\infty} \leq \varpi_{\max}. \tag{68}$$

Since $\varpi(\cdot)$ is differentiable, it is straightforward to check that $A(\cdot)$ is also differentiable, and given $\xi, \eta \in L^2(\Omega)$, we have

$$[A'(\xi)\eta]v = \big([\varpi'(\xi)\eta]v, \cdot\big)_{L^2}, \quad \forall v \in L^2(\Omega).$$

Moreover, we can verify

$$\|A'(\xi)\| = \sup_{\eta\in L^2(\Omega)} \frac{\|\varpi'(\xi)\eta\|_{L^\infty}}{\|\eta\|_{L^2}} = \|\varpi'(\xi)\|_{\mathcal{L}(L^2(\Omega);L^\infty(\Omega))} \leq \varpi'_\infty. \tag{69}$$

Thus, the differentiability of $S_h(\cdot)$ is a consequence of Proposition 2.11(ii).

(iii) Uniform boundedness of $S'_h(\cdot)$ is a consequence of Proposition 2.11(iii), using the fact that $A(\cdot)$, $A'(\cdot)$, and $\alpha_h^{-1} \equiv \varpi_{\min}^{-1}$ are all uniformly bounded (see the above expressions (68) and (69)).
On the other hand, the Lipschitz-continuity of $S'_h(\cdot)$ relies on the Lipschitz-continuity of $A'(\cdot)$ (by Proposition 2.11(iv)). The latter is true since

$$\|A'(\xi_1) - A'(\xi_2)\| = \sup_{\eta\in L^2(\Omega)} \frac{\|\varpi'(\xi_1)\eta - \varpi'(\xi_2)\eta\|_{L^\infty}}{\|\eta\|_{L^2}} \leq \varpi_L\|\xi_1 - \xi_2\|_{L^2}.$$

A.7. *Proof of Proposition 3.10*

(i) This is a well-known result from Babuška–Brezzi theory (see, e.g., [6]).

(ii) Observe that $u_h \equiv S_h(\cdot)$ satisfies

$$\|u_h\|_{\mathbb{V}} \leq \frac{1}{\alpha_h(\xi)} \sup_{v_h\in\mathbb{U}_h} \frac{b(u_h, \omega(\xi)v_h)}{\|v_h\|_{\mathbb{V}}} = \frac{1}{\alpha_h(\xi)} \sup_{v_h\in\mathbb{U}_h} \frac{f(\omega(\xi)v_h)}{\|v_h\|_{\mathbb{V}}} \leq \frac{\omega_\infty}{\alpha}\|f\|_{\mathbb{V}^*}.$$

(iii) Let $\xi, \eta \in \mathbb{X}$, $t \in \mathbb{R}$, and notice that for all $v_h \in \mathbb{U}_h$ we have

$$b\Big(S_h(\xi + t\eta) - S_h(\xi), \omega(\xi + t\eta)v_h\Big) = f\Big(\big(\omega(\xi + t\eta) - \omega(\xi)\big)v_h\Big) - b\Big(S_h(\xi), \big(\omega(\xi + t\eta) - \omega(\xi)\big)v_h\Big).$$

Thus, the derivative of S_h at ξ in the η direction is the solution of

$$b\big(S'_h(\xi)\eta, \omega(\xi)v_h\big) = f\left(\omega'(\xi)\eta\, v_h\right) - b\left(S_h(\xi), \omega'(\xi)\eta\, v_h\right), \quad \forall v_h \in \mathbb{U}_h.$$

Moreover,

$$\|S'_h(\xi)\eta\|_{\mathbb{V}} \leq \frac{\|f\|_{\mathbb{V}^*} + \|B\|_{\mathcal{L}(\mathbb{V},\mathbb{V}^*)}\|S_h(\xi)\|_{\mathbb{V}}}{\alpha}\|\omega'(\xi)\|_{\mathcal{L}(\mathbb{X},\mathbb{W})}\|\eta\|_{L^2(\Omega)},$$

which implies that $S_h(\cdot)$ is differentiable and $S'_h(\cdot)$ is uniformly bounded, whenever $\omega(\cdot)$ is differentiable with uniformly bounded $\omega'(\cdot)$.

On the other hand, observe that for all $\xi_1, \xi_2, \eta \in \mathbb{X}$, and all $v_h \in \mathbb{U}_h$, we have

$$\begin{aligned} b\Big(S_h'(\xi_1)\eta - S_h'(\xi_2)\eta, \omega(\xi_1)v_h\Big) = \;& f\Big(\big(\omega'(\xi_1)-\omega'(\xi_2)\big)\eta v_h\Big) \\ &+ b\Big(S_h(\xi_2)-S_h(\xi_1), \omega'(\xi_2)\eta v_h\Big) \\ &+ b\Big(S_h(\xi_1), \big(\omega'(\xi_2)-\omega'(\xi_1)\big)\eta v_h\Big) \\ &+ b\Big(S_h'(\xi_2)\eta, \big(\omega(\xi_2)-\omega(\xi_1)\big)v_h\Big). \end{aligned}$$

Therefore,

$$\begin{aligned} \|S_h'(\xi_1)-S_h'(\xi_2)\|_{\mathcal{L}(\mathbb{X},\mathbb{V})} \le \;& \tfrac{1}{\alpha}\|f\|_{\mathbb{V}^*}\,\|\omega'(\xi_1)-\omega'(\xi_2)\|_{\mathcal{L}(\mathbb{X},\mathbb{W})} \\ &+ \tfrac{1}{\alpha}\|B\|_{\mathcal{L}(\mathbb{V},\mathbb{V}^*)}\,\|S_h(\xi_2)-S_h(\xi_1)\|_{\mathbb{V}}\,\|\omega'(\xi_2)\|_{\mathcal{L}(\mathbb{X},\mathbb{W})} \\ &+ \tfrac{1}{\alpha}\|B\|_{\mathcal{L}(\mathbb{V},\mathbb{V}^*)}\,\|S_h(\xi_1)\|_{\mathbb{V}}\,\|\omega'(\xi_2)-\omega'(\xi_1)\|_{\mathcal{L}(\mathbb{X},\mathbb{W})} \\ &+ \tfrac{1}{\alpha}\|B\|_{\mathcal{L}(\mathbb{V},\mathbb{V}^*)}\,\|S_h'(\xi_2)\|_{\mathcal{L}(\mathbb{X},\mathbb{V})}\,\|\omega(\xi_2)-\omega(\xi_1)\|_{\mathbb{W}}. \end{aligned}$$

Thus, the Lipschitz-continuity of S_h' relies on: the Lipschitz-continuity of ω'; the mean value theorem; and the uniform boundedness of S_h, S_h', and ω'.

A.8. *Proof of* Proposition 3.17

(i) Making the identification $\hat{\mathbb{V}} \equiv \mathbb{V}_h$ and $a(\xi;\cdot,\cdot) \equiv (\cdot,\cdot)_{\mathbb{V},\xi}$, we observe that the well-posedness of (42) is a direct consequence of Proposition 2.9, using the fact that $(\cdot,\cdot)_{\mathbb{V},\xi}$ is an equivalent innerproduct, together with assumption (41)(b).

(ii) Using the hypothesis of this statement and the estimate (16) in Proposition 2.9(iii), we get the uniform bound

$$\|S_h(\xi)\|_{\mathbb{U}} \le \frac{1}{\beta_h}\frac{\tilde{C}_2}{\tilde{C}_1}\|f\|_{\mathbb{V}^*}, \quad \forall \xi \in \mathbb{X}.$$

(iii) Direct application of Proposition 2.11, noticing also that $\alpha_h^{-1}(\xi) \le \tilde{C}_1^{-2}$ and

$$\begin{aligned} \|A(\xi)\|_{\mathcal{L}(\mathbb{V};\mathbb{V}^*)} &= \sup_{v_1\in\mathbb{V}} \frac{\|(v_1,\cdot)_{\mathbb{V},\xi}\|_{\mathbb{V}^*}}{\|v_1\|_{\mathbb{V}}} \\ &\le \sup_{v_1\in\mathbb{V}} \frac{\tilde{C}_2^2}{\|v_1\|_{\mathbb{V},\xi}} \left(\sup_{v_2\in\mathbb{V}} \frac{|(v_1,v_2)_{\mathbb{V},\xi}|}{\|v_2\|_{\mathbb{V},\xi}}\right) \\ &= \tilde{C}_2^2\,. \end{aligned}$$

References

[1] C.F. Higham, D.J. Higham, Deep learning: An introduction for applied mathematicians, SIAM Rev. 61 (4) (2019) 860–891.

[2] W. E, Machine learning and computational mathematics, Commun. Comput. Phys. 28 (5) (2020) 1639–1670.

[3] G.E. Karniadakis, I.G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, L. Yang, Physics-informed machine learning, Nat. Rev. Phys. 3 (2021) 422–440.

[4] G.C.Y. Peng, M. Alber, A.B. Tepole, W.R. Cannon, S. De, S. Dura-Bernal, K. Garikipati, G. Karniadakis, W.W. Lytton, P. Perdikaris, L. Petzold, E. Kuhl, Multiscale modeling meets machine learning: What can we learn? Arch. Comput. Methods Eng. 28 (3) (2021) 1017–1037.

[5] J.T. Oden, J.N. Reddy, An Introduction to the Mathematical Theory of Finite Elements, Dover, Mineola, New York, 2011, Unabridged republication of the edition published by John Wiley and Sons, New York, 1976.

[6] A. Ern, J.-L. Guermond, Finite Elements II. Galerkin Approximation, Elliptic and Mixed PDEs, in: Texts in Applied Mathematics, vol. 73, Springer Nature, Switzerland, 2021.

[7] M. Raissi, P. Perdikaris, G.E. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, J. Comput. Phys. 378 (2019) 686–707.

[8] J.L. Guermond, A finite element technique for solving first-order PDEs in L^p, SIAM J. Numer. Anal. 42 (2) (2004) 714–737.

[9] E. Burman, A. Ern, Stabilized Galerkin approximation of convection-diffusion-reaction equations: discrete maximum principle and convergence, Math. Comp. 74 (2005) 1637–1652.

[10] V. John, P. Knobloch, On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I – A review, Comput. Methods Appl. Mech. Engrg. 196 (17) (2007) 2197–2215.

[11] J.A. Evans, T.J. Hughes, G. Sangalli, Enforcement of constraints and maximum principles in the variational multiscale method, Comput. Methods Appl. Mech. Engrg. 199 (1) (2009) 61–76.

[12] L. Demkowicz, J. Gopalakrishnan, I. Muga, J. Zitelli, Wavenumber explicit analysis of a DPG method for the multidimensional Helmholtz equation, Comput. Methods Appl. Mech. Engrg. 213–216 (2012) 126–138.

[13] D. Peterseim, Eliminating the pollution effect in Helmholtz problems by local subscale correction, Math. Comp. 86 (2017) 1005–1036.

[14] I. Brevis, I. Muga, K.G. van der Zee, A machine-learning minimal-residual (ML-MRes) framework for goal-oriented finite element discretizations, Comput. Math. Appl. 95 (2021) 186–199, Recent Advances in Least-Squares and Discontinuous Petrov–Galerkin Finite Element Methods.

[15] L. Demkowicz, J. Gopalakrishnan, Discontinuous Petrov–Galerkin (DPG) method, in: E. Stein, R. de Borst, T.J.R. Hughes (Eds.), Encyclopedia of Computational Mechanics, Second Edition, Wiley, 2017, Part 2 Fundamentals.

[16] I. Muga, K.G. van der Zee, Discretization of linear problems in Banach spaces: Residual minimization, nonlinear Petrov–Galerkin, and monotone mixed methods, SIAM J. Numer. Anal. 58 (6) (2020) 3406–3426.

[17] J. Xu, Finite neuron method and convergence analysis, Commun. Comput. Phys. 28 (5) (2020) 1707–1745.

[18] J. Pousin, Least squares formulations for some elliptic second order problems, feedforward neural network solutions and convergence results, J. Comput. Math. Data Sci. 2 (2022) 100023.

[19] J. Müller, M. Zeinhofer, Error estimates for the deep Ritz method with boundary penalty, in: B. Dong, Q. Li, L. Wang, Z.-Q.J. Xu (Eds.), Proceedings of Mathematical and Scientific Machine Learning, in: Proceedings of Machine Learning Research, vol. 190, PMLR, 2022, pp. 215–230.

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

[21] S. Mishra, R. Molinaro, Estimates on the generalization error of physics-informed neural networks for approximating PDEs, IMA J. Numer. Anal. (2022) 1–43, http://dx.doi.org/10.1093/imanum/drab093, in press.

[22] Z. Cai, J. Chen, M. Liu, Least-squares ReLU neural network (LSNN) method for linear advection-reaction equation, J. Comput. Phys. 443 (2021) 110514.

[23] K. Kergrene, S. Prudhomme, L. Chamoin, M. Laforest, A new goal-oriented formulation of the finite element method, Comput. Methods Appl. Mech. Engrg. 327 (2017) 256–276, Advances in Computational Mechanics and Scientific Computation—the Cutting Edge.

[24] A.N. Brooks, T.J. Hughes, Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations, Comput. Methods Appl. Mech. Engrg. 32 (1) (1982) 199–259.

[25] J. Barrett, K. Morton, Approximate symmetrization and Petrov-Galerkin methods for diffusion-convection problems, Comput. Methods Appl. Mech. Engrg. 45 (1) (1984) 97–122.

[26] J.T. Oden, Optimal h-p finite element methods, Comput. Methods Appl. Mech. Engrg. 112 (1994) 309–331.

[27] D. Ray, J.S. Hesthaven, An artificial neural network as a troubled-cell indicator, J. Comput. Phys. 367 (2018) 166–191.

[28] S. Mishra, A machine learning framework for data driven acceleration of computations of differential equations, Math. Eng. 1 (1) (2018) 118–146.

[29] Y. Bar-Sinai, S. Hoyer, J. Hickey, M.P. Brenner, Learning data-driven discretizations for partial differential equations, Proc. Natl. Acad. Sci. 116 (31) (2019) 15344–15349.

[30] N. Discacciati, J.S. Hesthaven, D. Ray, Controlling oscillations in high-order discontinuous Galerkin schemes using artificial viscosity tuned by neural networks, J. Comput. Phys. 409 (2020) 109304.

[31] Y. Wang, Z. Shen, Z. Long, B. Dong, Learning to discretize: Solving 1D scalar conservation laws via deep reinforcement learning, Commun. Comput. Phys. 28 (5) (2020) 2158–2179.

[32] L. Schwander, D. Ray, J.S. Hesthaven, Controlling oscillations in spectral methods by local artificial viscosity governed by neural networks, J. Comput. Phys. 431 (2021) 110144.

[33] T. Tassi, A. Zingaro, L. Dede', A machine learning approach to enhance the SUPG stabilization method for advection-dominated differential problems, Math. Eng. 5 (2) (2023) 1–26.

[34] K.J. Fidkowski, G. Chen, Metric-based, goal-oriented mesh adaptation using machine learning, J. Comput. Phys. 426 (2021) 109957.

[35] J. Bohn, M. Feischl, Recurrent neural networks as optimal mesh refinement strategies, Comput. Math. Appl. 97 (2021) 61–76.

[36] W. E., B. Yu, The Deep Ritz Method: A deep learning-based numerical algorithm for solving variational problems, Commun. Math. Sci. 6 (1) (2018) 1–12.

[37] J. Sirignano, K. Spiliopoulos, DGM: A deep learning algorithm for solving partial differential equations, J. Comput. Phys. 375 (2018) 1339–1364.

[38] J. Berg, K. Nyström, A unified deep artificial neural network approach to partial differential equations in complex geometries, Neurocomputing 317 (2018) 28–41.

[39] M. Ainsworth, J. Dong, Galerkin neural networks: A framework for approximating variational equations with error control, SIAM J. Sci. Comput. 43 (4) (2021) A2474–A2501.

[40] M. Liu, Z. Cai, J. Chen, Adaptive two-layer ReLU neural network: I. Best least-squares approximation, Comput. Math. Appl. 113 (2022) 34–44.

[41] C. Uriarte, D. Pardo, Á.J. Omella, A finite element based deep learning solver for parametric PDEs, Comput. Methods Appl. Mech. Engrg. 391 (2022) 114562.

[42] J. Hesthaven, S. Ubbiali, Non-intrusive reduced order modeling of nonlinear problems using neural networks, J. Comput. Phys. 363 (2018) 55–78.

[43] B. Khara, A. Balu, A. Joshi, S. Sarkar, C. Hegde, A. Krishnamurthy, B. Ganapathysubramanian, NeuFENet: Neural finite element solutions with theoretical bounds for parametric PDEs, 2021, arXiv:2110.01601.

[44] G. Teichert, A. Natarajan, A. Van der Ven, K. Garikipati, Machine learning materials physics: Integrable deep neural networks enable scale bridging by learning free energy functions, Comput. Methods Appl. Mech. Engrg. 353 (2019) 201–216.

[45] J. Berg, K. Nyström, Neural networks as smooth priors for inverse problems for PDEs, J. Comput. Math. Data Sci. 1 (2021) 100008.

[46] K. Xu, E. Darve, Physics constrained learning for data-driven inverse modeling from sparse observations, J. Comput. Phys. 453 (2022) 110938.

[47] H. You, Y. Yu, N. Trask, M. Gulian, M. D'Elia, Data-driven learning of nonlocal physics from high-fidelity synthetic data, Comput. Methods Appl. Mech. Engrg. 374 (2021) 113553.

[48] K. Bhattacharya, B. Hosseini, N.B. Kovachki, A.M. Stuart, Model reduction and neural networks for parametric PDEs, SMAI J. Comput. Math. 7 (2021) 121–157.

[49] L. Cao, T. O'Leary-Roseberry, P.K. Jha, J.T. Oden, O. Ghattas, Residual-based error correction for neural operator accelerated infinite-dimensional Bayesian inverse problems, 2022.

[50] S. Mishra, R. Molinaro, Estimates on the generalization error of physics-informed neural networks for approximating a class of inverse problems for PDEs, IMA J. Numer. Anal. 42 (2022) 981–1022.

[51] S. Berrone, C. Canuto, M. Pintore, Variational physics informed neural networks: the role of quadratures and test functions, 2021, https://arxiv.org/abs/2109.02035.

[52] J. Roth, M. Schröder, T. Wick, Neural network guided adjoint computations in dual weighted residual error estimation, SN Appl. Sci. 4 (2022) 62.

[53] P. Minakowski, T. Richter, A priori and a posteriori error estimates for the Deep Ritz method applied to the Laplace and Stokes problem, J. Comput. Appl. Math. (2022) 114845.

[54] J.L. Lions, Optimal Control of Systems Governed By Partial Differential Equations, Springer-Verlag, Berlin, 1971.

[55] M. Hinze, R. Pinnau, M. Ulbrich, S. Ulbrich, Optimization with PDE Constraints, Springer, 2009.

[56] F. Tröltzsch, Optimal Control of Partial Differential Equations: Theory, Methods and Applications, in: Graduate Studies in Mathematics, vol. 112, American Mathematical Society, Providence, 2010.

[57] A. Borzì, V. Schulz, Computational Optimization of Systems Governed By Partial Differential Equations, in: Siam series on Computational Science and Engineering, Society for Industrial and Applied Mathematics, 2012.

[58] R. Rannacher, B. Vexler, A priori error estimates for the finite element discretization of elliptic parameter identification problems with pointwise measurements, SIAM J. Control Optim. 44 (5) (2005) 1844–1863.

[59] P. Petersen, M. Raslan, F. Voigtlaender, Topological properties of the set of functions generated by neural networks of fixed size, Found. Comput. Math. 21 (5) (2021) 375—444.

[60] P. Bochev, M. Gunzburger, Chapter 12 - least-squares methods for hyperbolic problems, in: R. Abgrall, C.-W. Shu (Eds.), Handbook of Numerical Methods for Hyperbolic Problems, in: Handbook of Numerical Analysis, vol. 17, Elsevier, 2016, pp. 289–317.

[61] D. Yarotsky, Error bounds for approximations with deep ReLU networks, Neural Netw. 94 (2017) 103–114.

[62] I. Gühring, G. Kutyniok, P. Petersen, Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms, Anal. Appl. 18 (05) (2020) 803–859.

[63] S. Berrone, C. Canuto, M. Pintore, Solving PDEs by variational physics-informed neural networks: an a posteriori error analysis, 2022, https://arxiv.org/abs/2205.00786.

[64] D.A. Di Pietro, A. Ern, Mathematical Aspects of Discontinuous Galerkin Methods, in: Mathématiques et Applications, vol. 69, Springer, Berlin, 2012.

[65] I. Muga, M.J.W. Tyler, K.G. van der Zee, The discrete-dual minimal-residual method (DDMRes) for weak advection-reaction problems in Banach spaces, Comput. Methods Appl. Math. 19 (3) (2019) 557–579.

[66] L.A. Vese, C.L. Guyader, Variational Methods in Image Processing, in: Mathematical and Computational Imaging Sciences, Chapman and Hall/CRC, Boca Raton, 2016.

[67] H. Brezis, Functional Analysis, Sobolev Spaces and Partial Differential Equations, in: Universitext, Springer, New York, 2011.

[68] P.G. Ciarlet, Linear and Nonlinear Functional Analysis with Applications, SIAM, Philadelphia, 2013.