ESAIM: M2AN 56 (2022) 1173–1198
https://doi.org/10.1051/m2an/2022035

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

AN ULTRAWEAK SPACE-TIME VARIATIONAL FORMULATION FOR THE WAVE EQUATION: ANALYSIS AND EFFICIENT NUMERICAL SOLUTION

Julian Henning[1], Davide Palitta[2], Valeria Simoncini[2] and Karsten Urban[1,*]

Abstract. We introduce an ultraweak space-time variational formulation for the wave equation, prove its well-posedness (even in the case of minimal regularity) and optimal inf-sup stability. Then, we introduce a tensor product-style space-time Petrov–Galerkin discretization with optimal discrete inf-sup stability, obtained by a non-standard definition of the trial space. As a consequence, the numerical approximation error is equal to the residual, which is particularly useful for *a posteriori* error estimation. For the arising discrete linear systems in space and time, we introduce efficient numerical solvers that appropriately exploit the equation structure, either at the preconditioning level or in the approximation phase by using a tailored Galerkin projection. This Galerkin method shows competitive behavior concerning wall-clock time, accuracy and memory as compared with a standard time-stepping method in particular in low regularity cases. Numerical experiments with a 3D (in space) wave equation illustrate our findings.

Mathematics Subject Classification. 35L15, 65M15, 65M60.

Received July 26, 2021. Accepted April 7, 2022.

1. Introduction

The wave equation has extensively been studied in theory and numerical approximations. The aim of this paper is to introduce a (non-standard) variational Hilbert space setting for the wave equation and a corresponding Petrov–Galerkin discretization that is well-posed and optimally stable in the sense that the inf-sup constant is unity. A major source of motivation for this view point is model reduction of parameterized partial differential equations by the reduced basis method [18, 20, 29]. In that framework, the numerical approximation error is equal to the residual, which is particularly useful for *a posteriori* error estimation and model reduction.

Space-time variational methods have been introduced, *e.g.*, for parabolic problems [1, 33, 34], transport-dominated problems [9, 11, 15], the wave equation [5, 6, 16, 32, 36] and the Schrödinger equation [13], also partly with the focus of optimal inf-sup stability. The potential for efficient numerical solvers has been shown in [19, 27].

We follow the path of [9, 11] and introduce an *ultraweak* variational formulation in space and time by applying all derivatives onto the test functions using integration by parts. This means that the trial space is $L_2(I \times \Omega)$, where $I = (0, T)$ is the time interval and $\Omega \subset \mathbb{R}^d$ the domain in space. This is the "correct" space of minimal

Keywords and phrases. Wave equation, ultraweak formulation, tensorproduct, numerical solvers.

[1] Ulm University, Institute for Numerical Mathematics, Helmholtzstr. 18, 89081 Ulm, Germany.

[2] Università di Bologna, Centro AM^2, Dipartimento di Matematica, Piazza di Porta S. Donato 5, 40127 Bologna, Italy.

[*]Corresponding author: karsten.urban@uni-ulm.de

regularity for initial data $u_0 \in L_2(\Omega)$. Following [11], we employ specifically chosen test spaces so as to derive a well-posed variational problem. A Petrov–Galerkin method is then used for the discretization: inspired by [9], we first choose an appropriate test space and then define the (non-standard) trial space to preserve optimal inf-sup stability. After completion of this work, we learned that this approach is very closely related to the DPG* method [13, 21].

The aforementioned discretization results into a linear system of equations $\mathbf{B}_\delta \boldsymbol{u}_\delta = \boldsymbol{g}_\delta$, whose (stiffness) matrix $\mathbf{B}_\delta$ is a sum of tensor products and has large condition number, making the system solution particularly challenging. Memory and computational complexity are also an issue, as space-time discretizations in general lead to larger systems as compared to conventional time-stepping schemes, where a sequence of linear systems has to be solved, whose dimension corresponds to the spatial discretization only.

Building upon [19], we introduce matrix-based solvers that are competitive with respect to time-stepping schemes. In particular, we show that in case of minimal regularity the space-time method using fast matrix-based solvers outperforms a Crank–Nicolson time-stepping scheme.

The remainder of this paper is organized as follows: In Section 2, we review known facts concerning variational formulations in general and for the wave equation in particular. We derive an optimally inf-sup stable ultraweak variational form. Section 3 is devoted to the Petrov–Galerkin discretization, again allowing for an inf-sup constant equal to 1. The arising linear system of equations is derived in Section 4 and its efficient and stable numerical solution is discussed in Section 5. We show some results of numerical experiments for the 3D wave equation in Section 6. For proving the well-posedness of the proposed variational form we need a result concerning a semi-variational formulation of the wave equation, whose proof is given in Appendix A.

2. Variational formulations of the wave equation

We are interested in a general linear equation of wave type. To this end, consider a Gelfand triple of Hilbert spaces $V \hookrightarrow H \hookrightarrow V'$ and a positive, symmetric operator $A \in \mathcal{L}(D(A), H)$, where $D(A)$ is the domain of A to be detailed in (2.5) below[1]. Setting $I := (0, T)$, $T > 0$ and given $g \in L_2(I; V')$[2], $u_0 \in H$, $u_1 \in V'$, we look for $u(t) \in V$, $t \in I$ *a.e.*, such that

$$\ddot{u}(t) + A\,u(t) = f(t) \text{ in } V',\ t \in I \ \textit{a.e.}, \qquad u(0) = u_0 \in H,\ \dot{u}(0) = u_1 \in V'. \tag{2.1}$$

Note that the initial state is only in H (*e.g.*, $L_2(\Omega)$) and the initial velocity only in V' (*e.g.*, $H^{-1}(\Omega)$), which means very low regularity. Thus, without additional regularity, we cannot expect to get a smooth solution of (2.1). Such non-smooth data are in fact a physically relevant situation. We restrict ourselves to linear time-invariant (LTI) systems even though most of our results can be extended to the more general situation of a time-dependent operator $A(t)$.

2.1. Inf-sup-theory

We are interested in finding a well-posed weak (or variational) formulation of (2.1), *i.e.*, Hilbert spaces $\mathbb{U}$, $\mathbb{V}$ of functions and a bilinear form $b : \mathbb{U} \times \mathbb{V} \to \mathbb{R}$ such that

$$b(u, v) = g(v) \quad \forall v \in \mathbb{V}, \tag{2.2}$$

has a unique solution $u \in \mathbb{U}$ for all given functionals $g \in \mathbb{V}'$ and that u solves (2.1) in some appropriate weak sense. The well-posedness of (2.2) is fully described by the following well-known fundamental statement.

Theorem 2.1 (Nečas theorem, *e.g.*, [26], Thm. 2)**.** *Let $\mathbb{U}$, $\mathbb{V}$ be Hilbert spaces and $b : \mathbb{U} \times \mathbb{V} \to \mathbb{R}$ be a bilinear form, which is bounded, i.e.,*

$$\exists\, \gamma < \infty : \quad b(u, v) \le \gamma \|u\|_{\mathbb{U}}\, \|v\|_{\mathbb{V}}, \quad \text{for all} \ \ u \in \mathbb{U}, v \in \mathbb{V} \quad \text{(boundedness)}. \tag{C.1}$$

[1] We shall always denote by V' the dual space of V w.r.t. the pivot space H.

[2] For a definition of Bochner spaces, see Section 2.4 below.

Then, for all $g \in \mathbb{V}'$, the variational problem (2.2) *admits a unique solution $u^* \in \mathbb{U}$, which depends continuously on the data $g \in \mathbb{V}'$ if and only if*

$$\beta := \inf_{u \in \mathbb{U}} \sup_{v \in \mathbb{V}} \frac{b(u,v)}{\|u\|_{\mathbb{U}}\, \|v\|_{\mathbb{V}}} > 0 \quad \text{(inf-sup-condition);} \tag{C.2}$$

$$\forall\, 0 \neq v \in \mathbb{V} \quad \exists\, u \in \mathbb{U}: \quad b(u,v) \neq 0 \quad \text{(surjectivity).} \tag{C.3}$$

The inf-sup constant β (or some lower bound) also plays a crucial role for the numerical approximation of the solution $u \in \mathbb{U}$ since it enters the relation between the approximation error and the residual (by the Xu–Zikatanov lemma [35], see also below). This motivates our interest in the size of β: the larger[3], the better.

A standard tool (at least) for (i) proving the inf-sup-stability in (C.2); (ii) stabilizing finite-dimensional discretizations; and (iii) getting sharp bounds for the inf-sup constant; is to determine the so-called *supremizer*. To define it, let $b : \mathbb{U} \times \mathbb{V} \to \mathbb{R}$ be a generic bounded bilinear form and $0 \neq u \in \mathbb{U}$ be given. Then, the *supremizer* $s_u \in \mathbb{V}$ is defined as the unique solution of

$$(s_u, v)_{\mathbb{V}} = b(u,v) \qquad \forall v \in \mathbb{V}. \tag{2.3}$$

It is easily seen that

$$\sup_{v \in \mathbb{V}} \frac{b(u,v)}{\|v\|_{\mathbb{V}}} = \sup_{v \in \mathbb{V}} \frac{(s_u, v)_{\mathbb{V}}}{\|v\|_{\mathbb{V}}} = \|s_u\|_{\mathbb{V}}, \tag{2.4}$$

which justifies the name *supremizer*.

2.2. The semi-variational framework

We start presenting some facts from the analysis of semi-variational formulations of the wave equation, where we follow and slightly extend ([3], Chap. 8). The term *semi-variational* originates from the use of classical differentiation w.r.t. time and a variational formulation in the space variable. As above, we suppose that two real Hilbert spaces V and H are given, such that V is compactly imbedded in H. Let $a : V \times V \to \mathbb{R}$ be a continuous, coercive and symmetric bilinear form[4]. Next, let A be the operator on H associated with $a(\cdot,\cdot)$ in the following sense: We define the *domain* of A by

$$D(A) := \{u \in V : \exists f \in H \text{ such that } \; a(u,v) = (f,v)_H \,\forall v \in V\}, \tag{2.5}$$

and recall that for any $u \in D(A)$ there is a unique $f \in H$ such that $a(u,v) = (f,v)_H$ for all $v \in V$. Then, we define $A : D(A) \to H$ by $u \mapsto f =: Au$. By the spectral theorem there exists an orthonormal basis $\{e_n : n \in \mathbb{N}\}$ of H (the eigenvectors of A) and numbers $\lambda_n \in \mathbb{R}$ with $0 < \lambda_1 \leq \lambda_2 \leq \cdots$, $\lim_{n\to\infty} \lambda_n = \infty$, such that

$$V = \left\{ v \in H : \sum_{n=1}^{\infty} \lambda_n |(v, e_n)_H|^2 < \infty \right\}, \tag{2.6a}$$

$$D(A) = \{v \in H : Av \in H\} = \left\{ v \in H : \|v\|_{D(A)}^2 := \sum_{n=1}^{\infty} \lambda_n^2 |(v, e_n)_H|^2 < \infty \right\}, \tag{2.6b}$$

$$a(u,v) = \sum_{n=1}^{\infty} \lambda_n (u, e_n)_H \,(e_n, v)_H, \qquad u, v \in V, \tag{2.6c}$$

$$Av = \sum_{n=1}^{\infty} \lambda_n (v, e_n)_H \, e_n, \qquad v \in D(A). \tag{2.6d}$$

[3] *I.e.*, the closer to unity, resp. to γ.

[4] Note that most of what is said can be also extended to H-elliptic forms (Gårding inequality).

TABLE 1. Regularity statements for the wave equation – classical in time, variational in space.

s	u_0	u_1	f	w	$\dot{w}$	$\ddot{w}$
$=$	$\in$		$\in C([0,T];\cdot)$	$\in C([0,T];\cdot)$		
0	H	V'	V'	H	V'	$D(A)'$
1	V	H	H	V	H	V'
2	$D(A)$	V	V	$D(A)$	V	H

Note that $D(A)$ is dense in H, $e_n \in D(A)$ and $Ae_n = \lambda_n e_n$ for all $n \in \mathbb{N}$. For $s \in \mathbb{R}$, we define

$$H^s := \left\{ v = \sum_{n=1}^{\infty} v_n \, e_n : \, \|v\|_s^2 := \sum_{n=1}^{\infty} \lambda_n^s \, v_n^2 < \infty \right\} \tag{2.7}$$

and note that $H^0 = H$, $H^1 = V$ and $H^2 = D(A)$. Moreover, $(H^s)' \cong H^{-s}$, see Proposition A.1. We consider the non-homogeneous wave equation

$$\ddot{w}(t) + A\, w(t) = f(t), \quad t \in (0,T), \qquad w(0) = u_0, \dot{w}(0) = u_1. \tag{2.8}$$

Then the following result on the existence and uniqueness holds. Its proof is given in Appendix A.

Theorem 2.2. *Let $s \in \mathbb{R}_{\geq 0}$, $u_0 \in H^s$, $u_1 \in H^{s-1}$ and $f \in C\big([0,T]; H^{s-1}\big)$. Then (2.8) admits a unique solution*

$$w \in \mathcal{C}^s := C^2\big([0,T]; H^{s-2}\big) \cap C^1\big([0,T]; H^{s-1}\big) \cap C([0,T], H^s). \tag{2.9}$$

We note a simple consequence for the backward wave equation.

Corollary 2.3. *Let $s \in \mathbb{R}_{\geq 0}$, $u_0 \in H^s$, $u_1 \in H^{s-1}$ and $g \in C\big([0,T]; H^{s-1}\big)$. Then*

$$\ddot{w}(t) + A\, w(t) = g(t), \quad t \in (0,T), \qquad w(T) = u_0, \dot{w}(T) = u_1. \tag{2.10}$$

Admits a unique solution $w \in \mathcal{C}^s$, see (2.9).

Proof. By the mapping $t \mapsto T - t$ we can transform (2.10) into (2.8) and deduce the well-posedness from Theorem 2.2. □

Theorem 2.2 ensures that $B := \frac{\mathrm{d}^2}{\mathrm{d}t^2} + A$ is an isomorphism of $\mathcal{C}_0^s := \{v \in \mathcal{C}^s : v(0) = \dot{v}(0) = 0\}$ onto $C\big([0,T]; H^{s-2}\big)$ for any $s \geq 0$. We detail the involved spaces in Table 1, which also shows that we have to expect at most $w(t) \in H$, $t \in I$, in the semi-variational setting given the low regularity of the initial conditions in (2.1). Hence, in a variational space-time setting, we can only hope for $w(t) \in H$ for *almost all* $t \in I$.

2.3. Biharmonic problem and mixed form

For later reference, let us consider the bilinear form $q : D(A) \times D(A) \to \mathbb{R}$ defined by $q(u,v) := (Au, Av)_H$, $u, v \in D(A)$, which is of biharmonic type. In order to detail the associated operator Q, recall that we have a Gelfand quintuple $D(A) \hookrightarrow V \hookrightarrow H \hookrightarrow V' \hookrightarrow D(A)'$. The duality pairing of $D(A)$ and $D(A)'$ is denoted by $\langle \cdot, \cdot \rangle_{D(A)' \times D(A)}$. Then, $Q : D(A) \to D(A)'$ defined as $\langle Qu, v \rangle_{D(A)' \times D(A)} = q(u,v)$ for $u, v \in D(A)$.

The adjoint operator $A' : H \to D(A)'$ is given by $\langle A'h, w \rangle_{D(A)' \times D(A)} = (h, Aw)_H$ for $w \in D(A)$ and $h \in H$. Then, $A'A : D(A) \to D(A)'$ and we get for $u, v \in D(A)$ that $\langle A'Au, v \rangle_{D(A)' \times D(A)} = (Au, Av)_H = q(u,v) = \langle Qu, v \rangle_{D(A)' \times D(A)}$, hence $Q = A'A$. Next, we consider the following operator problem:

$$\text{Given } g \in D(A)', \quad \text{determine } z \in D(A) \text{ such that } Qz = g. \tag{2.11}$$

Introducing the auxiliary variable $u := -Az \in H$, we can rewrite this problem as

$$\begin{pmatrix} I & A \\ A' & 0 \end{pmatrix} \begin{pmatrix} u \\ z \end{pmatrix} = \begin{pmatrix} 0 \\ -g \end{pmatrix}, \tag{2.12}$$

which is easily seen to be equivalent to (2.11).

2.4. Towards space-time variational formulations

The semi-variational formulation described above cannot be written as a variational formulation in the form of (2.2), since $C^k([0,T];X)$ is not a Hilbert space, even if X is a Hilbert space of functions $\phi : \Omega \to \mathbb{R}$ in space, *e.g.*, $L_2(\Omega)$ or $H_0^1(\Omega)$. We need Lebesgue-type spaces for the temporal and spatial variables yielding the notion of *Bochner spaces*, denoted by $\mathcal{X} := L_2(I;X)$[5] and defined as

$$\mathcal{X} := L_2(I;X) := \left\{ v : I \to X : \|v\|^2_{L_2(I;X)} := \int_0^T \|v(t)\|_X^2 \,\mathrm{d}t < \infty \right\},$$

which are Hilbert spaces with the inner product $(w,v)_{\mathcal{X}} := \int_0^T (w(t),v(t))_X \,\mathrm{d}t$, where $(\cdot,\cdot)_X$ denotes the respective inner product in X. We will often use the specific cases $(\cdot,\cdot)_{\mathcal{V}}$ and $(\cdot,\cdot)_{\mathcal{H}}$ for $\mathcal{V} := L_2(I;V)$ as well as $\mathcal{H} := L_2(I;H)$. Sobolev–Bochner spaces, *e.g.*, $H^1(I;X)$, $H^2(I;X)$ can be defined accordingly using weak derivatives w.r.t. the time variable ([23], III.4.3).

We will derive a space-time variational formulation in Bochner spaces, *i.e.*, we multiply the partial differential equation in (2.1) with test functions in space *and* time and also integrate w.r.t. both variables. Now, the question remains how to apply integration by parts. One could think of performing integration by parts once w.r.t. all variables. This would yield a variational form in the Bochner space $H^1(I;V)$. However, we were not able to prove well-posedness in that setting. Hence, we suggest an *ultraweak* variational form, where all derivatives are put onto the test space by means of integration by parts. We thus define the trial space as

$$\mathbb{U} := \mathcal{H} = L_2(I;H) \tag{2.13}$$

and search for an appropriate test space $\mathbb{V}$ to guarantee the well-posedness of (2.2). Assuming that $v(T) = \dot{v}(T) = 0$ for any $v \in \mathbb{V}$ and performing integration by parts twice for both time and space variables, we obtain

$$b(u,v) := (u, \ddot{v} + Av)_{\mathcal{H}}, \qquad g(v) := (f,v)_{\mathcal{H}} + \langle u_1, v(0)\rangle - (u_0, \dot{v}(0))_H, \tag{2.14}$$

for $v \in \mathbb{V}$, where the space $\mathbb{V}$ still needs to be defined in such a way that all the assumptions of Theorem 2.1 are satisfied. It turns out that this is not a straightforward task. The duality pairing $\langle \cdot,\cdot \rangle$ is defined in (A.1) in the appendix.

The Lions-Magenes theory

Variational space-time problems for the wave equation within the setting (2.14) have already been investigated in the book [23] by Lions and Magenes. We are going to review some facts from Chapter III, Section 9, pp. 283–299 of [23]. The point of departure is the following adjoint-type problem.

For a given $\varphi \in L_2(I;H) = \mathbb{U}$, find $v : I \times \Omega \to \mathbb{R}$ such that

$$\ddot{v} + A\,v = \varphi, \qquad v(T) = \dot{v}(T) = 0. \tag{2.15}$$

It has been shown that the following space[6]

$$\mathbb{W} := \text{space described by the solution } v \text{ of (2.15) as } \varphi \text{ describes } L_2(I;H) \tag{2.16}$$

plays an important role for the analysis. It is known that $\mathbb{W} \subset C([0,T];V) \cap C^1([0,T];H) \cap H^2(I;V')$ and that $\frac{\mathrm{d}^2}{\mathrm{d}t^2} + A$ is an isomorphism of $\mathbb{W}$ onto $\mathbb{U}$.

[5] Spaces of space-time functions are denoted by calligraphic letters, spaces of functions in space only by plain letters.

[6] The definition (2.16) is literally cited from [23].

Theorem 2.4 ([23], Chap. 3, Thms. 8.1, 9.1)**.** *Let $a : V \times V \to \mathbb{R}$ satisfy a Gårding inequality and let $f \in L_2(I;H)$, $u_0 \in V$, $u_1 \in H$ be given. Then,*

(a) *there is a unique $u^* \in H^1(I;H) \cap L_2(I;V)$ such that $\ddot{u}^* + Au^* = f$, $u^*(0) = u_1$, $\dot{u}^*(0) = u_1$. In addition $u^* \in H^2(I;V')$;*
(b) *for any $\ell \in \mathbb{W}'$ there is a unique $u^* \in \mathbb{U}$ such that $b(u^*, v) = \ell(v)$ for all $v \in \mathbb{W}$.* □

Notice that the first statement is proven by deriving energy-type estimates for the uniqueness and a Faedo-Galerkin approximation for the existence. Let us comment on the previous theorem. First, we note that $u_0 \in V$, $u_1 \in H$ are "too smooth" initial conditions, we aim at (only) $u_0 \in H$, $u_1 \in V'$, see (2.14). As a consequence:

(1) Statement (a) in Thm. 2.4 results in a "too smooth" solution. In fact, we are interested in an ultraweak solution $u \in L_2(I;H)$, (a) is "too" much.
(2) Even though the stated solution in (b) has the "right" regularity, it is not clear how to associate the functional g in (2.14) to the dual space $\mathbb{W}'$, *i.e.*, how to interpret the three terms of g in (2.14) in the space $\mathbb{W}'$.

These issues are partly fixed by the following statement.

Theorem 2.5 ([23], Chap. III, Thms. 9.3, 9.4)**.** *Let the bilinear form $a(\cdot,\cdot)$ be coercive, $f \in L_2(I;V')$, $u_0 \in H$, $u_1 \in V'$. Then, there exists a unique $u^* \in L_\infty(I;H)$ with $\dot{u}^* \in L_\infty(I;V')$ such that $b(u^*, v) = g(v)$ for all $v \in \mathbb{W}_0 := \mathbb{W} \cap L_2(I;W)$ with $b(\cdot,\cdot)$ and g defined as in (2.14). Moreover, $u^* \in C^0(\bar{I};H) \cap C^1(\bar{I};V')$.* □

Even though the latter result uses the "right" smoothness of the data and also includes existence *and* uniqueness, we are not fully satisfied with regard to our goal of a well-posed variational formulation of the wave equation in Hilbert spaces. In fact, the "trial space" $L_\infty(I;H) \cap W^1_\infty(I;V')$ is not a Hilbert space and it is at least not straightforward to see how we can base a Petrov–Galerkin approximation on such a trial space. Hence, we follow a different path.

2.5. An optimally inf-sup stable ultraweak variational form

We are going to derive a well-posed ultraweak variational formulation (2.2) of (2.1), where $\mathbb{U} = L_2(I;H)$ and $b(\cdot,\cdot)$, $g(\cdot)$ are defined by (2.14). To this end, we will follow the framework presented in [11]. This approach is also called the *method of transposition* and already goes back to [23], see also *e.g.*, [2,8,24] for the corresponding finite element error analysis. For the presentation we will need the semi-variational formulation described above.

Let us restrict ourselves to $A = -\Delta$ acting on a convex domain $\Omega \subset \mathbb{R}^d$ and supplemented by homogeneous Dirichlet boundary conditions. This means that $H = L_2(\Omega)$, $V = H^1_0(\Omega)$ and $D(A) = H^2(\Omega) \cap H^1_0(\Omega)$. However, we stress the fact that most of what is said here can be also extended to other elliptic operators. Then, the starting point is the operator equation in the classical form, *i.e.*,

$$B_\circ u = g, \quad \text{where } B_\circ = \frac{\mathrm{d}^2}{\mathrm{d}t^2} + A_\circ, \qquad \Omega_T := (0,T) \times \Omega,$$

i.e., $A_\circ = -\Delta$ is also to be understood in the classical sense. Next, denote the classical domain of $B_\circ$ by $\mathcal{D}(B_\circ)$, where initial and boundary conditions are also imposed in $\mathcal{D}(B_\circ)$, *i.e.*, $\mathcal{D}(B_\circ) := \big\{v \in C\big(\bar{\Omega}_T\big) : B_\circ v \in C(\Omega_T), v(0) = 0, v(t,\cdot)_{|\partial\Omega} = 0 \ \forall t \in [0,T]\big\}$. Hence,

$$\mathcal{D}(B_\circ) = C^2(\Omega_T) \cap C^1_{\{0\}}\big([0,T]; C_0\big(\overline{\Omega}\big)\big) = \Big[C^2(I) \cap C^1_{\{0\}}([0,1])\Big] \otimes \big[C^2(\Omega) \cap C_0\big(\overline{\Omega}\big)\big],$$

where $C_0\big(\overline{\Omega}\big) := \big\{\phi \in C\big(\overline{\Omega}\big) : \phi_{|\partial\Omega} = 0\big\}$ models the homogeneous Dirichlet conditions, and for $t \in [0,T]$ and any function space X we define

$$C^1_{\{t\}}([0,T];X) := \big\{u \in C^1([0,T];X) : u(t) = \dot{u}(t) = 0 \ \text{ in } X\big\}.$$

The range $\mathcal{R}(B_\circ)$ in the classical sense then reads $\mathcal{R}(B_\circ) = C\big(\overline{\Omega}_T\big)$. As a next step, we determine the formal adjoint $B_\circ^*$ of $B_\circ$. Since

$$(B_\circ u, v)_{\mathcal{H}} = (u, B_\circ v)_{\mathcal{H}} \quad \text{for all } u, v \in C_0^\infty(\Omega_T),$$

the operator $B_\circ^*$ coincides with $B_\circ$ while acting on the space of functions with homogeneous terminal conditions $u(T) = \dot{u}(T) = 0$ instead of initial conditions. This means that $\mathcal{R}(B_\circ^*) = C\big(\overline{\Omega}_T\big)$ and

$$\mathcal{D}(B_\circ^*) = C^2(\Omega_T) \cap C^1_{\{T\}}\big([0,T]; C_0\big(\overline{\Omega}\big)\big) = \Big[C^2(I) \cap C^1_{\{T\}}([0,1])\Big] \otimes \big[C^2(\Omega) \cap C_0\big(\overline{\Omega}\big)\big].$$

Following [11], we need to verify the following conditions

(B^*1) $B_\circ^*$ is injective on the dense subspace $\mathcal{D}(B_\circ^*) \subset L_2(I;H)$ and
(B^*2) $\mathcal{R}(B_\circ^*) \hookrightarrow L_2(I;H)$ is densely imbedded.

Since $C\big(\overline{\Omega}_T\big) \cong C\big([0,T]; C\big(\overline{\Omega}\big)\big) \hookrightarrow L_2(I;H)$ is dense, (B^*2) is immediate. In order to prove (B^*1), first note that

$$\mathcal{D}(B_\circ^*) \subset \mathcal{C}_T^2 := \mathcal{C}^2 \cap C^1_{\{T\}}([0,T]; V). \tag{2.17}$$

Let us denote the continuous extension of $B_\circ^*$ from $\mathcal{D}(B_\circ^*)$ to $\mathcal{C}_T^2$ also by $B_\circ^*$. Corollary 2.3 implies that this continuous extension $B_\circ^*$ is an isomorphism from $\mathcal{C}_T^2$ onto $C([0,T];V)$ (here we need the semi-variational theory). This implies that $B_\circ^*$ is injective on $\mathcal{D}(B_\circ^*)$, *i.e.*, (B^*1). Now, the properties (B^*1) and (B^*2) ensure that

$$\|v\|_{\mathbb{V}} := \|B_\circ^* v\|_{\mathcal{H}} \tag{2.18}$$

is a norm on $D(B_\circ^*) = \mathcal{C}_T^2$. Then, we set

$$\mathbb{V} := \operatorname{clos}_{\|\cdot\|_{\mathbb{V}}}(\mathcal{C}_T^2) \subset L_2(I;H), \quad (v,w)_{\mathbb{V}} := (B^*v, B^*w)_{\mathcal{H}},\ v,w \in \mathbb{V}, \tag{2.19}$$

which is a Hilbert space, where B^* is to be understood as the continuous extension of $B_\circ^*$ from $\mathcal{C}_T^2$ to $\mathbb{V}$[7]. Now, we are ready to prove our first main result.

Theorem 2.6. *Let $f \in L_2(I;V')$, $u_0 \in H$ and $u_1 \in V'$. Moreover, let $\mathbb{V}$, $b(\cdot,\cdot)$, and $g(\cdot)$ be defined as in* (2.19) *and* (2.14)*, respectively. Then, the variational problem*

$$b(u,v) = g(v) \quad \text{for all } v \in \mathbb{V}, \tag{2.20}$$

admits a unique solution $u^* \in \mathbb{U}$. In particular,

$$\beta := \inf_{u\in\mathbb{U}} \sup_{v\in\mathbb{V}} \frac{b(u,v)}{\|u\|_{\mathbb{U}}\,\|v\|_{\mathbb{V}}} = \sup_{u\in\mathbb{U}} \sup_{v\in\mathbb{V}} \frac{b(u,v)}{\|u\|_{\mathbb{U}}\,\|v\|_{\mathbb{V}}} = 1. \tag{2.21}$$

Proof. We are going to show the conditions (C.1)–(C.3) of Theorem 2.1 above.

(C.1) *Boundedness*: let $u \in \mathbb{U}$, $v \in \mathbb{V}$, then by Cauchy–Schwarz' inequality

$$b(u,v) = (u, \ddot{v} + Av)_{\mathcal{H}} \le \|u\|_{\mathcal{H}}\,\|\ddot{v} + Av\|_{\mathcal{H}} = \|u\|_{\mathbb{U}}\,\|v\|_{\mathbb{V}},$$

i.e., the continuity constant is unity.

[7] For an approach alternative to the completion in (2.19), we refer to [13].

(C.2) *Inf-sup*: let $0 \neq u \in \mathbb{U}$ be given. We consider the supremizer $s_u \in \mathbb{V}$ defined as $(s_u, v)_{\mathbb{V}} = b(u, v) = (u, \ddot{v} + Av)_{\mathcal{H}}$ for all $v \in \mathbb{V}$. Since by definition of the inner product $(s_u, v)_{\mathbb{V}} = (\ddot{s}_u + As_u, \ddot{v} + Av)_{\mathcal{H}}$ for all $v \in \mathbb{V}$ we get $\ddot{s}_u + As_u = u$ in $\mathcal{H}$. Then, by (2.4),

$$\sup_{v \in \mathbb{V}} \frac{b(u,v)}{\|v\|_{\mathbb{V}}} = \sup_{v \in \mathbb{V}} \frac{(s_u, v)_{\mathbb{V}}}{\|v\|_{\mathbb{V}}} = \|s_u\|_{\mathbb{V}} = \|\ddot{s}_u + As_u\|_{\mathcal{H}} = \|u\|_{\mathcal{H}},$$

i.e., $\beta = 1$ for the inf-sup constant.

(C.3) *Surjectivity*: let $0 \neq v \in \mathbb{V}$ be given. Then, there is a sequence $(v_n)_{n\in\mathbb{N}} \subset \mathcal{C}_T^2$ with $v_n \neq 0$, converging towards v in $\mathbb{V}$. Since $B_\circ^*$ is an isometric isomorphism of $\mathcal{C}_T^2$ onto $C([0,T];V)$, there is a unique $u_n := B_\circ^* v_n = \ddot{v}_n + Av_n \in C([0,T];V)$. Hence $0 \neq \|v_n\|_{\mathcal{C}^2} = \|u_n\|_{C([0,T];V)}$. By possibly taking a subsequence, $(u_n)_{n\in\mathbb{N}}$ converges to a unique limit $u_v \in L_2(I;H)$. We take the limit as $n \to \infty$ on both sides of $u_n = \ddot{v}_n + Av_n$ and obtain $0 \neq u_v = B^* v = \ddot{v} + Av \in L_2(I;H) = \mathbb{U}$. Finally, $b(u_v, v) = (u_v, B^* v)_{\mathcal{H}} = (u_v, u_v)_{\mathcal{H}} = \|u_v\|_{\mathbb{U}}^2 > 0$, which proves surjectivity and concludes the proof.

□

Remark 2.7. The essence of the above proof is the fact that $\mathbb{U}$ and $\mathbb{V}$ are related as $\mathbb{U} = B^*(\mathbb{V})$ and noting that B^* coincides with B, while B acts on functions with homogeneous *initial* conditions whereas B^* on functions with homogeneous *terminal* conditions.

3. Petrov–Galerkin discretization

We determine a numerical approximation to the solution of a variational problem of the general form (2.2). To this end, one chooses finite-dimensional trial and test spaces, $\mathbb{U}_\delta \subset \mathbb{U}$, $\mathbb{V}_\delta \subset \mathbb{V}$, respectively, where δ is a discretization parameter to be explained later. For convenience, we assume that their dimension is equal, *i.e.*, $\mathcal{N}_\delta := \dim \mathbb{U}_\delta = \dim \mathbb{V}_\delta$. The Petrov–Galerkin method then reads

$$\text{find } u_\delta \in \mathbb{U}_\delta: \quad b(u_\delta, v_\delta) = g(v_\delta) \quad \text{for all } v_\delta \in \mathbb{V}_\delta. \tag{3.1}$$

As opposed to the coercive case, the well-posedness of (3.1) is not inherited from that of (2.20). In fact, in order to ensure uniform stability (*i.e.*, stability independent of the discretization parameter δ), the spaces $\mathbb{U}_\delta$ and $\mathbb{V}_\delta$ need to be appropriately chosen in the sense that the discrete inf-sup (or LBB – Ladyshenskaja–Babuška–Brezzi) condition holds, *i.e.*, there exists a $\beta_\circ > 0$ such that

$$\beta_\delta := \inf_{u_\delta \in \mathbb{U}_\delta} \sup_{v_\delta \in \mathbb{V}_\delta} \frac{b(u_\delta, v_\delta)}{\|u_\delta\|_{\mathbb{U}} \, \|v_\delta\|_{\mathbb{V}}} \geq \beta_\circ > 0, \tag{3.2}$$

where the crucial point is that $\beta_\circ$ is independent of δ. The size of $\beta_\circ$ is also relevant for the error analysis, since the Xu–Zikatanov lemma [35] yields a best approximation result

$$\|u^* - u_\delta^*\|_{\mathbb{U}} \leq \frac{1}{\beta_\circ} \inf_{w_\delta \in \mathbb{U}_\delta} \|u^* - w_\delta\|_{\mathbb{U}} \tag{3.3}$$

for the "exact" solution u^* of (2.20) and the "discrete" solution u_δ^* of (3.1). This is also the key for an optimal error/residual relation, which is important for *a posteriori* error analysis (also within the reduced basis method).

The key idea, as already stated earlier, is to *first* choose sufficiently smooth test functions, namely H^2 in space and time. This can be done, *e.g.*, by choosing at least quadratic splines. Then, the trial functions are the image of the test functions under the adjoint wave operator.

3.1. A stable Petrov–Galerkin space-time discretization

In order to use a straightforward finite element discretization, we use tensor product subspaces $\mathbb{U}_\delta \subset \mathbb{U}$ and $\mathbb{V}_\delta \subset \mathbb{V}$ with a possibly large inf-sup lower bound $\beta_\circ$ in (3.2). Constructing such a stable pair of trial and test spaces is again a nontrivial task, not only for the wave equation. It is a common approach to choose some trial approximation space $\mathbb{U}_\delta$ (*e.g.*, by splines) and then (try to) construct an appropriate according test space $\mathbb{V}_\delta$ in such a way that (3.2) is satisfied. This can be done, *e.g.*, by computing the supremizers for all basis functions in $\mathbb{U}_\delta$ and then define $\mathbb{V}_\delta$ as the linear span of these supremizers. However, this would amount to solve the original problem $\mathcal{N}_\delta$ times, which is way too costly. We mention that this approach indeed works within the discontinuous Galerkin (dG) method, see, *e.g.*, [10, 12, 16]. We will follow a different path, also used in [9] for transport problems. We first construct a test space $\mathbb{V}_\delta$ by a standard approach and then define a stable trial space $\mathbb{U}_\delta$ in a second step. This implies that the trial functions are no longer "simple" splines but they arise from the application of the adjoint operator B^* (which is here the same as the primal one B except for initial/terminal conditions) to the test basis functions.

Finite elements in time.

We start with the temporal discretization. We choose some integer $N_t > 1$ and set $\Delta t := T/N_t$. This results in a temporal "triangulation"

$$\mathcal{T}_{\Delta t}^{\text{time}}:=\{t^{k-1}:=(k-1)\Delta t < t \le k\,\Delta t=:t^k, 1 \le k \le N_t\}$$

in time. Then, we set

$$R_{\Delta t} := \text{span}\{\varrho^1, \dots, \varrho^{N_t}\} \subset H^2_{\{T\}}(I), \tag{3.4}$$

e.g., piecewise quadratic splines on $\mathcal{T}_{\Delta t}^{\text{time}}$ with standard modification in terms of multiple knots at the right end point of $\bar{I} = [0, T]$.

Example 3.1. Denote by S^k the quadratic B-spline corresponding to the nodes t^{k-2}, t^{k-1}, t^k and t^{k+1}, where we extend the node sequence outside $\bar{I}$ in an obvious manner. Then, $\varrho^k := S^{k-1}$, $k = 3, \dots, N_t$ are $H^2_0(I)$-functions which are fully supported in I. The remaining two basis functions on the left end point of the interval I, *i.e.*, ϱ^1, ϱ^2, can be formed by using $t^0 = 0$ as double and triple node, respectively. Thus, we get a discretization in $H^2_{\{T\}}(I)$ of dimension N_t. We show an example for $T = 1$ and $\Delta t = \frac{1}{4}$ (*i.e.*, $N_t = 4$) in Figure 1, the test functions in the center, optimal trial functions on the right.

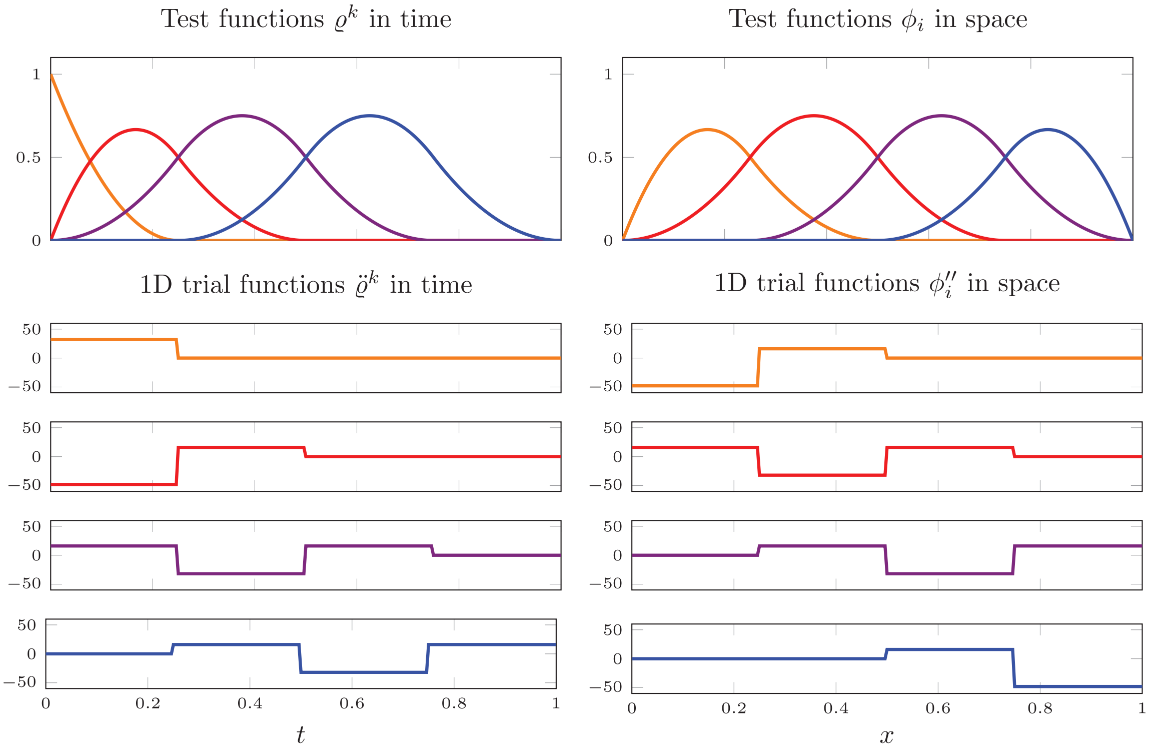
Discretization in space

For the space discretization, we choose any conformal finite element space

$$Z_h := \text{span}\{\phi_1, \dots, \phi_{N_h}\} \subset H^1_0(\Omega) \cap H^2(\Omega), \tag{3.5}$$

e.g., piecewise quadratic finite elements with homogeneous Dirichlet boundary conditions.

Example 3.2. As an example for the space discretization, let us detail the univariate (1D) case $\Omega = (0, 1)$. Define $x_j := j\,h$, $j = 0, \dots, N_h := \frac{1}{h}$, and denote by S^j the quadratic B-spline corresponding to the nodes x_{j-2}, x_{j-1}, x_j, and x_{j+1}. The B-splines $\phi_j := S_{j+1}$, $j = 2, \dots, N_h - 1$, are supported in Ω. We define the two boundary functions ϕ_1 and ϕ_{N_h} as the quadratic B-spline w.r.t. the nodes $(0, 0, x_1, x_2)$ and $(x_{N_h-2}, x_{N_h-1}, 1, 1)$ (*i.e.*, with double nodes), respectively, such that the homogeneous boundary conditions are satisfied. We obtain a discretization of dimension N_h. We show an example for $\Omega = (0, 1)$ and $h = \frac{1}{4}$ (*i.e.*, $N_h = 4$) in Figure 1, the test functions on the left. The arising trial functions are depicted on the right and turn out to be identical to the time discretization trial functions in Example 3.1.


Test functions ϱ^k in time

Test functions ϕ_i in space

1D trial functions $\ddot{\varrho}^k$ in time

1D trial functions ϕ_i'' in space

FIGURE 1. Discretization with quadratic test functions in time and 1D-space for $h = \Delta t = \frac{1}{4}$ with the corresponding trial functions in the *second* to *fifth row*. *First row*: functions on the left span $\mathbb{V}^{\mathrm{IVP}}_{\Delta t}$ in (3.9b), the second column corresponds to $\mathbb{V}^{\mathrm{BVP}}_h$ in (3.9a). *Rows* 2–5: functions in $\mathbb{U}^{\mathrm{IVP}}_{\Delta t}$ (*left*) and $\mathbb{U}^{\mathrm{BVP}}_h$ (*right*), see Section 3.2.

Test and trial space in space and time

Then, we define the test space as

$$\mathbb{V}_\delta := R_{\Delta t} \otimes Z_h = \operatorname{span}\big\{\varphi_\nu := \varrho^k \otimes \phi_i : \, k = 1, \ldots, N_t, \, i = 1, \ldots, N_h, \nu = (k, i)\big\}, \tag{3.6}$$

which is a tensor product space of dimension $\mathcal{N}_\delta = N_t \, N_h$, $\delta = (\Delta t, h)$, satisfying $\mathbb{V}_\delta \subset \mathbb{V}$.

The trial space $\mathbb{U}_\delta$ is constructed by applying the adjoint operator B^* to each test basis function, *i.e.*, for $\mu = (\ell, j)$ and $A = -\Delta$

$$\psi_\mu := B^*(\varphi_\mu) = B^*\big(\varrho^\ell \otimes \phi_j\big) = (\partial_{tt} + A)\big(\varrho^\ell \otimes \phi_j\big) = \ddot{\varrho}^\ell \otimes \phi_j + \varrho^\ell \otimes A\phi_j,$$

i.e.,

$$\mathbb{U}_\delta := B^*(\mathbb{V}_\delta) = \operatorname{span}\{\psi_\mu : \, \mu = 1, \ldots, \mathcal{N}_\delta\}.$$

Since B^* is an isomorphism of $\mathbb{V}$ onto $L_2(I; H)$, the functions ψ_μ are in fact linearly independent. An example of a single trial function is shown in Figure 2.

Proposition 3.3. *For the space* $\mathbb{V}_\delta$ *defined in* (3.6) *and* $\mathbb{U}_\delta := B^*(\mathbb{V}_\delta)$, *we have*

$$\beta_\delta := \inf_{u_\delta \in \mathbb{U}_\delta} \sup_{v_\delta \in \mathbb{V}_\delta} \frac{b(u_\delta, v_\delta)}{\|u_\delta\|_{\mathbb{U}} \, \|v_\delta\|_{\mathbb{V}}} = 1.$$

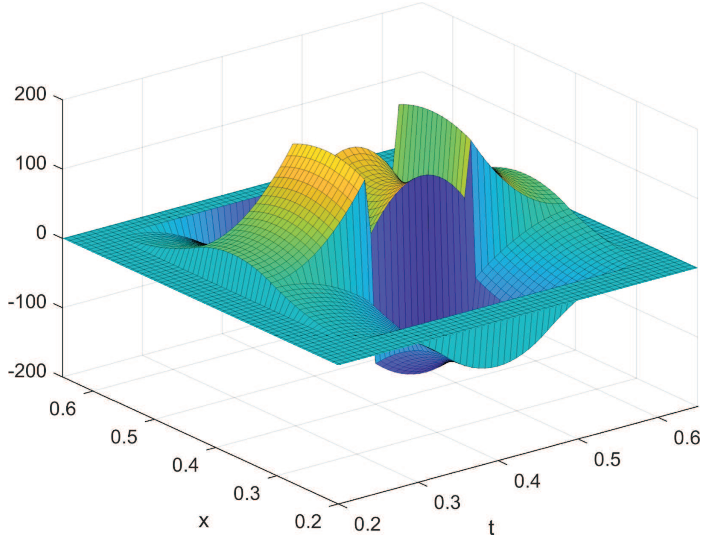


FIGURE 2. Sample trial function for $I = \Omega = (0,1)$, $\Delta t = h = \frac{1}{8}$.

Proof. Let $0 \neq u_\delta \in \mathbb{U}_\delta \subset L_2(I;H)$. Then, since $\mathbb{U}_\delta = B^*(\mathbb{V}_\delta)$ there exists a unique $z_\delta \in \mathbb{V}_\delta$ such that $B^* z_\delta = u_\delta$. Hence

$$\sup_{v_\delta \in \mathbb{V}_\delta} \frac{b(u_\delta, v_\delta)}{\|u_\delta\|_{\mathbb{U}}\, \|v_\delta\|_{\mathbb{V}}} \geq \frac{b(u_\delta, z_\delta)}{\|u_\delta\|_{\mathbb{U}}\, \|z_\delta\|_{\mathbb{V}}} = \frac{(u_\delta, B^* z_\delta)_{\mathcal{H}}}{\|u_\delta\|_{\mathbb{U}}\, \|z_\delta\|_{\mathbb{V}}} = \frac{(u_\delta, u_\delta)_{\mathcal{H}}}{\|u_\delta\|_{\mathcal{H}}\, \|B^* z_\delta\|_{\mathcal{H}}} \frac{\|u_\delta\|_{\mathcal{H}}^2}{\|u_\delta\|_{\mathcal{H}}\, \|u_\delta\|_{\mathcal{H}}} = 1.$$

On the other hand, by the Cauchy–Schwarz inequality, we have

$$\sup_{v_\delta \in \mathbb{V}_\delta} \frac{b(u_\delta, v_\delta)}{\|u_\delta\|_{\mathbb{U}}\, \|v_\delta\|_{\mathbb{V}}} = \sup_{v_\delta \in \mathbb{V}_\delta} \frac{(u_\delta, B^* v_\delta)_{\mathcal{H}}}{\|u_\delta\|_{\mathbb{U}}\, \|v_\delta\|_{\mathbb{V}}} \leq \sup_{v_\delta \in \mathbb{V}_\delta} \frac{\|u_\delta\|_{\mathcal{H}}\, \|B^* v_\delta\|_{\mathcal{H}}}{\|u_\delta\|_{\mathcal{H}}\, \|B^* v_\delta\|_{\mathcal{H}}} = 1,$$

which proves the claim. □

3.2. Optimal ultraweak discretization of ordinary differential equations

For the understanding of our subsequent numerical investigations, it is worth considering the univariate case, *i.e.*, ordinary differential equations (ODEs) of the form

$$-u''(x) = f(x), \qquad x \in (0,1), \tag{3.7}$$

with either boundary or second order initial conditions, namely

$$u(0) = u(1) = 0 \qquad \text{for a space-like problem, or} \tag{3.8a}$$

$$u(0) = 0, \quad u'(0) = 0 \qquad \text{for a time-like problem.} \tag{3.8b}$$

Using the framework above, we obtain $b(u,v) := -(u, v'')_{L_2(0,1)}$ and $\mathbb{U} := L_2(0,1)$ in both cases. Moreover, in this univariate setting, we can identify the test space $\mathbb{V}$ given in (2.19) as follows

$$\mathbb{V}^{\mathrm{BVP}} := H_0^1(0,1) \cap H^2(0,1), \quad \text{for (3.8a)}, \tag{3.9a}$$

$$\mathbb{V}^{\mathrm{IVP}} := H^2_{\{T\}}(0,1) \qquad \text{for (3.8b)}, \tag{3.9b}$$

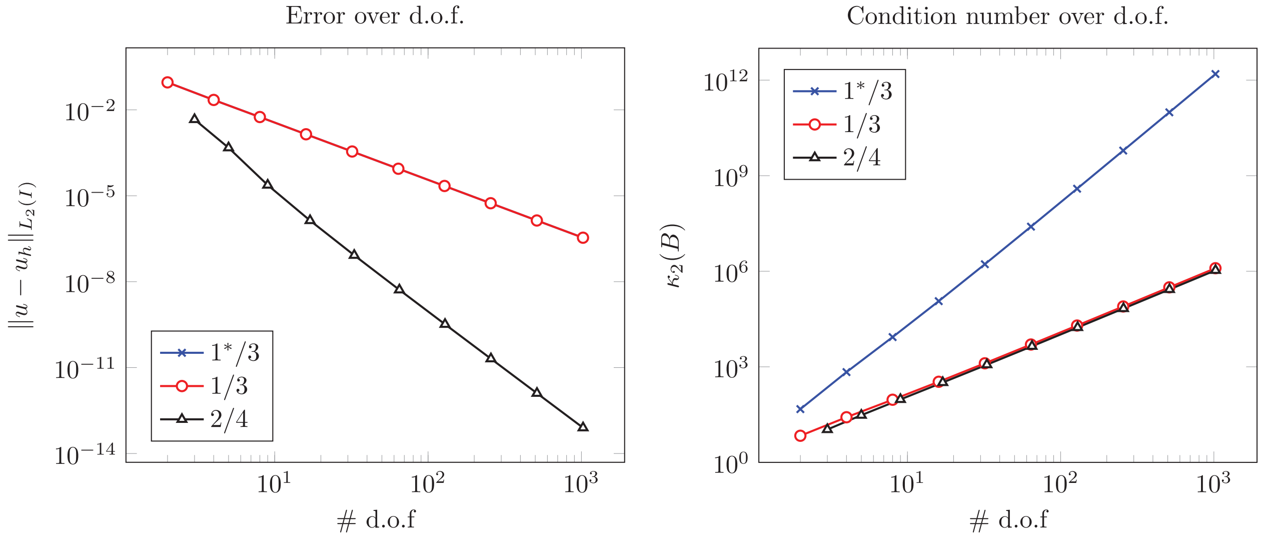


FIGURE 3. Initial value problem (3.8b). B-spline discretization of order $r_{\text{ansatz}}/r_{\text{test}}$, where * means that $U_{\Delta t} = B^*(V_{\Delta t})$.

where $H^2_{\{T\}}(0,1) := \left\{\vartheta \in H^2(I)\colon \vartheta(T) = \dot{\vartheta}(T) = 0\right\}$, which makes the discretization particularly straightforward. In fact, we use B-spline bases of different orders $r \geq 1$ (*i.e.*, polynomial degree $r-1$) to obtain discrete test spaces $\mathbb{V}_h^{\mathrm{BVP}} := \{\phi_i : i = 1, \ldots, N_h\}$ and $\mathbb{V}_{\Delta t}^{\mathrm{IVP}} = \left\{\varrho^k : k = 1, \ldots, N_t\right\}$, respectively. The boundary conditions for (3.8) can be realized by multiple knots and then omitting those B-splines at the boundaries which do not satisfy the particular boundary condition, see again Figure 1. The trial spaces are then formed by the second derivatives of the test functions, *i.e.*, $\mathbb{U}_h^{\mathrm{BVP}} := \{\phi_i'' : i = 1, \ldots, N_h\}$ and $\mathbb{U}_{\Delta t}^{\mathrm{IVP}} = \left\{\ddot{\varrho}^k : k = 1, \ldots, N_t\right\}$. Those functions are also depicted in Figure 1.

We did experiments for a whole variety of problems admitting solutions of different smoothness both for the boundary value ((3.7), (3.8a)) and the initial value problem ((3.7), (3.8b)). The differences were negligible, so that we only report the results for the initial value problem (3.8b). We investigate the L_2-error and the condition number of the stiffness matrix over the degrees of freedom (d.o.f.) for different type of discretizations, namely

- $1^*/3$: quadratic spline test functions and inf-sup-optimal trial functions as image of the test functions under the adjoint operator;
- $r_{\text{ansatz}}/r_{\text{test}}$: splines of order r_{ansatz} for the trial and of order r_{test} for the test functions (here $1/3$ and $2/4$). We obtain "standard" spline spaces for the trial space, not related to the test space through the image of the adjoint operator – and thus not necessarily inf-sup optimal.

The results are shown in Figure 3. The errors for $1^*/3$ and $1/3$ are the same, so that the blue line is not visible (in fact, the spanned spaces coincide with different bases). We obtain linear convergence for these cases and quadratic order for $2/4$. Concerning the condition numbers, we see that the inf-sup optimal case in fact gives rise to significantly larger condition numbers than the "standard" ones.

It is worth mentioning that we got $\beta_\delta \equiv 1$ in *all* cases. This means in particular that the ansatz spaces generated by the inf-sup-optimal setting $1^*/3$ are identical with those for the $1/3$ case. After observing this numerically, we have also proven this observation. However, we stress the fact that this is a pure univariate fact, *i.e.*, for the ODE. It is no longer true in the PDE case as we shall also see below.

4. Derivation and properties of the algebraic linear system

4.1. The linear system

To derive the stiffness matrix, we first use arbitrary spaces induced by $\{\psi_\mu := \sigma^\ell \otimes \xi_j : \mu = 1, \dots, \mathcal{N}_\delta\}$ for the trial and $\{\varphi_\nu = \varrho^k \otimes \phi_i : \nu = 1, \dots, \mathcal{N}_\delta\}$ for the test space an call this the "general" case. Using $[\mathbb{B}_\delta]_{\mu,\nu} = [\mathbb{B}_\delta]_{(\ell,j),(k,i)}$ we get

$$\begin{aligned}
[\mathbb{B}_\delta]_{(\ell,j),(k,i)} &= b(\psi_\mu, \varphi_\nu) = (\psi_\mu, B^*\varphi_\nu)_{\mathcal{H}} = \big(\sigma^\ell \otimes \xi_j, \ddot{\varrho}^k \otimes \phi_i + \varrho^k \otimes A\phi_i\big)_{\mathcal{H}} \\
&= \big(\sigma^\ell, \ddot{\varrho}^k\big)_{L_2(I)} (\xi_j, \phi_i)_{L_2(\Omega)} + \big(\sigma^\ell, \varrho^k\big)_{L_2(I)} (\xi_j, A\phi_i)_{L_2(\Omega)},
\end{aligned}$$

so that

$$\mathbb{B}_\delta = \tilde{\boldsymbol{N}}_{\Delta t} \otimes \tilde{\boldsymbol{M}}_h + \tilde{\boldsymbol{M}}_{\Delta t} \otimes \tilde{\boldsymbol{N}}_h \text{in this “general” case,} \tag{4.1}$$

where $\big[\tilde{\boldsymbol{M}}_{\Delta t}\big]_{\ell,k} := \big(\sigma^\ell, \varrho^k\big)_{L_2(I)}$, $\big[\tilde{\boldsymbol{M}}_h\big]_{j,i} := (\xi_j, \phi_i)_{L_2(\Omega)}$, $\big[\tilde{\boldsymbol{N}}_{\Delta t}\big]_{\ell,k} := \big(\sigma^\ell, \ddot{\varrho}^k\big)_{L_2(I)}$ and $\big[\tilde{\boldsymbol{N}}_h\big]_{j,i} := (\xi_j, A\phi_i)_{L_2(\Omega)}$.

On the other hand, in the specific inf-sup optimal case $\psi_\mu = B^*(\varphi_\mu)$, we get the representation

$$\begin{aligned}
[\mathbb{B}_\delta]_{(\ell,j),(k,i)} &= b(\psi_\mu, \varphi_\nu) = (\psi_\mu, B^*\varphi_\nu)_{\mathcal{H}} = (B^*\varphi_\mu, B^*\varphi_\nu)_{\mathcal{H}} \\
&= \big(\ddot{\varrho}^\ell \otimes \phi_j + \varrho^\ell \otimes A\phi_j, \ddot{\varrho}^k \otimes \phi_i + \varrho^k \otimes A\phi_i\big)_{\mathcal{H}} \\
&= \big(\ddot{\varrho}^\ell, \ddot{\varrho}^k\big)_{L_2(I)} (\phi_j, \phi_i)_{L_2(\Omega)} + \big(\varrho^\ell, \varrho^k\big)_{L_2(I)} (A\phi_j, A\phi_i)_{L_2(\Omega)} \\
&\quad + \big(\ddot{\varrho}^\ell, \varrho^k\big)_{L_2(I)} (\phi_j, A\phi_i)_{L_2(\Omega)} + \big(\varrho^\ell, \ddot{\varrho}^k\big)_{L_2(I)} (A\phi_j, \phi_i)_{L_2(\Omega)}
\end{aligned}$$

so that

$$\mathbb{B}_\delta = \boldsymbol{Q}_{\Delta t} \otimes \boldsymbol{M}_h + \boldsymbol{N}_{\Delta t} \otimes \boldsymbol{N}_h^\top + \boldsymbol{N}_{\Delta t}^\top \otimes \boldsymbol{N}_h + \boldsymbol{M}_{\Delta t} \otimes \boldsymbol{Q}_h, \text{for } \psi_\mu = B^*(\varphi_\mu), \tag{4.2}$$

where

$$\begin{aligned}
&[\boldsymbol{Q}_{\Delta t}]_{\ell,k} := \big(\ddot{\varrho}^\ell, \ddot{\varrho}^k\big)_{L_2(I)}, && [\boldsymbol{M}_{\Delta t}]_{\ell,k} := \big(\varrho^\ell, \varrho^k\big)_{L_2(I)}, && [\boldsymbol{N}_{\Delta t}]_{\ell,k} := \big(\ddot{\varrho}^\ell, \varrho^k\big)_{L_2(I)}, \\
&[\boldsymbol{Q}_h]_{j,i} := (A\phi_j, A\phi_i)_{L_2(\Omega)}, && [\boldsymbol{M}_h]_{j,i} := (\phi_j, \phi_i)_{L_2(\Omega)}, && [\boldsymbol{N}_h]_{j,i} := (A\phi_j, \phi_i)_{L_2(\Omega)}.
\end{aligned}$$

We stress that $\mathbb{B}_\delta$ is symmetric and positive definite for $A = -\Delta$. Finally, let us now detail the right-hand side. Recall from (2.14), that $g(v) = (f, v)_{\mathcal{H}} + \langle u_1, v(0)\rangle - (u_0, \dot{v}(0))_H$. Hence,

$$\begin{aligned}
[\boldsymbol{g}_\delta]_\nu = [\boldsymbol{g}_\delta]_{(k,i)} &= (f, \varphi_\nu)_{\mathcal{H}} + \langle u_1, \varphi_\nu(0)\rangle_{V' \times V} - (u_0, \dot{\varphi_\nu}(0))_H \\
&= \big(f, \varrho^k \otimes \phi_i\big)_{\mathcal{H}} + \langle u_1, \varphi_\nu(0)\rangle_{V' \times V} - (u_0, \dot{\varphi_\nu}(0))_H \\
&= \int_0^T \int_\Omega f(t,x)\, \varrho^k(t)\, \phi_i(x)\, \mathrm{d}x\, \mathrm{d}t + \int_\Omega \big[u_1(x)\, \varrho^k(0) - u_0(x)\, \dot{\varrho}^k(0)\big] \phi_i(x)\, \mathrm{d}x.
\end{aligned}$$

Using appropriate quadrature formulae results in a numerical approximation, which we will again denote by $\boldsymbol{g}_\delta$. Then, solving the linear system $\mathbb{B}_\delta \boldsymbol{u}_\delta = \boldsymbol{g}_\delta$ yields the expansion coefficients of the desired approximation $u_\delta \in \mathbb{U}_\delta$ as follows: Let $\boldsymbol{u}_\delta = (u_\mu)_{\mu=1,\dots,\mathcal{N}_\delta}$, $\mu = (k,i)$, then

$$u_\delta(t,x) = \sum_{\mu=1}^{\mathcal{N}_\delta} u_\mu\, \psi_\mu(t,x) = \sum_{k=1}^{N_t} \sum_{i=1}^{N_h} u_{k,i}\, \sigma^k(x)\, \xi_i(x),$$

in the general case, and for the special one, *i.e.*, $\psi_\mu = B^*(\varphi_\mu)$,

$$u_\delta(t,x) = \sum_{\mu=1}^{\mathcal{N}_\delta} u_\mu\, \psi_\mu(t,x) = \sum_{k=1}^{N_t} \sum_{i=1}^{N_h} u_{k,i} \big(\ddot{\varrho}^k(t)\, \phi_i(x) + \varrho^k(t)\, A\phi_i(x)\big).$$

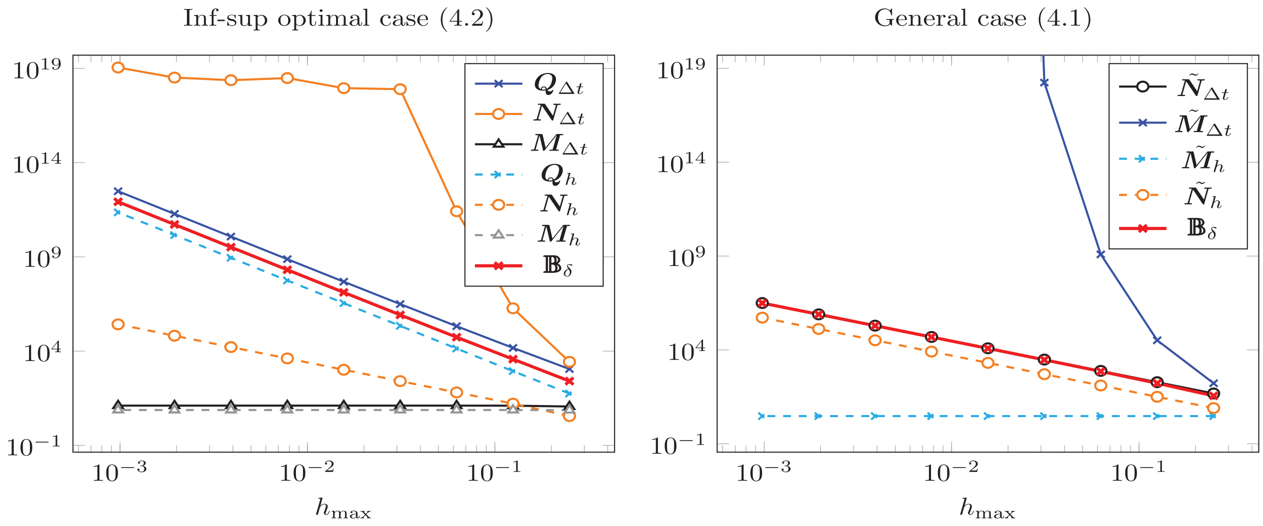


FIGURE 4. Condition numbers of involved matrices, for the inf-sup-optimal case (4.2) (*left*) and the general case (4.1) (*right*).

4.2. Stability *vs.* conditioning

The (discrete) inf-sup constant refers to the stability of the discrete system, being included in the error/residual relation

$$\|u^* - u_\delta^*\|_{\mathbb{U}} \leq \frac{1}{\beta} \sup_{v \in \mathbb{V}} \frac{g(v) - b(u_\delta^*, v)}{\|v\|_{\mathbb{V}}} = \frac{1}{\beta} \|r_\delta\|_{\mathbb{V}'},$$

where the residual $r_\delta \in \mathbb{V}'$ is defined as usual by $r_\delta(v) := g(v) - b(u_\delta^*, v)$, $v \in \mathbb{V}$. Hence, β is a measure for the stability; its value is the minimal generalized eigenvalue of a generalized eigenvalue problem. This has no effect on the condition number $\kappa(\mathbb{B}_\delta)$, which instead governs the accuracy of direct solvers and convergence of iterative methods in the symmetric case.

Conditioning of the matrices

We report on the condition numbers of the matrices involved in (4.1) and (4.2). In Figure 4, we see the asymptotic behavior of the different matrices. Most matrices show a "normal" scaling w.r.t. the order of the differential operator. However, there are two components, namely $\tilde{\mathbf{M}}_{\Delta t}$ in the general case and $\mathbf{N}_{\Delta t}$ in the inf-sup-optimal case, which show a very poor scaling as the mesh size tends to zero (here indicated by $h_{\max}$ but used for both Δt and h). This effect comes from the initial condition, namely the first column of the matrices. On the other hand, $\kappa(\mathbb{B}_\delta)$ scales like a stiffness matrix of a 4th order problem. Since $\mathbb{B}_\delta$ is a sum of tensor products involving some ill-conditioned components, we need a structure-aware preconditioning.

Preconditioning

Let $\mathbb{M}_\delta := \boldsymbol{M}_{\Delta t} \otimes \boldsymbol{M}_h$ and $\mathbb{K}_\delta := \boldsymbol{N}_{\Delta t} \otimes \boldsymbol{M}_h + \boldsymbol{M}_{\Delta t} \otimes \boldsymbol{N}_h$. Then

$$\mathbb{K}_\delta^\top \mathbb{M}_\delta^{-1} \mathbb{K}_\delta = \left(\boldsymbol{N}_{\Delta t}^\top \boldsymbol{M}_{\Delta t}^{-1} \boldsymbol{N}_{\Delta t}\right) \otimes \boldsymbol{M}_h + \boldsymbol{N}_{\Delta t} \otimes \boldsymbol{N}_h^\top + \boldsymbol{N}_{\Delta t}^\top \otimes \boldsymbol{N}_h + \boldsymbol{M}_{\Delta t} \otimes \left(\boldsymbol{N}_h^\top \boldsymbol{M}_h^{-1} \boldsymbol{N}_h\right),$$

so that $\mathbb{K}_\delta^\top \mathbb{M}_\delta^{-1} \mathbb{K}_\delta = \mathbb{B}_\delta$ if and only if $\boldsymbol{Q}_{\Delta t} = \boldsymbol{N}_{\Delta t}^\top \boldsymbol{M}_{\Delta t}^{-1} \boldsymbol{N}_{\Delta t}$ and $\boldsymbol{Q}_h = \boldsymbol{N}_h^\top \boldsymbol{M}_h^{-1} \boldsymbol{N}_h$.

Even if we cannot hope that those relations hold exactly in general, we are going to describe situations in which at least spectral equivalence holds. To this end, we will closely follow [4, 30] in a slightly generalized setting. We recall the biharmonic-type problem (2.11) along with its equivalent mixed form (2.12). Let us abbreviate $Z := D(A)$ and let $Z_h := \operatorname{span}\{\phi_1, \dots, \phi_{N_h}\} \subset Z$ be some discretization as in (3.5). Moreover, let $H_h := \operatorname{span}\{\xi_1, \dots, \xi_{n_h}\} \subset H$ be some finite-dimensional approximation space for the auxiliary variable. Then, setting

$$\boldsymbol{M}_h := \big[(\xi_i, \xi_j)_H\big]_{i,j=1,\dots,n_h}, \quad \boldsymbol{A}_h := \big[(A\phi_k, \xi_j)_H\big]_{k=1,\dots,N_h, j=1,\dots,n_h},$$

the discrete form of (2.12) aims to determine $\boldsymbol{u}_h \in \mathbb{R}^{n_h}$ and $\boldsymbol{z}_h \in \mathbb{R}^{N_h}$ such that

$$\begin{pmatrix} \boldsymbol{M}_h & \boldsymbol{A}_h \\ \boldsymbol{A}_h^\top & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{u}_h \\ \boldsymbol{z}_h \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ -\boldsymbol{g}_h \end{pmatrix}, \tag{4.3}$$

where $\boldsymbol{g}_h = [\langle g, \phi_k \rangle_{Z' \times Z}]_{k=1,\dots,N_h}$. Note that $\boldsymbol{M}_h$ is symmetric and positive definite. The corresponding discrete operators are defined as follows

$$\begin{aligned} A_h : Z_h \to H_h : \quad & (A_h z_h, u_h)_H := (A z_h, u_h)_H, \; u_h \in H_h, z_h \in Z_h, \\ M_h : H_h \to H_h : \quad & (M_h u_h, v_h)_H := (u_h, v_h)_H, \; u_h, v_h \in M_h. \end{aligned}$$

The stiffness matrix for the biharmonic-type problems reads as follows: $\boldsymbol{Q}_h := [(A\phi_k, A\phi_\ell)_H]_{k,\ell=1,\dots,N_h}$. Finally, we define discrete norms on Z_h by $\|z_h\|_{Z_h}^2 := \boldsymbol{z}_h^\top \boldsymbol{Q}_h \boldsymbol{z}_h$ for $z_h = \sum_{k=1}^{N_h} (\boldsymbol{z}_h)_k \, \phi_k \in Z_h$, $\boldsymbol{z}_h \in \mathbb{R}^{N_h}$ and $\|u_h\|_{M_h}^2 = \boldsymbol{u}_h^\top \boldsymbol{M}_h \boldsymbol{u}_h$ for $u_h = \sum_{i=1}^{n_h} (\boldsymbol{u}_h)_i \, \xi_i \in M_h$, $\boldsymbol{u}_h \in \mathbb{R}^{n_h}$.

Proposition 4.1. *Let A_h be* bounded, *i.e., there exists a constant $0 < \Gamma < \infty$ such that $(A_h z_h, u_h)_H \le \Gamma \|z_h\|_{Z_h} \|u_h\|_{M_h}$ for all $u_h \in H_h$ and $z_h \in Z_h$, and* uniformly inf-sup stable, *i.e.,*

$$\inf_{z_h \in Z_h} \sup_{u_h \in H_h} \frac{(A_h z_h, u_h)_H}{\|z_h\|_{Z_h} \|u_h\|_{M_h}} \ge \gamma > 0. \tag{4.4}$$

Then, $\boldsymbol{Q}_h$ and $\boldsymbol{A}_h \boldsymbol{M}_h^{-1} \boldsymbol{A}_h^\top$ are spectrally equivalent, i.e.,

$$\gamma^2 \, \boldsymbol{z}_h^\top \boldsymbol{Q}_h \boldsymbol{z}_h \le \boldsymbol{z}_h^\top \boldsymbol{A}_h \boldsymbol{M}_h^{-1} \boldsymbol{A}_h^\top \boldsymbol{z}_h \le \Gamma^2 \, \boldsymbol{z}_h^\top \boldsymbol{Q}_h \boldsymbol{z}_h \quad \text{for all } \boldsymbol{z}_h \in \mathbb{R}^{N_h}.$$

Proof. The proof follows the lines in 1.9–1.12 from [30]. Let $Z_h \ni z_h = \sum_{k=1}^{N_h} (\boldsymbol{z}_h)_k \, \phi_k$ and $M_h \ni u_h = \sum_{i=1}^{n_h} (\boldsymbol{u}_h)_i \, \xi_i$. Then, by (4.4)

$$\begin{aligned} \gamma \left(\boldsymbol{z}_h^\top \boldsymbol{Q}_h \boldsymbol{z}_h \right)^{1/2} = \gamma \|z_h\|_{Z_h} \le \sup_{u_h \in H_h} \frac{(A_h z_h, u_h)_H}{\|u_h\|_{M_h}} &= \max_{\boldsymbol{u}_h \in \mathbb{R}^{n_h}} \frac{\boldsymbol{z}_h^\top \boldsymbol{A}_h \boldsymbol{u}_h}{\left(\boldsymbol{u}_h^\top \boldsymbol{M}_h \boldsymbol{u}_h \right)^{1/2}} \\ &= \max_{\boldsymbol{v}_h = \boldsymbol{M}_h^{1/2} \boldsymbol{u}_h \in \mathbb{R}^{n_h}} \frac{\boldsymbol{z}_h^\top \boldsymbol{A}_h \boldsymbol{M}_h^{-1/2} \boldsymbol{v}_h}{\left(\boldsymbol{v}_h^\top \boldsymbol{v}_h \right)^{1/2}} = \left(\boldsymbol{z}_h^\top \boldsymbol{A}_h \boldsymbol{M}_h^{-1} \boldsymbol{A}_h^\top \boldsymbol{z}_h \right)^{1/2}, \end{aligned}$$

since it is easily seen that the maximum is attained for $\boldsymbol{v}_h = \boldsymbol{M}_h^{-1/2} \boldsymbol{A}_h^\top \boldsymbol{z}_h$. This proves the first inequality. Using the boundedness of A_h yields $\Gamma \left(\boldsymbol{z}_h^\top \boldsymbol{Q}_h \boldsymbol{z}_h \right)^{1/2} = \Gamma \|z_h\|_{Z_h} \ge \sup_{u_h \in H_h} \frac{(A_h z_h, u_h)_H}{\|u_h\|_{M_h}}$, so that the second inequality follows the lines above. □

Remark 4.2. In Section 4 of [4], the assumptions above have been shown within the so-called *Ciarlet–Raviart method*, where $A = -\Delta$ with homogeneous Dirichlet boundary conditions on a bounded convex polygon $\Omega \subset \mathbb{R}^2$. Then, $D(A) = H^2(\Omega) \cap H_0^1(\Omega)$ and $H = L_2(\Omega)$ – exactly our setting for the wave equation.

Let $\{\mathcal{T}_h\}_{0<h<1}$ be a family of shape regular and quasi uniform triangulations of Ω consisting of triangles of diameter less or equal to h. The next piece consists of mesh dependent norms and spaces defined as $H_h^2 := \{u \in H^1(\Omega) : u|_T \in H^2(T), T \in \mathcal{T}_h\}$, $\Gamma_h := \bigcup_{T\in\mathcal{T}_h} \partial T$ and $\|u\|_{2,h}^2 := \sum_{T\in\mathcal{T}_h} \|u\|_{2,T}^2 + h^{-1}\int_{\Gamma_h} \left|J\frac{\partial u}{\partial \nu}\right|^2 \mathrm{d}s$, where

$$J\frac{\partial u}{\partial \nu}\Big|_{T'} := \begin{cases} \frac{\partial u_{|T_1}}{\partial \nu^1} + \frac{\partial u_{|T_2}}{\partial \nu^2}, & \text{if } T' = \partial T^1 \cap \partial T^2 \text{ is an interior edge of } \mathcal{T}_h, \\ \frac{\partial u}{\partial \nu}, & \text{if } T' \text{ is a boundary edge of } \mathcal{T}_h, \end{cases}$$

and ν^j denotes the unit outward normal of T^j. Next, let

$$\|u\|_{0,h}^2 := \|u\|_{L_2(\Omega)}^2 + h\int_{\Gamma_h} |u(s)|^2 \,\mathrm{d}s, \qquad u \in H^1(\Omega)$$

and define H_h^0 as the completion of $H^1(\Omega)$ w.r.t. $\|\cdot\|_{0,h}$. Then $H_h^0 \cong L_2(\Omega) \oplus L_2(\Gamma_h)$. For $S_h := \{v \in C^0(\bar\Omega) : v|_T \in \mathcal{P}_k, T \in \mathcal{T}_h\}$, $k \geq 1$ and $\mathcal{P}_k$ denoting the space of polynomials of degree k or less, we have that $S_h \subset H_h^0 \cap H_h^2$. For $k = 3$, we get that $S_h = Z_h$ with Z_h defined in (3.5). The discrete operator A_h is induced by the bilinear form $a_h : H_h^0 \times H_h^2 \cap H_0^1(\Omega) \to \mathbb{R}$ defined by

$$a_h(u_h, w_h) := \sum_{T\in\mathcal{T}_h} \int_T u_h(x)\,\Delta w_h(x)\,\mathrm{d}x - \int_{\Gamma_h} u_h\Big(J\frac{\partial w_h}{\partial \nu}\Big)\,\mathrm{d}s$$

and M_h is induced by $m_h(u,v) := (u,v)_{L_2(\Omega)}$. The discrete spaces arise there from a shape regular and quasi uniform triangulation of Ω as well as mesh-dependent inner products and norms. The boundedness of A_h is immediate. The inf-sup-stability (4.4) was proven in Theorem 3 from [4].

Noting that $a_h(u_h, w_h) = -(\nabla u_h, \nabla w_h)_{L_2(\Omega)}$ for $u \in H^1(\Omega)$ and $w_h \in H_h^2$, we obtain that the matrices $\boldsymbol{Q}_h$ and $\boldsymbol{N}_h^\top \boldsymbol{M}_h^{-1} \boldsymbol{N}_h$ defined in Section 4.1 are in fact spectrally equivalent.

We observed the spectral equivalence for the spatial matrices also in our numerical experiments. However, we saw that this is not true for the temporal matrices in the sense that $\boldsymbol{Q}_{\Delta t}$ and $\boldsymbol{N}_{\Delta t}^\top \boldsymbol{M}_{\Delta t}^{-1} \boldsymbol{N}_{\Delta t}$ are not spectrally equivalent.

5. Solution of the algebraic linear system

To derive preconditioning strategies and the new projection method, we rewrite the linear system $\mathbb{B}_\delta \boldsymbol{u}_\delta = \boldsymbol{g}_\delta$ as a linear matrix equation, so as to exploit the structure of the Kronecker problem. Let $\boldsymbol{x} = \mathrm{vec}(\boldsymbol{X})$ be the operator stacking the columns of $\boldsymbol{X}$ one after the other, then it holds that $(\boldsymbol{B} \otimes \boldsymbol{A})\boldsymbol{x} = \mathrm{vec}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}^\top)$ for given matrices $\boldsymbol{A}, \boldsymbol{X}$, and $\boldsymbol{B}$ of conforming dimensions. Hence, the vector system is written as

$$\mathcal{A}(\boldsymbol{U}) = \boldsymbol{G} \quad \text{with} \quad \mathcal{A}(\boldsymbol{U}) = \boldsymbol{M}_h \boldsymbol{U} \boldsymbol{Q}_{\Delta t}^\top + \boldsymbol{N}_h^\top \boldsymbol{U} \boldsymbol{N}_{\Delta t}^\top + \boldsymbol{N}_h \boldsymbol{U} \boldsymbol{N}_{\Delta t} + \boldsymbol{Q}_h \boldsymbol{U} \boldsymbol{M}_{\Delta t}, \tag{5.1}$$

where $\boldsymbol{g} = \mathrm{vec}(\boldsymbol{G})$ and the symmetry of some of the matrices has been exploited.

In the following we describe two distinct approaches: First, we recall the matrix-oriented conjugate gradient method, preconditioned by two different operator-aware strategies. Then we discuss a procedure that directly deals with (5.1).

5.1. Preconditioned conjugate gradients

Since $\mathbb{B}_\delta$ is symmetric and positive definite, the preconditioned conjugate gradient (PCG) method can be applied directly to (5.1), yielding a matrix-oriented implementation of PCG, see Algorithm 1. Here $\mathrm{tr}(\boldsymbol{X})$ denotes the trace of the square matrix $\boldsymbol{X}$. In exact precision arithmetic, this formulation, gives the same iterates as the standard vector form, while exploiting matrix-matrix computations [22]. This can easily be seen by exploiting

Algorithm 1. Matrix-oriented PCG.

```
Input: $\boldsymbol{U}_0$
set $\boldsymbol{R}_0 = \boldsymbol{G} - \mathcal{A}(\boldsymbol{U}_0)$, $\boldsymbol{Z}_0 = \mathcal{P}^{-1}(\boldsymbol{R}_0)$, $\boldsymbol{P}_0 = \boldsymbol{Z}_0$, $\gamma_0 = \mathrm{tr}\big(\boldsymbol{R}_0^\top \boldsymbol{Z}_0\big)$
for $k = 0, 1, \ldots$ do
    $\delta = \mathrm{tr}\big(\boldsymbol{P}_k^\top \mathcal{A}(\boldsymbol{P}_k)\big)$, $\alpha = \gamma_k/\delta$
    $\boldsymbol{X}_{k+1} = \boldsymbol{X}_k + \alpha \boldsymbol{P}_k$
    $\boldsymbol{R}_{k+1} = \boldsymbol{G} - \mathcal{A}(\boldsymbol{X}_{k+1})$
    $\boldsymbol{Z}_{k+1} = \mathcal{P}^{-1}(\boldsymbol{R}_{k+1})$
    $\gamma_{k+1} = \mathrm{tr}\big(\boldsymbol{R}_{k+1}^\top \boldsymbol{Z}_{k+1}\big)$, $\beta = \gamma_{k+1}/\gamma_k$
    $\boldsymbol{P}_{k+1} = \boldsymbol{Z}_{k+1} + \beta \boldsymbol{P}_k$
end for
```

the properties of the Kronecker product and the matrix inner product, namely $\mathrm{vec}(AXB) = (B^T \otimes A)\mathrm{vec}(X)$ and $\mathrm{vec}(A)^T\mathrm{vec}(B) = \mathrm{tr}(A^T B)$ for conforming A, B and X; see, *e.g.*, Section 1.3.7 of [17][8].

5.1.1. Sylvester operator preconditioning.

A natural preconditioning strategy consists of taking the leading part of the coefficient matrix, in terms of order of the differential operators. Hence, setting $\mathbb{P} = \boldsymbol{Q}_{\Delta t} \otimes \boldsymbol{M}_h + \boldsymbol{M}_{\Delta t} \otimes \boldsymbol{Q}_h$, we have (see also [19])

$$\boldsymbol{z}_{k+1} = \mathbb{P}^{-1}\boldsymbol{r}_{k+1} \quad \Leftrightarrow \quad \boldsymbol{Z}_{k+1} = \mathcal{P}^{-1}(\boldsymbol{R}_{k+1}),$$

with $\boldsymbol{r}_{k+1} = \mathrm{vec}(\boldsymbol{R}_{k+1})$ and $\boldsymbol{z}_{k+1} = \mathrm{vec}(\boldsymbol{Z}_{k+1})$. Applying $\mathcal{P}^{-1}$ corresponds to solving the generalized Sylvester equation $\boldsymbol{M}_h \boldsymbol{Z} \boldsymbol{Q}_{\Delta t}^\top + \boldsymbol{Q}_h \boldsymbol{Z} \boldsymbol{M}_{\Delta t} = \boldsymbol{R}_{k+1}$. For small size problems in space, this can be carried out by means of the Bartels-Stewart method [7], which entails the computation of two Schur decompositions, performed before the PCG iteration is started. For fine discretizations in space, iterative procedures need to be used. For these purposes, we use a Galerkin approach based on the rational Krylov subspace [14], only performed on the spatial matrices; see [31] for a general discussion. A key issue is that this class of iterative methods requires the right-hand side to be low rank; we deliberately set the rank to be at most twenty. Hence, the Sylvester solver is applied after a rank truncation of $\boldsymbol{R}_{k+1}$, which thus becomes part of the preconditioning application.

5.1.2. $\mathbb{K}_\delta^\top \mathbb{M}_\delta^{-1} \mathbb{K}_\delta$*-preconditioning.*

To derive a preconditioner that takes full account of the coefficient matrix we employ the operator $\mathbb{K}_\delta^\top \mathbb{M}_\delta^{-1} \mathbb{K}_\delta$ in Section 4.2. Thanks to the spectral equivalence in Proposition 4.1, PCG applied to the resulting preconditioned operator appears to be optimal, in the sense that the number of iterations to reach the required accuracy is independent of the spatial mesh size; see Table 2.

In vector form this preconditioner is applied as $\boldsymbol{z}_{k+1} = \big(\mathbb{K}_\delta^\top \mathbb{M}_\delta^{-1} \mathbb{K}_\delta\big)^{-1} \boldsymbol{r}_{k+1}$. However, this operation can be performed without explicitly using the Kronecker form of the involved matrices, with significant computational and memory savings. We observe that

$$\mathbb{K}_\delta = \boldsymbol{N}_{\Delta t} \otimes \boldsymbol{M}_h + \boldsymbol{M}_{\Delta t} \otimes \boldsymbol{N}_h = \big(\boldsymbol{N}_{\Delta t}\boldsymbol{M}_{\Delta t}^{-1} \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{N}_h \boldsymbol{M}_h^{-1}\big)(\boldsymbol{M}_{\Delta t} \otimes \boldsymbol{M}_h) =: \widehat{\mathbb{K}}_\delta \mathbb{M}_\delta.$$

Moreover, due to the transposition properties of the Kronecker product, $\mathbb{K}_\delta^\top = \widehat{\mathbb{K}}_\delta^\top \mathbb{M}_\delta$. Hence, $\mathbb{K}_\delta^\top \mathbb{M}_\delta^{-1} \mathbb{K}_\delta = \widehat{\mathbb{K}}_\delta^\top \widehat{\mathbb{K}}_\delta \mathbb{M}_\delta$. Therefore,

$$\boldsymbol{Z}_{k+1} = \mathcal{P}^{-1}(\boldsymbol{R}_{k+1}) \;\Leftrightarrow\; \boldsymbol{z}_{k+1} = \mathbb{M}_\delta^{-1} \widehat{\mathbb{K}}_\delta^{-1} \Big(\widehat{\mathbb{K}}_\delta^\top\Big)^{-1} \boldsymbol{r}_{k+1},$$

We next observe that the equation $\widehat{\mathbb{K}}_\delta^\top \boldsymbol{w} = \boldsymbol{r}_{k+1}$ can be written as the following Sylvester matrix equation

$$\boldsymbol{W}\boldsymbol{M}_{\Delta t}^{-1}\boldsymbol{N}_{\Delta t} + \boldsymbol{N}_h^\top \boldsymbol{M}_h^{-1} \boldsymbol{W} = \boldsymbol{R}_{k+1} \tag{5.2}$$

[8] The matrix-oriented version of PCG is also used to exploit low rank representations of the iterates, in case the starting residual is low rank and the final solution can be well approximated by a low rank matrix; see, *e.g.*, [22]. We will not exploit this setting here.

and analogously for $\widehat{\mathbb{K}}_\delta \widehat{\boldsymbol{w}} = \boldsymbol{w}$, that is

$$\widehat{\boldsymbol{W}} \boldsymbol{M}_{\Delta t}^{-1} \boldsymbol{N}_{\Delta t}^\top + \boldsymbol{N}_h \boldsymbol{M}_h^{-1} \widehat{\boldsymbol{W}} = \boldsymbol{W}. \tag{5.3}$$

Finally, the preconditioned matrix is obtained as $\boldsymbol{Z}_{k+1} = \boldsymbol{M}_h^{-1} \widehat{\boldsymbol{W}} \boldsymbol{M}_{\Delta t}^{-1}$.

Summarizing, the application of the preconditioning operator amounts to the solution of the two Sylvester matrix equations (5.2), (5.3), and the product $\boldsymbol{Z}_{k+1} = \boldsymbol{M}_h^{-1} \widehat{\boldsymbol{W}} \boldsymbol{M}_{\Delta t}^{-1}$. The overall computational cost of this operation depends on the cost of solving the two matrix equations. For small dimensions in space, once again a Schur-decomposition based method can be used [7]; we recall here that thanks to the discretization employed, we do not expect to have large dimensions in time, as matrices of size at most $\mathcal{O}(100)$ arise. Also in this case, for fine discretizations in space we use an iterative method (Galerkin) based on the rational Krylov subspace [14], only performed on the spatial matrices, with the truncation of the corresponding right-hand side, $\boldsymbol{R}_{k+1}$ and $\boldsymbol{W}$, respectively, so as to have at most rank equal to twenty. Allowing a larger rank did not seem to improve the effectiveness of the preconditioner. Several implementation enhancements can be developed to make the action of the preconditioner more efficient, since most operations are repeated at each PCG iteration with the same matrices.

5.2. Galerkin projection

An alternative to PCG consists of attacking the original multi-term matrix equation directly. Thanks to the symmetry of $\boldsymbol{N}_h$ we rewrite the matrix equation (5.1) as

$$\boldsymbol{M}_h \boldsymbol{U} \boldsymbol{Q}_{\Delta t}^\top + \boldsymbol{N}_h^\top \boldsymbol{U} \big(\boldsymbol{N}_{\Delta t}^\top + \boldsymbol{N}_{\Delta t} \big) + \boldsymbol{Q}_h \boldsymbol{U} \boldsymbol{M}_{\Delta t} = \boldsymbol{G}, \tag{5.4}$$

with $\boldsymbol{G}$ of low rank, that is $\boldsymbol{G} = \boldsymbol{G}_1 \boldsymbol{G}_2^\top$. Note that this is an assumption on the data. In particular, we assume that the right-hand side $g(v)$ in (2.14) can be discretized in a way such that the matrix form $\boldsymbol{G}$ of $\boldsymbol{g}_\delta$ has low rank. This happens for instance when g is a separable function of x and t, or it can be well approximated by a separable function; other scenarios can also lead to a low-rank $\boldsymbol{G}$.

Consider two appropriately selected vector spaces $\mathcal{V}_k$, $\mathcal{W}_k$ of dimensions much lower than N_h, N_t, respectively, and let $\boldsymbol{V}_k$, $\boldsymbol{W}_k$ be the matrices whose orthonormal columns span the two corresponding spaces. We look for a low rank approximation of $\boldsymbol{U}$ as $\boldsymbol{U}_k = \boldsymbol{V}_k \boldsymbol{Y}_k \boldsymbol{W}_k^\top$. To determine $\boldsymbol{Y}_k$ we impose an orthogonality (Galerkin) condition on the residual

$$\boldsymbol{R}_k := \boldsymbol{G}_1 \boldsymbol{G}_2^\top - \boldsymbol{M}_h \boldsymbol{U}_k \boldsymbol{Q}_{\Delta t}^\top - \boldsymbol{N}_h^\top \boldsymbol{U}_k \big(\boldsymbol{N}_{\Delta t}^\top + \boldsymbol{N}_{\Delta t} \big) - \boldsymbol{Q}_h \boldsymbol{U}_k \boldsymbol{M}_{\Delta t}. \tag{5.5}$$

with respect to the generated space pair $(\boldsymbol{V}_k, \boldsymbol{W}_k)$. Using the matrix Euclidean inner product, this corresponds to imposing that $\boldsymbol{V}_k^\top \boldsymbol{R}_k \boldsymbol{W}_k = 0$. Substituting $\boldsymbol{R}_k$ and $\boldsymbol{U}_k$ into this matrix equation, we obtain the following *reduced* matrix equation, of the same type as (5.4) but of much smaller size,

$$\begin{aligned}\big(\boldsymbol{V}_k^\top \boldsymbol{M}_h \boldsymbol{V}_k\big) \boldsymbol{Y}_k \big(\boldsymbol{Q}_{\Delta t}^\top \boldsymbol{W}_k\big) + \big(\boldsymbol{V}_k^\top \boldsymbol{N}_h^\top \boldsymbol{V}_k\big) \boldsymbol{Y}_k \big(\boldsymbol{W}_k^\top \big(\boldsymbol{N}_{\Delta t}^\top + \boldsymbol{N}_{\Delta t}\big) \boldsymbol{W}_k\big)& \\ + \big(\boldsymbol{V}_k^\top \boldsymbol{Q}_h \boldsymbol{V}_k\big) \boldsymbol{Y}_k \big(\boldsymbol{W}_k^\top \boldsymbol{M}_{\Delta t} \boldsymbol{W}_k\big) = \big(\boldsymbol{V}_k^\top \boldsymbol{G}_1\big)\big(\boldsymbol{G}_2^\top \boldsymbol{W}_k\big).&\end{aligned}$$

The small dimensional matrix $\boldsymbol{Y}_k$ is thus obtained by solving the Kronecker form of this equation[9]. The described Galerkin reduction strategy has been thorough exploited and analyzed for Sylvester equations, and more recently successfully applied to multi-term equations, see, *e.g.*, [28]. The key problem-dependent ingredient is the choice of the spaces $\mathcal{V}_k$, $\mathcal{W}_k$, so that they well represent spectral information of the "left-hand" and "right-hand" matrices in (5.4). A well established choice is (a combination of) rational Krylov subspaces [31]. More precisely, for the spatial approximation we generate the growing space range $(\boldsymbol{V}_k)$ as

$$\widehat{\boldsymbol{V}}_{k+1} = \Big[\boldsymbol{V}_k, (\boldsymbol{Q}_h + s_k \boldsymbol{M}_h)^{-1} \boldsymbol{v}_k, (\boldsymbol{N}_h + \sqrt{s_k} \boldsymbol{M}_h)^{-1} \boldsymbol{v}_k\Big], \qquad \boldsymbol{V}_1 = \boldsymbol{G}_1,$$

[9] To this end, Algorithm 1 with a preconditioning strategy similar to the ones described in Sections 5.1.1 and 5.1.2 can be employed as well.

where $\boldsymbol{v}_k$ is the kth column of $\boldsymbol{V}_k$, so that $\boldsymbol{V}_{k+1}$ is obtained by orthogonalizing the new columns inserted in $\widehat{\boldsymbol{V}}_{k+1}$. The matrix $\widehat{\boldsymbol{V}}_{k+1}$ grows at most by two vectors at a time. For each k, the parameter s_k can be chosen either *a priori* or dynamically, with the same sign as the spectrum of $\boldsymbol{Q}_h$ ($\boldsymbol{N}_h$). Here s_k is cheaply determined using the adaptive strategy in [14]. Since $\boldsymbol{N}_h$ represents an operator of the second order, the value $\sqrt{s_k}$ resulted to be appropriate; a specific computation of the parameter associated with $\boldsymbol{N}_h$ can also be included, at low cost. Analogously,

$$\widehat{\boldsymbol{W}}_{k+1} = \left[\boldsymbol{W}_k, (\boldsymbol{Q}_{\Delta t} + \ell_k \boldsymbol{M}_{\Delta t})^{-1}\boldsymbol{w}_k, \left((\boldsymbol{N}_{\Delta t} + \boldsymbol{N}_{\Delta t}^\top) + \sqrt{\ell_k}\boldsymbol{M}_{\Delta t}\right)^{-1}\boldsymbol{w}_k\right], \qquad \boldsymbol{W}_1 = \boldsymbol{G}_2,$$

where $\boldsymbol{w}_k$ is the kth column of $\boldsymbol{W}_k$, and $\boldsymbol{W}_{k+1}$ is obtained by orthogonalizing the new columns inserted in $\widehat{\boldsymbol{W}}_{k+1}$. The choice of $\ell_k > 0$ is made as for s_k.

Remark 5.1. This approach yields the vector approximation $\boldsymbol{u}_k = (\boldsymbol{W}_k \otimes \boldsymbol{V}_k)\boldsymbol{y}_k$, with $\boldsymbol{y}_k = \text{vec}(\boldsymbol{Y}_k)$ that is, the approximation space range $(\boldsymbol{W}_k \otimes \boldsymbol{V}_k)$ is more structured than that generated by PCG applied to $\mathcal{A}$. Experimental evidence shows that this structure-aware space requires significantly smaller dimension to achieve similar accuracy. This is theoretically clear in the Sylvester equation case [31], while it is an open problem for the multi-term linear equation setting.

Remark 5.2. For fine space discretizations, the most expensive step of the Galerkin projection is the solution of the linear systems with $(\boldsymbol{Q}_h + s_k\boldsymbol{M}_h)$ and $\left(\boldsymbol{N}_h + \sqrt{s_k}\boldsymbol{M}_h\right)$. Depending on the size and sparsity, these systems can be solved by either a sparse direct method or by an iterative procedure; see [31] and references therein.

6. Numerical experiments

We report some results of our extensive numerical experiments for the wave equation (2.1) with $A = -c^2\Delta$, $H = L_2(\Omega)$, $\Omega = (0,1)^3$, $c \neq 0$ being the wave speed, $V = H_0^1(\Omega)$ and $I = (0,1)$, *i.e.*, $T = 1$. We set $u_1 \equiv f \equiv 0$ and choose the initial value u_0 in such a way that the respective solutions have different regularity (cases 1 and 2 below).

If $\boldsymbol{U}_k$ denotes the current approximate solution computed at iteration k, Algorithm 1 and the Galerkin method are stopped as soon as (i) $\mathcal{E}_k < 10^{-5}$, where the backward error $\mathcal{E}_k$ is defined as

$$\mathcal{E}_k = \frac{\|\boldsymbol{R}_k\|_F}{\|\boldsymbol{G}\|_F + \|\boldsymbol{U}_k\|_F(\|\boldsymbol{M}_h\|_F\|\boldsymbol{Q}_{\Delta t}\|_F + \|\boldsymbol{Q}_h\|_F\|\boldsymbol{M}_{\Delta t}\|_F + 2\|\boldsymbol{N}_h\|_F\|\boldsymbol{N}_{\Delta t}\|_F)},$$

and $\boldsymbol{R}_k$ is the residual matrix defined in (5.5), and (ii) $\|\boldsymbol{R}_k\|_F/\|\boldsymbol{G}\|_F < 10^{-2}$ for the relative residual norm. For the Galerkin approach the computation of $\mathcal{E}_k$ simplifies thanks to the low-rank format of the involved quantities (for instance, $\boldsymbol{R}_k$ does not need to be explicitly formed to compute its norm). Moreover, the linear systems in the rational Krylov subspace basis construction are solved by the vector PCG method with a tolerance $\epsilon = 10^{-8}$; see Remark 5.2.

We compared the space-time method with the classical Crank–Nicolson time stepping scheme, in terms of approximation accuracy and CPU time. The $N_h \times N_h$ linear systems involved in the time marching scheme are solved by means of the vector PCG method with tolerance $\epsilon = 10^{-6}$. The time stepping scheme is also used to compute the reference solutions. To this end, we chose 1024 time steps and 64 unknowns in every space dimension, resulting in $2.68 \cdot 10^8$ degrees of freedom. For the error calculation, we evaluated the solutions on a grid of 64 points in every dimension and approximated the L_2 error through the $1.67 \cdot 10^7$ query points.

The code[10] is run in `Matlab` and the B-spline implementation is based on [25][11]. To explore the potential of the new ultraweak method on low-regularity solutions, we only concentrate on experiments with lower regularity

[10] The whole code can be found under https://github.com/j-henning/waveRB.

[11] Executed on the BwUniCluster 2.0 on instances with 32 GB of RAM on two cores of an Intel Xeon Gold 6230.

TABLE 2. Case 1: Continuous solution, $c^2 = 1$.

Refinement		2	3	4	5
Unknowns	Time	4	8	16	32
	Space	64	515	4096	32 768
Time stepping	Wall time [s]	$1.87 \cdot 10^{-2}$	$2.00 \cdot 10^{-2}$	$6.82 \cdot 10^{-1}$	$8.85 \cdot 10^{0}$
	L_2 error	$4.27 \cdot 10^{-1}$	$7.10 \cdot 10^{-2}$	$2.85 \cdot 10^{-2}$	$1.20 \cdot 10^{-2}$
Space-time	L_2 error	$5.31 \cdot 10^{-2}$	$4.48 \cdot 10^{-2}$	$3.45 \cdot 10^{-2}$	$2.50 \cdot 10^{-2}$
PCG $(\mathbb{K}_\delta^\top \mathbf{M}_\delta^{-1} \mathbb{K}_\delta)$	Wall time [s]	$9.59 \cdot 10^{-2}$	$9.73 \cdot 10^{-1}$	$2.17 \cdot 10^{1}$	$1.07 \cdot 10^{3}$
	# Iterations	9	7	7	6
PCG (Sylvester)	Wall time [s]	$3.01 \cdot 10^{-3}$	$3.85 \cdot 10^{-2}$	$1.55 \cdot 10^{0}$	$8.57 \cdot 10^{1}$
	# Iterations	9	9	13	17
Galerkin	Wall time [s]	$2.24 \cdot 10^{-1}$	$6.21 \cdot 10^{-1}$	$1.20 \cdot 10^{0}$	$1.61 \cdot 10^{1}$
	# Iterations	2	6	15	24

solutions, in particular a solution which is continuous with discontinuous derivative (Case 1) and a discontinuous solution (Case 2). This is realized through the choice of u_0. On the other hand, for smooth solutions the time-stepping method would be expected to be more accurate, due to its second-order convergence, compared to the ultraweak method, as long as the latter uses piecewise constant trial functions.

We describe our results for the 3D setting, with $\Omega = (0,1)^3$. The data are summarized as follows[12]

	Case 1	Case 2
$u_0(r)$	$\left(1 - \frac{5}{\sqrt{2}} r\right) \cdot \mathbb{1}_{r<\sqrt{2}/5}$	$\mathbb{1}_{r<\sqrt{2}/5}$
u	$\in C\big(\bar{I} \times \bar{\Omega}\big) \setminus C^1(I \times \Omega)$	$\notin C\big(\bar{I} \times \bar{\Omega}\big)$

We use tensor product spaces for the spatial discretization for both approaches. In the space-time setting we use B-splines in each direction for the test functions. For the time-stepping method, we use a Galerkin approach in which the trial and test functions are given by B-splines. Hence, the radial symmetry cannot be exploited by either methods, and the tensor product approach provides no limitation. All tables show the matrix dimensions N_t in the time space ("Time") and N_h in the spatial space ("Space"). We display results for uniform discretizations in space and time, where $N_h = N_t^3$, but stress the fact that our space-time discretization is *unconditionally* stable, *i.e.*, for any combination of N_t and N_h.

6.1. Case 1: Continuous, but not continuously differentiable solution

The L_2-error, the number of iterations and the wall-clock time (using the Matlab tic-toc commands) for all described methods are displayed in Table 2 for the case $c^2 = 1$. We start by comparing the three iterative methods for solving the ultraweak space-time discretization. Comparing the performance of the two preconditioners in Sections 5.1.1 and 5.1.2, we note that they are applied inexactly as described above. In spite of this, we notice that the optimal operator-preconditioner is able to maintain mesh independence, thus experimentally confirming our theoretical results. Unfortunately, the high complexity of this preconditioner results in excessive computational costs on the finer discretizations. The Sylvester preconditioner is the fastest method for small discretization sizes, whereas the Galerkin method needs smaller wall times for larger systems. The data of Table 2 is also visualized in the first column of Figure 5, where we see that the time-stepping method outperforms the ultraweak space-time approach both w.r.t. accuracy and efficiency.

This changes somewhat for larger values of the wave speed c^2, see the results in Figure 5. We see that the error of the space-time method is basically independent of the wave speed, whereas the time-stepping method suffers

[12] We use polar coordinates, *i.e.*, $r := \|\boldsymbol{x} - \mathbf{c}\|$, $\boldsymbol{x} \in \Omega$, with center $\boldsymbol{c} = (c_i)_{i=1,\dots,d}$, $c_i = 0.5$.

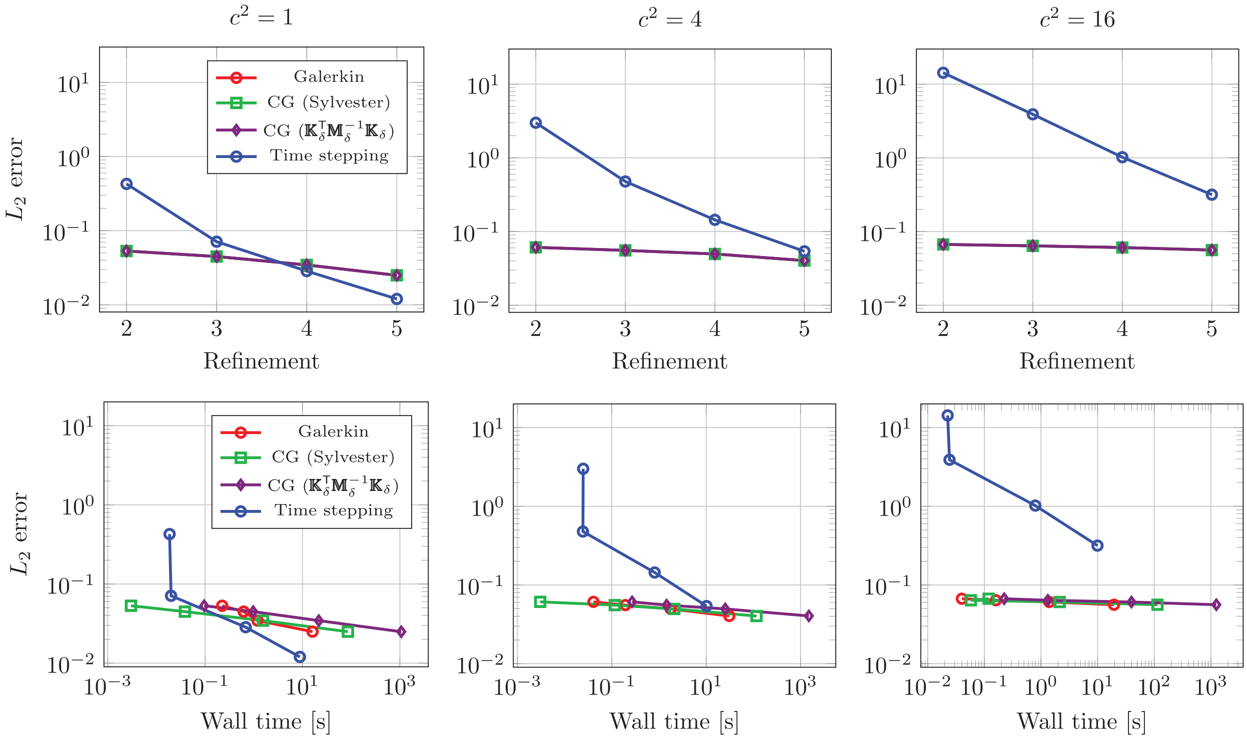


FIGURE 5. Case 1: Continuous solution. L_2-error over refinement (*top row*) and wall time (*bottom row*) for the cases $c^2 = 1, 4, 16$ (*left, center, right*).

from the CFL-stability condition. On the other hand, the rate of convergence of the time-stepping method is clearly better than the low-order space-time method using discontinuous ansatz functions. In order to get a convergence order of the space-time method comparable to the rate of the Crank–Nicolson scheme, we would at least need to use quartic test functions.

6.2. Case 2: Discontinuous solution

For the case of a discontinuous solution, our results are shown in Figure 6. The number of iterations for the PCG and the Galerkin methods behave similar as in case 1, so that we do not monitor all numbers. Again, we observe that the performance of the ultraweak space-time method is basically independent of the wave speed. Moreover, due to the fact that the solution is discontinuous, the rate of convergence of the time stepping scheme is no longer optimal. As a consequence, the ultraweak space-time method outperforms the time stepping approach w.r.t. accuracy and efficiency. The benefit is even larger for increasing wave speed numbers.

7. Conclusions

Our theoretical results and numerical experience show that the proposed ultraweak variational space-time method, when equipped with appropriate linear algebra solvers, is significantly more accurate and efficient than the Crank–Nicolson scheme on problems with low regularity and high wave speed.

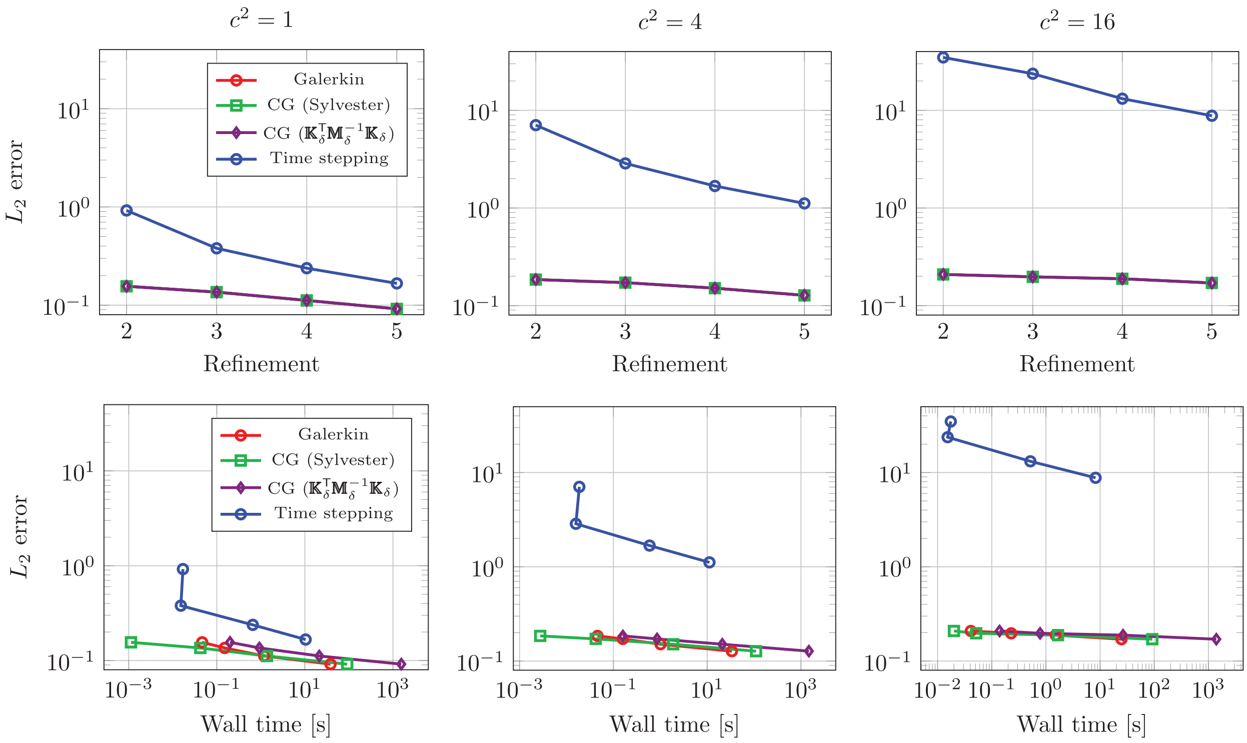


FIGURE 6. Case 2: Discontinuous solution. L_2-error over refinement (*top row*) and wall time (*bottom row*) for the cases $c^2 = 1, 4, 16$ (*left, center, right*).

APPENDIX A. PROOF OF THEOREM 2.2

We collect the proof of the well-posedness for the semi-variational setting in Section 2.2. Even though the proofs are based upon rather classical spectral theory, we could not find them in the desired form in the literature.

Proposition A.1. *Let $s \in \mathbb{R}^+$. The mapping $w \mapsto \langle \cdot, w\rangle$, $w \in H^{-s}$, where*

$$\langle \cdot, \cdot \rangle : H^s \times H^{-s} \to \mathbb{R}, \quad \langle v, w\rangle := \sum_{n=1}^{\infty} v_n \, w_n \tag{A.1}$$

is an isometric isomorphism from H^{-s} to $(H^s)'$, i.e., $(H^s)' \cong H^{-s}$.

Proof. First, note that H^s is a Hilbert space with the inner product $(v, w)_s := \sum_{n=1}^{\infty} \lambda_n^s \, v_n \, w_n$ and $H^s \hookrightarrow H \hookrightarrow H^{-s}$ with continuous embeddings. Let $v \in H^s$, $w \in H^{-s}$, then by Hölder's inequality

$$\langle v, w\rangle = \sum_{n=1}^{\infty} \lambda_n^{s/2} v_n \, \lambda_n^{-s/2} w_n \leq \left(\sum_{n=1}^{\infty} \lambda_n^s \, v_n^2\right)^{1/2} \left(\sum_{n=1}^{\infty} \lambda_n^{-s} \, w_n^2\right)^{1/2} = \|v\|_s \, \|w\|_{-s} < \infty.$$

Hence, $\langle \cdot, w\rangle \in (H^s)'$ and $\|\langle \cdot, w\rangle\|_{(H^s)'} = \sup_{v \in H^s} \frac{\langle v, w\rangle}{\|v\|_s} \leq \|w\|_{-s}$. On the other hand, given $w \in H^{-s}$, set $\tilde{v}_n := \lambda^{-s} w_n$ and $\tilde{v} := \sum_{n=1}^{\infty} \tilde{v}_n e_n$. Then, $\|\tilde{v}\|_s^2 = \sum_{n=1}^{\infty} \lambda_n^s \, (\lambda_n^{-s} w_n)^2 = \sum_{n=1}^{\infty} \lambda_n^{-s} (w_n)^2 = \|w\|_{-s}^2 < \infty$, *i.e.*, $\tilde{v} \in H^{-s}$. Moreover $\langle \tilde{v}, w\rangle = \sum_{n=1}^{\infty} \tilde{v}_n \, w_n = \sum_{n=1}^{\infty} \lambda_n^{-s} (w_n)^2 = \|w\|_{-s}^2 = \|\tilde{v}\|_s \, \|w\|_{-s}$. If $w \neq 0$, we get that $\|\langle \cdot, w\rangle\|_{(H^s)'} = \sup_{v \in H^s} \frac{\langle v, w\rangle}{\|v\|_s} \geq \frac{\langle \tilde{v}, w\rangle}{\|\tilde{v}\|_s} = \|w\|_{-s}$ with equality for $w = 0$. Hence, $\|\langle \cdot, w\rangle\|_{(H^s)'} = \|w\|_{-s}$ for all $w \in H^{-s}$. □

Now we start by considering the following *homogeneous* abstract second order initial value problem. Let $u_0 \in D(A)$ and $u_1 \in H$. The goal is to find a function $w \in C^2([0,T], H)$ such that $w(t) \in D(A)$ for $t \in [0,T]$ and satisfying

$$\ddot{w}(t) + Aw(t) = 0, \quad t \in (0,T), \qquad w(0) = u_0,\ \dot{w}(0) = u_1, \tag{A.2}$$

where the spaces u_0 and u_1 reside in will be specified later. It is easily seen that (a) $u_0 = e_n$, $u_1 = 0$ yields $w(t) = \cos\big(\sqrt{\lambda_n} t\big) e_n$ and (b) $u_0 = 0$ and $u_1 = e_n$ gives rise to $w(t) = \lambda_n^{-1/2} \sin\big(\sqrt{\lambda_n} t\big) e_n$.

We can now express the general solution of (A.2) as a series of solutions of these special types and prove the following theorem.

Theorem A.2 (Homogeneous wave equation). *Let $s \in \mathbb{R}_{\geq 0}$, $u_0 \in H^s$ and $u_1 \in H^{s-1}$. Then* (A.2) *admits a unique solution $w \in \mathcal{C}^s$, see* (2.9).

Proof. Uniqueness: Let $w \in \mathcal{C}^s$ be a solution of (A.2), then $w(t) \in H$ for all $t \in [0,T]$. Set $w_n(t) := \langle w(t), e_n \rangle = (w(t), e_n)_H$ for $n \in \mathbb{N}$ and $t \in [0,T]$. Since $w \in \mathcal{C}^s$, in particular $\ddot{w}(t) \in H^{s-2}$, we get by $e_n \in D(A) = H^2$ the fact $w_n \in C^2([0,T])$ with derivative $\ddot{w}_n(t) = \langle \ddot{w}(t), e_n \rangle = -\langle Aw(t), e_n \rangle = -\sum_{k=1}^\infty \lambda_k (w(t), e_k)_H \langle e_k, e_n \rangle = -\lambda_n (w(t), e_n)_H = -\lambda_n w_n(t)$, $t \in (0,T)$, and initial values $w_n(0) = (w(0), e_n)_H = (u_0, e_n)_H$, $\dot{w}_n(0) = \langle \dot{w}(0), e_n \rangle = \langle u_1, e_n \rangle$. This is an initial value problem of a second order linear ode with the unique solution

$$w_n(t) = \cos\Big(\sqrt{\lambda_n t}\Big)(u_0, e_n)_H + \lambda_n^{-1/2} \sin\Big(\sqrt{\lambda_n t}\Big)\langle u_1, e_n \rangle, \tag{A.3}$$

which is easily verified. Since

$$w(t) = \sum_{n=1}^\infty w_n(t) e_n \tag{A.4}$$

is the unique expansion of $w(t)$ in H with respect to the orthonormal basis $\{e_n : n \in \mathbb{N}\}$, the uniqueness statement has been proved.

Existence: we now define $w_n(t)$ by (A.3) and (A.4). Then, for all $t \in [0,T]$,

$$\begin{aligned}
\|w(t)\|_s^2 &\leq 2 \sum_{n=1}^\infty \lambda_n^s \left|\cos\Big(t\sqrt{\lambda_n}\Big)\right|^2 |(u_0, e_n)_H|^2 + 2 \sum_{n=1}^\infty \lambda_n^{s-1} \left|\sin\Big(t\sqrt{\lambda_n}\Big)\right|^2 |\langle u_1, e_n \rangle|^2 \\
&\leq 2 \sum_{n=1}^\infty \lambda_n^s |(u_0, e_n)_H|^2 + 2 \sum_{n=1}^\infty \lambda_n^{s-1} |\langle u_1, e_n \rangle|^2 = 2\|u_0\|_s^2 + 2\|u_1\|_{s-1}^2 < \infty,
\end{aligned}$$

uniformly in $t \in [0,T]$, so that $w \in C([0,T]; H^s)$. Next

$$\begin{aligned}
\|\dot{w}(t)\|_{s-1}^2 &\leq 2 \sum_{n=1}^\infty \lambda_n^{s-1}\, \lambda_n \left|\sin\Big(t\sqrt{\lambda_n}\Big)\right|^2 |(u_0, e_n)_H|^2 + 2 \sum_{n=1}^\infty \lambda_n^{s-1}\, \lambda_n^{-1} \lambda_n \left|\cos\Big(t\sqrt{\lambda_n}\Big)\right|^2 |\langle u_1, e_n \rangle|^2 \\
&\leq 2 \sum_{n=1}^\infty \lambda_n^s |(u_0, e_n)_H|^2 + 2 \sum_{n=1}^\infty \lambda_n^{s-1} |\langle u_1, e_n \rangle|^2 = 2\|u_0\|_s^2 + 2\|u_1\|_{s-1}^2 < \infty,
\end{aligned}$$

so that $w \in C^1\big([0,T]; H^{s-1}\big)$ and similarly

$$\begin{aligned}
\|\ddot{w}(t)\|_{s-2}^2 &\leq 2 \sum_{n=1}^\infty \lambda_n^{s-2}\, \lambda_n^2 \left|\cos\Big(t\sqrt{\lambda_n}\Big)\right|^2 |(u_0, e_n)_H|^2 + 2 \sum_{n=1}^\infty \lambda_n^{s-2}\, \lambda_n^{-1} \lambda_n^2 \left|\sin\Big(t\sqrt{\lambda_n}\Big)\right|^2 |\langle u_1, e_n \rangle|^2 \\
&\leq 2 \sum_{n=1}^\infty \lambda_n^s |(u_0, e_n)_H|^2 + 2 \sum_{n=1}^\infty \lambda_n^{s-1} |\langle u_1, e_n \rangle|^2 = 2\|u_0\|_s^2 + 2\|u_1\|_{s-1}^2 < \infty,
\end{aligned}$$

which shows that $w \in C^2\big([0,T];H^{s-2}\big)$. We conclude that $w \in \mathcal{C}^s$. Finally, we have $\ddot{w}(t) = \sum_{n=1}^\infty \ddot{w}_n(t)e_n = \sum_{n=1}^\infty w_n(t)\lambda_n e_n = -Aw(t)$ by definition of A. In addition, $w(0) = \sum_{n=1}^\infty (u_0, e_n)_H e_n = u_0$ and $\dot{w}(0) = \sum_{n=1}^\infty \langle u_1, e_n\rangle\, e_n = u_1$. This shows that w solves (A.2), and we have proved existence of solutions. □

We are now in the position to prove Theorem 2.2 for the wave equation with inhomogeneous right-hand side.

Proof of Theorem 2.2. Since the difference of two solutions of (2.8) is a solution of the homogeneous problem (A.2), uniqueness follows from Theorem A.2. Moreover, since the homogeneous problem has a solution, in order to prove existence for (2.8), we may and will assume that $u_0 = u_1 = 0$.

Next, we set $f_n(t) := \langle f(t), e_n\rangle$, which is well-defined since $e_n \in D(A) = H^2$ and $f(t) \in H^{s-1}$, $s \geq 0$. Then, $f_n \in C([0,T])$. We set $w_n(t) = \lambda_n^{-1/2} \int_0^t \sin\big(\sqrt{\lambda_n}(t-\tau)\big)\, f_n(\tau)\,\mathrm{d}\tau$, and $w(t) := \sum_{n=1}^\infty w_n(t)\, e_n$. By Hölder's inequality, we have, for all $t \in [0,T]$

$$\lambda_n^s\, w_n(t)^2 \leq \lambda_n^{s-1} \int_0^T \sin\Big(\sqrt{\lambda_n}(t-\tau)\Big)^2 \mathrm{d}\tau \int_0^T f_n(\tau)^2\,\mathrm{d}\tau \leq T\lambda_n^{s-1} \int_0^T f_n(\tau)^2\,\mathrm{d}\tau,$$

so that

$$\|w(t)\|_s^2 \leq T\sum_{n=1}^\infty \lambda_n^{s-1}\int_0^T f_n(\tau)^2\,\mathrm{d}\tau = T\int_0^T \sum_{n=1}^\infty \lambda_n^{s-1} f_n(\tau)^2\,\mathrm{d}\tau = T\int_0^T \|f(\tau)\|_{s-1}^2\,\mathrm{d}\tau,$$

which is finite uniformly in $t \in [0,T]$ since $f \in C\big([0,T];H^{s-1}\big)$, so that $w \in C([0,T];H^s)$. Next, we note that $\dot{w}_n(t) = \int_0^t \cos\big(\sqrt{\lambda_n}(t-s)\big)\, f_n(s)\,\mathrm{d}s$, so that similar as above

$$\lambda_n^{s-1}\, w_n(t)^2 \leq \lambda_n^{s-1} \int_0^T \cos\Big(\sqrt{\lambda_n}(t-\tau)\Big)^2 \mathrm{d}\tau \int_0^T f_n(\tau)^2\,\mathrm{d}\tau \leq T\lambda_n^{s-1} \int_0^T f_n(\tau)^2\,\mathrm{d}\tau,$$

which yields

$$\|\dot{w}(t)\|_{s-1}^2 \leq T\sum_{n=1}^\infty \lambda_n^{s-1}\int_0^T f_n(\tau)^2\,\mathrm{d}\tau = T\int_0^T \sum_{n=1}^\infty \lambda_n^{s-1} f_n(\tau)^2\,\mathrm{d}\tau = T\int_0^T \|f(\tau)\|_{s-1}^2\,\mathrm{d}\tau,$$

which again is finite uniformly in $t \in [0,T]$, so that $w \in C^1\big([0,T];H^{s-1}\big)$. In order to prove $w \in C^2\big([0,T];H^{s-2}\big)$ (and thus $w \in \mathcal{C}^s$), we note that $\ddot{w}_n + \lambda_n\, w_n = f_n$, $w_n(0) = \dot{w}_n(0) = 0$. Hence,

$$\|\ddot{w}(t)\|_{s-2}^2 \leq 2\sum_{n=1}^\infty \lambda_n^{s-2}\lambda_n^2 w_n(t)^2 + 2\sum_{n=1}^\infty \lambda_n^{s-2} f_n(t)^2 = 2\,\|w(t)\|_s^2 + 2\,\|f(t)\|_{s-2}^2 < \infty$$

uniformly in $t \in [0,T]$, so that $w \in C^2\big([0,T];H^{s-2}\big)$. Finally

$$\begin{aligned}\ddot{w}(t) &= \sum_{n=1}^\infty \langle \ddot{w}_n(t), e_n\rangle\, e_n = -\sum_{n=1}^\infty \lambda_n\,\langle w_n(t), e_n\rangle\, e_n + \sum_{n=1}^\infty \langle f_n(t), e_n\rangle\, e_n = -\sum_{n=1}^\infty \lambda_n\,(w_n(t), e_n)_H\, e_n + f(t)\\ &= -Aw(t) + f(t)\end{aligned}$$

for all $t \in (0,T)$. Since $w_n(0) = 0 = \dot{w}_n(0)$, we obtain $w(0) = \dot{w}(0) = 0$, so that w solves (2.8) for $u_0 = u_1 = 0$, which concludes the proof. □

Acknowledgements. The authors are grateful to Wolfgang Arendt (Ulm University) for various inspiring discussions concerning the analytical aspects. Furthermore, the authors acknowledge support by the state of Baden-Württemberg through bwHPC. The second and third authors are members of the Italian research group Indam-GNCS, whose support is gratefully acknowledged. Part of this work was carried out while the second author was affiliated with the Max Planck Institute for Dynamics of Complex Technical Systems in Magdeburg, Germany.

References

[1] R. Andreev, Stability of sparse space-time finite element discretizations of linear parabolic evolution equations. *IMA J. Numer. Anal.* **33** (2013) 242–260.

[2] T. Apel, S. Nicaise and J. Pfefferer, Discretization of the Poisson equation with non-smooth data and emphasis on non-convex domains. *Numer. Meth. Part. Diff. Equ.* **32** (2016) 1433–1454.

[3] W. Arendt and K. Urban, Partial Differential Equations: An Analytic and Numerical Approach. Springer (2022) to appear. Translated from the German by J.B. Kennedy.

[4] I. Babuška, J. Osborn and J. Pitkäranta, Analysis of mixed methods using mesh dependent norms. *Math. Comput.* **35** (1980) 1039–1062.

[5] L. Bales and I. Lasiecka, Continuous finite elements in space and time for the nonhomogeneous wave equation. *Comput. Math. Appl.* **27** (1994) 91–102.

[6] L. Bales and I. Lasiecka, Negative norm estimates for fully discrete finite element approximations to the wave equation with nonhomogeneous L_2 Dirichlet boundary data. *Math. Comput.* **64** (1995) 89–115.

[7] R.H. Bartels and G.W. Stewart, Algorithm 432: solution of the matrix equation $AX + XB = C$. *Comm. ACM* **15** (1972) 820–826.

[8] M. Berggren, Approximations of very weak solutions to boundary-value problems. *SIAM J. Numer. Anal.* **42** (2004) 860–877.

[9] J. Brunken, K. Smetana and K. Urban, (Parametrized) First order transport equations: realization of optimally stable Petrov–Galerkin methods. *SIAM J. Sci. Comput.* **41** (2019) A592–A621.

[10] T. Bui-Thanh, L. Demkowicz and O. Ghattas, Constructively well-posed approximation methods with unity inf-sup and continuity constants for partial differential equations. *Math. Comput.* **82** (2013) 1923–1952.

[11] W. Dahmen, C. Huang, C. Schwab and G. Welper, Adaptive Petrov–Galerkin methods for first order transport equations. *SIAM J. Numer. Anal.* **50** (2012) 2420–2445.

[12] L. Demkowicz and J. Gopalakrishnan, A class of discontinuous Petrov–Galerkin methods. II. Optimal test functions. *Numer. Meth. Part. Diff. Equ.* **27** (2011) 70–105.

[13] L. Demkowicz, J. Gopalakrishnan, S. Nagaraj and P. Sepúlveda, A spacetime DPG method for the Schrödinger equation. *SIAM J. Numer. Anal.* **55** (2017) 1740–1759.

[14] V. Druskin and V. Simoncini, Adaptive rational Krylov subspaces for large-scale dynamical systems. *Syst. Control Lett.* **60** (2011) 546–560.

[15] T. Ellis, J. Chan and L. Demkowicz, Robust DPG methods for transient convection-diffusion. In: Building Bridges: Connections and Challenges in Modern Approaches to Numerical Partial Differential Equations. Vol. 114. Springer (2016) 179–203.

[16] J. Ernesti and C. Wieners, Space-time discontinuous Petrov–Galerkin methods for linear wave equations in heterogeneous media. *Comput. Methods Appl. Math.* **19** (2019) 465–481.

[17] G. Golub and C.F. Van Loan, Matrix Computations, 4th edition. The Johns Hopkins University Press (2013).

[18] B. Haasdonk, Reduced Basis Methods for Parametrized PDEs – a tutorial. In: Model Reduction and Approximation, edited by P. Benner, A. Cohen, M. Ohlberger and K. Willcox. Chapter 2. SIAM (2017) 65–136.

[19] J. Henning, D. Palitta, V. Simoncini and K. Urban, Matrix oriented reduction of space-time Petrov–Galerkin variational problems. In: Numerical Mathematics and Advanced Applications ENUMATH 2019, edited by F.J. Vermolen and C. Vuik. Vol. 139 of *Lect. Notes Comput. Sci. Eng.* Springer (2021) 1049–1057.

[20] J.S. Hesthaven, G. Rozza and B. Stamm, Certified Reduced Basis Methods for Parametrized Partial Differential Equations. Springer (2016).

[21] B. Keith, A priori error analysis of high-order LL* (FOSLL*) finite element methods. *Comput. Math. Appl.* **103** (2021) 12–18.

[22] D. Kressner and C. Tobler, Low-rank tensor Krylov subspace methods for parametrized linear systems. *SIAM. J. Matrix Anal. Appl.* **32** (2011) 1288–1316.

[23] J.-L. Lions and E. Magenes, Non-homogeneous Boundary Value Problems and Applications. Vol. I. Springer (1972). Translated from the French by P. Kenneth.

[24] S. May, R. Rannacher and B. Vexler, Error analysis for a finite element approximation of elliptic Dirichlet boundary control problems. *SIAM J. Control Optim.* **51** (2013) 2585–2611.

[25] C. Mollet, *Parabolic PDEs in space-time formulations: stability for Petrov–Galerkin discretizations with B-splines and existence of moments for problems with random coefficients*. Ph.D. thesis, Universität zu Köln (2016).

[26] R.H. Nochetto, K.G. Siebert and A. Veeser, Theory of adaptive finite element methods: an introduction. In: Multiscale, Nonlinear and Adaptive Approximation, edited by R.A. DeVore and A. Kunoth. Springer (2009) 409–542.

[27] D. Palitta, Matrix equation techniques for certain evolutionary partial differential equations. *J. Sci. Comput.* **87** (2021) 1–36.

[28] C.E. Powell, D. Silvester and V. Simoncini, An efficient reduced basis solver for stochastic Galerkin matrix equations. *SIAM J. Sci. Comput.* **39** (2017) A141–A163.

[29] A. Quarteroni, A. Manzoni and F. Negri, Reduced Basis Methods for Partial Differential Equations: An Introduction. Springer (2016).

[30] D. Silvester and M. Mihajlović, A black-box multigrid preconditioner for the biharmonic equation. *BIT* **44** (2004) 151–163.

[31] V. Simoncini, Computational methods for linear matrix equations. *SIAM Rev.* **58** (2016) 377–441.

[32] O. Steinbach and M. Zank, A generalized inf-sup stable variational formulation for the wave equation. *J. Math. Anal. Appl.* **505** (2022) 24.

[33] K. Urban and A.T. Patera, A new error bound for reduced basis approximation of parabolic partial differential equations. *C.R. Math. Acad. Sci. Paris* **350** (2012) 203–207.

[34] K. Urban and A.T. Patera, An improved error bound for reduced basis approximation of linear parabolic problems. *Math. Comput.* **83** (2014) 1599–1615.

[35] J. Xu and L. Zikatanov, Some observations on Babuška and Brezzi theories. *Numer. Math.* **94** (2003) 195–202.

[36] M. Zank, The Newmark method and a space-time FEM for the second-order wave equation. In: Numerical Mathematics and Advanced Applications ENUMATH 2019, edited by F.J. Vermolen and C. Vuik. Vol. 139 of *Lect. Notes Comput. Sci. Eng.* Springer (2021) 1225–1233.