NUMERICAL SOLUTION OF PARABOLIC PROBLEMS BASED ON A WEAK SPACE-TIME FORMULATION

STIG LARSSON AND MATTEO MOLTENI

Abstract. We investigate a weak space-time formulation of the heat equation and its use for the construction of a numerical scheme. The formulation is based on a known weak space-time formulation, with the difference that a pointwise component of the solution, which in other works is usually neglected, is now kept. We investigate the role of such a component by first using it to obtain a pointwise bound on the solution and then deploying it to construct a numerical scheme. The scheme obtained, besides being quasi-optimal in the L^2 sense, is also pointwise superconvergent in the temporal nodes. We prove *a priori* error estimates and we present numerical experiments to empirically support our findings.

1. Introduction

In this article we study a numerical scheme to solve linear parabolic problems, based on a weak space-time formulation. The equation we consider, in its strong form and under assumptions that we specify in Section 2, is

$$\begin{aligned} &\dot{u}(t) + Au(t) = f(t), \quad t \in (0,T], \\ &u(0) = u_0. \end{aligned} \tag{1.1}$$

During the last decades several authors have dealt with the space-time formulation of this problem. The main idea of a space-time formulation is to integrate the equation (1.1) in both the spatial and the temporal dimensions after multiplying it by a suitable space and time dependent test function. By doing the same with the initial condition, and by adding up the equations, we achieve the first space-time formulation of the problem, also called the primal formulation in other articles.

By means of a formal integration by parts of the term containing the time derivative, we achieve the weak space-time formulation of the problem, sometimes also called the second formulation (see [And13], [And16], [Mol13], [SS13], [CSt11]) or natural formulation (see [Tan13]). For both formulations, the main tool to prove the existence and uniqueness of the solution is the Banach–Nečas–Babuška theorem, see Theorem 1 below.

Although such a theory was originally used to deal with mixed formulations of elliptic problems, from the late eighties it has also been used in connection with parabolic problems. A first analysis of numerics for evolution problems based on space-time formulations can be found in [BJ89, BJ90].

Date: March 9, 2016.

2010 *Mathematics Subject Classification.* 65M15, 65M60.

Key words and phrases. inf-sup, space-time, superconvergence, quasi-optimality, finite element, error estimate, Petrov–Galerkin.

In [SSt09], a discretization of evolution equations based on the primal formulation of the problem is discussed. The problem is restated as a bi-infinite matrix problem and discretized by an adaptive wavelet method. The proof of well-posedness of the abstract problem presented in the appendix of this article is of great relevance, since many other articles explicitly refer to it. In [SS13] the second space-time formulation is deployed to construct adaptive numerical schemes; this choice allows the authors to apply the theory presented in previous paper to parabolic PDE's in infinite dimensions, where the solution is in general not regular enough to allow the use of the first space-time formulation

In [CSt11], the second space-time formulation is used to further investigate what was studied in [SSt09], under the extra assumption that the bi-infinite matrix system is truly sparse.

In [And12, And13] the stability of space-time Petrov–Galerkin discretizations of the problem is studied for both the first and the second formulations. A possible selection of stable space-time trial and test spaces is presented, and a CFL condition is derived. Such a condition is shown to be necessary when trial and test spaces are chosen to be piecewise polynomials. In [And16] the author proposes a Petrov–Galerkin space-time discretization of the heat equation on an unbounded time interval by means of Laguerre polynomials. Both the first and the second space-time formulations are investigated.

In [Mol13] the author considers suitable hierarchical families of discrete spaces, both of finite element and wavelet type, and investigates the required number of extra layers in order to guarantee uniform boundedness of the discrete inf-sup constant in the second space-time formulation.

In [UP14], the second space-time formulation is used as a natural framework in which the reduced basis method can be investigated, allowing the authors to derive sharp a posteriori error bounds.

However, in all the works on the second space-time formulation, the authors choose to neglect a term that naturally arises from the integration by parts. This is achieved by using test functions which vanish at the final time instant. Although this is justified because the neglected term is a pointwise version of the term which is kept, the neglected term can play an important role, as noticed, for example, in [LM16], where the second space-time formulation is used to study a stochastic variant of (1.1).

By keeping such a term in the current paper, not only do we have a framwork for stochastic evolution equations, but we also obtain estimates in the $L^\infty((0,T);H)$-norm in addition to the natural $L^2((0,T);V)$-norm and we can construct a numerical scheme that is superconvergent at the temporal mesh points.

The paper is structured as follows. In Section 2 we present the abstract framework for the weak space-time formulation based on the Banach–Necǎs–Babuška "inf-sup" theorem. Section 3 introduces the Petrov–Galerkin approximation based on piecewise polynomials in space and time. The trial functions are discontinuous of degree $q \geq 0$ in time while the test functions are continuous of degree $q+1$. The possibility of extracting point values at the temporal nodes is emphasized. Section 4 is devoted to the *a priori* error estimates based on quasi-optimality. A CFL condition is required. The temporal order in the natural norm is $q+1$. However, we note that the piecewise constant approximation ($q=0$) is of second order in time by a comparison with the Crank–Nicolson method. In Section 5 we give a direct proof of this by showing that our method is actually superconvergent of order $2(q+1)$ at the temporal nodes. The proof is based on separating the temporal and spatial error and a duality argument. We only present the analysis of the temporally semidiscrete part. The proof avoids the use of a CFL condition, which is not available for pure time discretizations. The temporal convergence rates are demonstrated in numerical experiments in Section 6.

2. The abstract problem

2.1. **An abstract framework.** We assume that a Gelfand triple $V \hookrightarrow H \hookrightarrow V^*$ is given, where V and H are separable Hilbert spaces such that V is densely embedded into H. We assume that the operator A, which appears in (1.1), is associated to a symmetric bilinear form $a(\cdot,\cdot)$ that satisfies the following conditions:

$$\text{(boundedness)} \qquad |a(u,v)| \le A_{\max}\|u\|_V\|v\|_V, \quad u,v\in V,$$

$$\text{(coercivity)} \qquad a(v,v) \ge A_{\min}\|v\|_V^2, \quad v\in V,$$

for some positive constants $A_{\max}$ and $A_{\min}$. We introduce the Lebesgue-Bochner spaces

$$\mathcal{Y}^t = L^2((0,t);V), \quad \mathcal{X}^t = L^2((0,t);V)\cap H^1((0,t);V^*),$$

with norms defined by

$$\|y\|_{\mathcal{Y}^t}^2 := \|y\|_{L^2((0,t);V)}^2 = \int_0^t \|y(s)\|_V^2\,\mathrm{d}s,$$
$$\|x\|_{\mathcal{X}^t}^2 := \|x(0)\|_H^2 + \|x\|_{L^2((0,t);V)}^2 + \|\dot{x}\|_{L^2((0,t);V^*)}.$$

We use the notation $\mathcal{Y}_H^t$ for the space $\mathcal{Y}^t\times H$ endowed with the product norm, and we use the convention that $\mathcal{Y}=\mathcal{Y}^T$, $\mathcal{Y}_H=\mathcal{Y}\times H$, and $\mathcal{X}=\mathcal{X}^T$, when $t=T$. We recall that the space $\mathcal{X}^t$ is densely embedded in $\mathscr{C}([0,t];H)$, So that pointwise values of $x\in\mathcal{X}$ make sense. With the present choice of norm the embedding constant is 1; in particular, it does not depend on t or V, see [LM16].

The first space-time formulation of (1.1) reads:

$$u\in\mathcal{X} : \mathscr{B}(u,y)=\mathscr{F}(y), \quad \forall y=(y_1,y_2)\in\mathcal{Y}_H. \tag{2.1}$$

Here we use the bilinear form:

$$\mathscr{B}\colon \mathcal{X}\times\mathcal{Y}_H\to\mathbb{R},$$
$$\mathscr{B}(x,y) := \int_0^T {}_{V^*}\langle \dot{x}+Ax, y_1\rangle_V\,\mathrm{d}s + \langle x(0), y_2\rangle_H,$$

and the load functional

$$\mathscr{F}\in\mathcal{Y}_H^*, \quad \mathscr{F}(y) := \int_0^T {}_V\langle f, y_1\rangle_{V^*}\,\mathrm{d}t + \langle u_0, y_2\rangle_H.$$

If we integrate by parts and swap the test and trial spaces, then we obtain the weak (or second) space-time formulation

$$u=(u_1,u_2)\in\mathcal{Y}_H : \mathscr{B}^*(u,x)=\mathscr{F}(x), \quad \forall x\in\mathcal{X}, \tag{2.2}$$

where the bilinear form and the load functional are now:

$$\mathscr{B}^*\colon \mathcal{Y}_H\times\mathcal{X}\to\mathbb{R},$$
$$\mathscr{B}^*(y,x) := \int_0^T {}_V\langle y_1, -\dot{x}+Ax\rangle_{V^*}\,\mathrm{d}s + \langle y_2, x(T)\rangle_H, \tag{2.3}$$

$$\mathscr{F} \in \mathcal{X}^*, \quad \mathscr{F}(x) := \int_0^T {}_V\langle f, x\rangle_{V^*} \, \mathrm{d}t + \langle u_0, x(0)\rangle_H. \tag{2.4}$$

It is easy to see that the second component u_2 of the solution u to (2.2) depends on the final time instant T. We can think of parametrizing (2.2) over $t \in [0,T]$ and reformulate it as a family of problems:

$$u = (u_1, u_2(t)) \in \mathcal{Y}_H^t : \mathscr{B}_t^*(u,x) = \mathscr{F}_t(x), \quad \forall x \in \mathcal{X}^t, \tag{2.5}$$

where $\mathscr{F}_t$ and $\mathscr{B}_t^*$ are as before, but restricted to the spaces $\mathcal{Y}_H^t$ and $\mathcal{X}^t$.

If the right-hand side of (1.1) is regular enough, as in § 2.2.1 below, then u_1 has a square integrable weak derivative and therefore belongs to the space $\mathcal{X} \subset \mathscr{C}([0,T];H)$, and $u_1 = u_2$. However, if the right-hand side is less regular, as in § 2.2.3 and § 2.2.4, then u_1 need not be differentiable nor continuous, but u_2 is a continuous time-dependent H-valued version of u_1:

$$\int_0^T \|u_1(t) - u_2(t)\|_H^2 \, \mathrm{d}t = 0.$$

The second component u_2 is often omitted in other works (e.g., [SS13], [Mol13]), where the following weak space-time formulation is used:

$$u \in \mathcal{Y} : \mathscr{B}^*(u,x) = \mathscr{F}(x), \quad \forall x \in \mathcal{X}_{0,\{T\}} := \{x \in \mathcal{X} : x(T) = 0\}.$$

We keep u_2 in order to be able to extract point values.

In order to appreciate the weak space-time formulation, we briefly recall the two main advantages that we want to exploit: larger variety of source terms and the possibility to obtain pointwise bounds.

2.2. **A larger variety of right-hand sides.** First of all, the weak-space time formulation allows the use of a broad family of possible source terms.

2.2.1. *Regular right-hand side.* The basic case that we analyse is given by

$$\mathscr{F}_t(x) = \int_0^t {}_{V^*}\langle f(s), x(s)\rangle_V \, \mathrm{d}s + \langle u_0, x(0)\rangle_H, \quad t \in [0,T], \tag{2.6}$$

for some $f \in L^2((0,T);V^*)$ and $u_0 \in H$. In this case, we have $u_2 = u_1 \in \mathcal{X}$. Indeed, by taking $x \in \mathscr{C}_0^\infty([0,t];V)$ in (2.5), we obtain

$$\int_0^t \langle u_1(s), -\dot{x}(s)\rangle_H \, \mathrm{d}s = \int_0^t {}_{V^*}\langle f(s) - Au_1(s), x(s)\rangle_V \, \mathrm{d}s.$$

Thus, u_1 has a weak derivative $\dot{u}_1 = f - Au_1 \in L^2((0,T);V^*)$, so that $u_1 \in \mathcal{X}$. Then we can integrate by parts in (2.5) and conclude that $u_2 = u_1$ and that they both belong to $\mathcal{X} \subset \mathscr{C}([0,T];H)$.

2.2.2. *Piecewise regular right-hand side.* A more general case is offered by

$$\mathscr{F}_t(x) = \int_0^t {}_{V^*}\langle f(s), x(s)\rangle_V \, \mathrm{d}s + \sum_{t_i \le t} \langle \zeta_i, x(t_i)\rangle_H, \quad t \in [0,T],$$

for some $f \in L^2((0,T);V^*)$, $\{\zeta_i\}_{i=1,\ldots,M} \in H$ and $\{t_i\}_{i=1,\ldots,M} \subset [0,T]$.

In this case the conclusions presented above only hold piecewise. In particular, the values of ζ_i represent the jumps of the solution at time t_i.

2.2.3. *Stochastic integral.* A more general example is represented by a functional which is defined ω-wise, for ω in a complete probability space $(\Omega, \Sigma_t, \mathbb{P})$, and of the form $\mathscr{F}_t + \mathscr{W}_t$. Here $\mathscr{F}_t$ is as in § 2.2.1 and $\mathscr{W}_t$ is a weak stochastic integral with respect to an H-valued Wiener process W, with operator-valued integrand Ψ:

$$\mathscr{W}_t(x) = \int_0^t \langle \Psi(s)\, \mathrm{d}W(s), x(s)\rangle_H, \quad t \in [0,T]. \tag{2.7}$$

The details of such an equation have been presented in [LM16] and we refrain from recalling them here. It holds that u_1 and u_2 are versions of each other, in the sense that $u_1 \in L^2((0,T);V)$, $u_2 \in \mathscr{C}([0,T];H)$ almost surely and $u_1 = u_2$ in $L^2(\Omega \times (0,T);H)$. This case represents an important example in which the weak space-time formulation cannot be replaced by the first space-time formulation, since the Wiener process is nowhere differentiable and therefore $u_1 \notin \mathcal{X}$.

2.2.4. *Nowhere differentiable right-hand side.* The most general type of right-hand side that we can handle has the form

$$\mathscr{F}_t(x) = \int_0^t {}_V\langle g(s), -\dot{x}(s)\rangle_{V^*}\, \mathrm{d}s - \langle g(t), x(t)\rangle_H + \langle g(0), x(0)\rangle_H, \quad t \in [0,T], \tag{2.8}$$

for a function $g \in L^2((0,T);V) \cap \mathscr{C}([0,T];H)$, and with g nowhere differentiable, so that we are not in one of the first two cases in this list.

Similar conclusions to the ones obtained for the stochastic integral hold even in this case. We have that $u_1 \notin \mathcal{X}$, that $u_1 = u_2$ in $L^2((0,T);H)$, and that $u_2 \in \mathscr{C}([0,T];H)$. In case g is smooth it is easy to see that integration by parts leads to a right-hand side of the same form as in (2.6).

We want to stress that both in the case of a right-hand side of the form (2.7) or (2.8) the presence of u_2 is important, since point values $u_1(t)$ of u_1 are not well defined.

2.3. **Point values and decompositions.** Another important advantage offered by the weak formulation is that the solution is not required to be continuous in its first component u_1. This allows us to split the time interval and to solve local problems, where information is passed from one time interval to the next through $u_2(t)$, see (2.15) and § 3.2 below. This can be exploited even on a discrete level, by solving problems with different spatial discretizations on each time interval, since the passage of information between two different intervals occurs only by means of the second component of the solution, u_2. This ensures a flexibility in the choice of the spatial grid, which could in principle change at each interval and still not cause any sort of variational crime, since the discrete spaces would still be proper subspaces of the continuous ones.

2.4. **The inf-sup theorem.** We recall the following theorem (see [BA72, EG04]):

Theorem 1 (Banach–Nečas–Babuška (BNB))**.** *Let V and W be Hilbert spaces. Given a bilinear form $\mathscr{B}\colon W \times V \to \mathbb{R}$, such that*

$$C_B := \sup_{0 \neq w \in W} \sup_{0 \neq v \in V} \frac{\mathscr{B}(w,v)}{\|w\|_W \|v\|_V} < \infty, \tag{BDD}$$

the associated linear operator $B\colon W \to V^*$, defined by

$${}_{V^*}\langle Bw, v\rangle_V := \mathscr{B}(w,v),\ \forall w \in W, \forall v \in V,$$

is boundedly invertible if and only if the following two conditions are satisfied:

$$\text{(BNB1)} \qquad c_B := \inf_{0\neq w\in W} \sup_{0\neq v\in V} \frac{\mathscr{B}(w,v)}{\|w\|_W\|v\|_V} > 0,$$

$$\text{(BNB2)} \qquad \forall v \in V, \quad \sup_{0\neq w\in W} \mathscr{B}(w,v) > 0.$$

The constant c_B is called the *inf-sup constant*, while the constant C_B is called the *boundedness constant*. Since $c_B^{-1} = \|B^{-1}\|_{\mathscr{L}(V^*,W)} = \|(B^*)^{-1}\|_{\mathscr{L}(W^*,V)}$, it follows that (BNB1)–(BNB2) are equivalent to

$$\inf_{0\neq w\in W} \sup_{0\neq v\in V} \frac{\mathscr{B}(w,v)}{\|w\|_W\|v\|_V} = \inf_{0\neq v\in V} \sup_{0\neq w\in W} \frac{\mathscr{B}(w,v)}{\|w\|_W\|v\|_V} > 0. \tag{2.9}$$

This allows to swap the spaces where the infimum and the supremum are taken.

We now have to show that $\mathscr{B}_t^*$ in (2.3) satisfies the assumptions of the BNB theorem on the spaces $\mathcal{Y}_H^t$ and $\mathcal{X}^t$. The proof follows the same line as the one presented [SSt09]; we omit the proof of the (BNB2) since it does not contain any quantitative information. In order to obtain sharper bounds for C_B and c_B, we introduce equivalent norms. This is of particular relevance in this new formulation, since we want to have a constant 1 in front of the pointwise term u_2, in order to exploit the temporal decomposition, which we present in the next section.

In virtue of the properties of A, fractional powers are well defined and the norms of V and V^* are equivalent to $\|A^{\frac12}\cdot\|_H$ and $\|A^{-\frac12}\cdot\|_H$, respectively. For a more detailed explanation of this fact we refer to [CDD$^+$14]. We therefore introduce equivalent norms on $\mathcal{X}^t$ and $\mathcal{Y}_H^t$, respectively, as follows:

$$\begin{aligned} |x|_{\mathcal{X}^t}^2 &:= \|x(0)\|_H^2 + \int_0^t \Big(\|A^{\frac12}x(s)\|_H^2 + \|A^{-\frac12}\dot{x}(s)\|_H^2\Big)\,\mathrm{d}s,\\ |y|_{\mathcal{Y}_H^t}^2 &:= \|y_2\|_H^2 + \int_0^t \|A^{\frac12}y_1(s)\|_H^2\,\mathrm{d}s. \end{aligned}$$

Lemma 2. *The norm* $|||\cdot|||_{\mathcal{X}^t}$*, defined by*

$$|||x|||_{\mathcal{X}^t}^2 := \|x(t)\|_H^2 + \int_0^t \|A^{\frac12}x(s) - A^{-\frac12}\dot{x}(s)\|_H^2\,\mathrm{d}s,$$

is equal to the norm $|\cdot|_{\mathcal{X}^t}$*, for every* $t\in[0,T]$.

Proof. We have

$$\begin{aligned} |||x|||_{\mathcal{X}^t}^2 &= \|x(t)\|_H^2 + \int_0^t \Big(\|A^{\frac12}x(s)\|_H^2 + \|A^{-\frac12}\dot{x}(s)\|_H^2 - 2\,{}_V\langle x(s),\dot{x}(s)\rangle_{V^*}\Big)\,\mathrm{d}s\\ &= \|x(0)\|_H^2 + \int_0^t \Big(\|A^{\frac12}x(s)\|_H^2 + \|A^{-\frac12}\dot{x}(s)\|_H^2\Big)\,\mathrm{d}s = |x|_{\mathcal{X}^t}^2, \end{aligned}$$

because $\langle A^{\frac12}x(s), A^{-\frac12}\dot{x}(s)\rangle_H = {}_V\langle x(s),\dot{x}(s)\rangle_{V^*} = \frac12\frac{\mathrm{d}}{\mathrm{d}t}\|x(s)\|_H^2$. □

We now compute c_B and C_B for $\mathscr{B}_t^*$ with respect to $|\cdot|_{\mathcal{Y}_H^t}$ and $|\cdot|_{\mathcal{X}^t}$.

Theorem 3. *The bilinear form $\mathscr{B}_t^*(\cdot,\cdot)$ satisfies the following:*

$$C_B := \sup_{0\neq y\in\mathcal{Y}^t\times H}\ \sup_{0\neq x\in\mathcal{X}^t} \frac{\mathscr{B}_t^*(y,x)}{|y|_{\mathcal{Y}_H^t}|x|_{\mathcal{X}^t}} = 1, \tag{2.10}$$

$$c_B := \inf_{0\neq y\in\mathcal{Y}^t\times H}\ \sup_{0\neq x\in\mathcal{X}^t} \frac{\mathscr{B}_t^*(y,x)}{|y|_{\mathcal{Y}_H^t}|x|_{\mathcal{X}^t}} = 1. \tag{2.11}$$

Proof. We first notice that

$$\begin{aligned}|\mathscr{B}_t^*(y,x)| &\leq \int_0^t |{}_V\langle y_1(s), -\dot{x}(s)+Ax(s)\rangle_{V^*}|\,\mathrm{d}s + |\langle y_2, x(t)\rangle_H| \\ &\leq \int_0^t \|A^{\frac{1}{2}}y_1(s)\|_H\,\|-A^{-\frac{1}{2}}\dot{x}(s)+A^{\frac{1}{2}}x(s)\|_H\,\mathrm{d}s + \|y_2\|_H\|x(t)\|_H \\ &\leq |y|_{\mathcal{Y}_H^t}|\!|\!|x|\!|\!|_{\mathcal{X}^t} = |y|_{\mathcal{Y}_H^t}|x|_{\mathcal{X}^t}.\end{aligned}$$

This proves $C_B \leq 1$. To show $c_B \geq 1$, we use the second variant in (2.9) and prove

$$\forall x\in\mathcal{X}^t,\ \exists y_x\in\mathcal{Y}_H^t : \mathscr{B}_t^*(y_x,x) \geq |y_x|_{\mathcal{Y}_H^t}|\!|\!|x|\!|\!|_{\mathcal{X}^t} = |y_x|_{\mathcal{Y}_H^t}|x|_{\mathcal{X}^t}.$$

For $x\in\mathcal{X}^t$ we choose $y_x = \big(x - A^{-1}\dot{x}, x(t)\big)$, which belongs to $\mathcal{Y}_H^t$, since

$$|y_x|^2_{\mathcal{Y}_H^t} = \|x(t)\|_H^2 + \|A^{\frac{1}{2}}(x-A^{-1}\dot{x})\|^2_{L^2((0,t);H)} = |\!|\!|x|\!|\!|^2_{\mathcal{X}^t} = |x|^2_{\mathcal{X}^t}. \tag{2.12}$$

By expanding the bilinear form and using(2.12), we have

$$\begin{aligned}\mathscr{B}_t^*(y_x,x) &= \int_0^t \langle x(s)-A^{-1}\dot{x}(s), -\dot{x}(s)+Ax(s)\rangle_H\,\mathrm{d}s + \|x(t)\|_H^2 \\ &= \int_0^t \|A^{\frac{1}{2}}x(s) - A^{-\frac{1}{2}}\dot{x}(s)\|_H^2\,\mathrm{d}s + \|x(t)\|_H^2 = |\!|\!|x|\!|\!|^2_{\mathcal{X}^t} = |x|_{\mathcal{X}^t}|y_x|_{\mathcal{Y}_H^t}.\end{aligned}$$

Hence, $c_B \geq 1$. Since $c_B \leq C_B$, we conclude that they are both equal to 1. □

As a consequence, since the bilinear form fulfils the hypothesis of the the BNB theorem, the operator $B_t \in \mathcal{L}(\mathcal{Y}_H^t, (\mathcal{X}^t)^*)$ associated with $\mathscr{B}_t^*(\cdot,\cdot)$ via

$$\mathscr{B}_t^*(y,x) = {}_{(\mathcal{X}^t)^*}\langle B_t y, x\rangle_{\mathcal{X}^t}$$

is boundedly invertible, and $|y|_{\mathcal{Y}_H^t} \leq \|\mathscr{F}\|_{(\mathcal{X}^t,|\cdot|_{\mathcal{X}^t})^*}$. We note that for a right-hand side of the form § 2.2.1, $\mathscr{F}_t$ belongs to the dual space of $(\mathcal{X}^t, |\cdot|_{\mathcal{X}^t})$ for any $t\leq T$, if $f\in L^2((0,T);V^*)$ and $u_0\in H$. In fact,

$$\begin{aligned}\|\mathscr{F}_t\|_{(\mathcal{X}^t,|\cdot|_{\mathcal{X}^t})^*} &\leq \Big[\int_0^t \|A^{-\frac{1}{2}}f(s)\|_H^2\,\mathrm{d}s + \|u_0\|_H^2\Big]^{\frac{1}{2}} \\ &\leq \Big[A_{\min}^{-1}\int_0^t \|f(s)\|_{V^*}^2\,\mathrm{d}s + \|u_0\|_H^2\Big]^{\frac{1}{2}}.\end{aligned} \tag{2.13}$$

By combining the BNB theorem with (2.13), we thus achieve the estimate

$$\int_0^t \|A^{\frac{1}{2}}u_1(s)\|_H^2\,\mathrm{d}s + \|u_2(t)\|_H^2 \leq \int_0^t \|A^{-\frac{1}{2}}f(s)\|_H^2\,\mathrm{d}s + \|u_0\|_H^2.$$

In particular, by using the equivalence between $|\cdot|_{\mathcal{Y}_H^t}$ and $\|\cdot\|_{\mathcal{Y}_H^t}$, and the last bound in (2.13), we obtain that:

$$A_{\min}\int_0^t \|u_1(s)\|_V^2\,\mathrm{d}s + \|u_2(t)\|_H^2 \leq A_{\min}^{-1}\int_0^t \|f(s)\|_{V^*}^2\,\mathrm{d}s + \|u_0\|_H^2. \tag{2.14}$$

We emphasize that we have a constant 1 in front of u_2. Therefore, we can split and recompose the problem as we please, and the bounds for the norms will compose accordingly, without accumulation of constants. More precisely, if we consider the same problem on $[0,r]$ with initial data $u_0 \in H$, and on $[r,t]$ with initial data given by the $u_2(r) \in H$ previously obtained, then we have the two local bounds:

$$\begin{aligned} A_{\min} \int_0^r \|u_1(s)\|_V^2 \,\mathrm{d}s + \|u_2(r)\|_H^2 &\leq A_{\min}^{-1} \int_0^r \|f(s)\|_{V^*}^2 \,\mathrm{d}s + \|u_0\|_H^2, \\ A_{\min} \int_r^t \|u_1(s)\|_V^2 \,\mathrm{d}s + \|u_2(t)\|_H^2 &\leq A_{\min}^{-1} \int_r^t \|f(s)\|_{V^*}^2 \,\mathrm{d}s + \|u_2(r)\|_H^2, \end{aligned} \tag{2.15}$$

which sum up to the global bound (2.14). We have thus a local inf-sup theory consistent with the global one, which can be exploited to derive local estimates which can be put together to build global estimates.

We summarize this in the following theorem.

Theorem 4 (Existence and uniqueness)**.** *For a right-hand side of the form § 2.2.1, with $u_0 \in H$ and $f \in L^2((0,T);V^*)$, there exists a unique solution $u = (u_1, u_2)$ in $L^2((0,T);V) \times \mathscr{C}([0,T];H)$ to Problem* (2.5)*. Its norm satisfies the following bound:*

$$A_{\min} \int_0^T \|u_1(t)\|_V^2 \,\mathrm{d}t + \sup_{t\in[0,T]} \|u_2(t)\|_H^2 \leq A_{\min}^{-1} \int_0^T \|f(t)\|_{V^*}^2 \,\mathrm{d}t + \|u_0\|_H^2,$$

and in particular it holds that $u_1 = u_2 \in \mathcal{X}$.

Remark 5. *In case the right-hand side is not the one introduced in § 2.2.1, we still obtain existence and uniqueness as in Theorem 4, but the bounds of the norms are modified according to the bounds that can be obtained for $\|\mathscr{F}_t\|_{(\mathcal{X}^t,|\cdot|_{\mathcal{X}^t})^*}$. The modifications for the cases presented in § 2.2.2 or § 2.2.4 are easy to derive, while for the case of § 2.2.3 the theory required is more involved and we refer to* [LM16] *for the details.*

2.5. **Further spatial regularity.** In order to measure spatial regularity use the spaces $\dot{H}^\gamma = D(A^{\frac{\gamma}{2}})$ with norms $\|v\|_{\dot{H}^\gamma} = \|A^{\frac{\gamma}{2}} v\|_H$ for $\gamma \in \mathbb{R}$.

Theorem 6 (Spatial regularity)**.** *Assume $\beta \geq 0$. The bilinear form defining problem* (2.5) *is bounded and satisfies the inf-sup conditions on the couple of spaces $L^2((0,t);\dot{H}^{\beta+1}) \times \dot{H}^\beta$ and $L^2((0,t);\dot{H}^{1-\beta}) \cap H^1((0,t);\dot{H}^{-1-\beta})$. In particular, for a right-hand side of the form § 2.2.1, if $f \in L^2((0,T);\dot{H}^{\beta-1})$ and $u_0 \in \dot{H}^\beta$, there exists a unique solution $u = (u_1,u_2) \in L^2((0,T);\dot{H}^{\beta+1}) \times \mathscr{C}([0,T];\dot{H}^\beta)$ to* (2.5)*. Its norm satisfies the following bound:*

$$\int_0^T \|u_1(t)\|_{\dot{H}^{\beta+1}}^2 \,\mathrm{d}t + \sup_{t\in[0,T]} \|u_2(t)\|_{\dot{H}^\beta}^2 \leq \int_0^T \|f(t)\|_{\dot{H}^{\beta-1}}^2 \,\mathrm{d}t + \|u_0\|_{\dot{H}^\beta}^2,$$

and it holds, in particular, that $u_1 = u_2 \in L^2((0,T);\dot{H}^{\beta+1}) \cap H^1((0,T);\dot{H}^{\beta-1})$.

3. Discretization

We start this section by introducing a discretization based on test functions which are piecewise linear in time and trial functions which are piecewise constant in time. The scheme that we obtain turns out to be a modification of the Crank–Nicolson scheme, namely with a first step of Euler backward and a final step of Euler forward.

3.1. Discretization with polynomials of lowest degree in time.

We consider a partition of the time interval $[0,T]$, given by $\mathcal{T}_k = \{0 = t_0 < \cdots < t_i < t_{i+1} < \cdots < t_N = T\}$, with $k_i = t_{i+1} - t_i$, and $k = \max_i k_i$. We denote by $\mathcal{T}_k^n$ the partition $\mathcal{T}_k$ restricted to the interval $[0, t_n]$. We denote by I_i the interval $[t_i, t_{i+1}]$, by S_k the space of continuous piecewise linear functions with respect to $\mathcal{T}_k$, and by Q_k the space of piecewise constant functions for the same partition, with the convention that S_k^n and Q_k^n refer to the partition $\mathcal{T}_k^n$. We introduce V_h as a standard finite element space of continuous piecewise polynomials of degree less or equal to p, over a quasi-uniform family of triangulations of the spatial domain, with mesh size h. Since temporal discretization is our main concern, we assume that $p \geq 1$ is sufficiently large for our analysis to make sense.

The finite-dimensional subspaces that we use are defined as $\mathcal{Y}_{h,k} := Q_k \otimes V_h$, and $\mathcal{X}_{h,k} := S_k \otimes V_h$; consistently with the notation introduced above we introduce the family of spaces $\mathcal{Y}_{h,k}^n$ and $\mathcal{X}_{h,k}^n$.

We denote the standard basis of piecewise linear "hat" functions generating S_k by $\{\phi_i\}_{i=0}^N$ and the standard basis of piecewise constant functions generating Q_k by $\{\psi_i\}_{i=0}^{N-1}$. We denote by $\mathscr{B}^*$ and $\mathscr{F}$ the bilinear form and the load functional defined in (2.3)–(2.4). If we start from the formulation in (2.2), then the discretized problem can be written as:

$$U \in \mathcal{Y}_{h,k} \times V_h \colon \mathscr{B}^*(U, X) = \mathscr{F}(X), \quad \forall X \in \mathcal{X}_{h,k}. \tag{3.1}$$

For a formal proof of the existence and uniqueness of a solution to the discrete problem in (3.1), we follow [UP14], where the authors show that the inf-sup condition holds, and that the discrete inf-sup constant is the same as the inf-sup-constant obtained in the continuous case. However, in order to do so, the space $\mathcal{X}_{h,k}$ is endowed with a different norm, depending on the discretization:

$$\|X\|_{\mathcal{X}_k}^2 := \|X(0)\|_H^2 + \sum_{i=0}^{N-1} \int_{I_i} \left(\|\dot{X}\|_{V^*}^2 + \|\Pi_i X\|_V^2 \right) \mathrm{d}s,$$

and similarly

$$|X|_{\mathcal{X}_k}^2 := \|X(0)\|_H^2 + \sum_{i=0}^{N-1} \int_{I_i} \left(\|A^{-\frac{1}{2}} \dot{X}\|_H^2 + \|A^{\frac{1}{2}} \Pi_i X\|_H^2 \right) \mathrm{d}s,$$

where Π is the orthogonal projection, defined locally by $\big(\Pi_i X\big)(t) = \frac{1}{k_i} \int_{I_i} X(s)\,\mathrm{d}s$, $t \in I_i$.

We can now repeat the argument of Lemma 2 and Theorem 3 in $(\mathcal{Y}_{h,k} \times V_h, |\cdot|_{\mathcal{Y}_H})$ and $(\mathcal{X}_{h,k}, |\cdot|_{\mathcal{X}_k})$, and obtain inf-sup constant $c_B = 1$ and boundedness constant $C_B = 1$ (cf. Lemma 13 and Theorem 14 below). What remains now is to bound $\mathscr{F}$ with respect to the modified norm $|\cdot|_{\mathcal{X}_k}$ instead of $|\cdot|_{\mathcal{X}}$. Comparing the two norms, we note that, for all $X \in \mathcal{X}_{h,k}$,

$$\sum_{i=0}^{N-1} \int_{I_i} \left(\|\dot{X}\|_{V^*}^2 + \|X\|_V^2 \right) \mathrm{d}s \leq c_{\mathrm{S}}^2 \sum_{i=0}^{N-1} \int_{I_i} \left(\|\dot{X}\|_{V^*}^2 + \|\Pi_i X\|_V^2 \right) \mathrm{d}s,$$

since $\mathcal{X}_{h,k}$ is finite-dimensional, and where c_{S} is in general not uniform in the choice of the spaces. This leads to the equivalence of norms:

$$|X|_{\mathcal{X}_k} \leq |X|_{\mathcal{X}} \leq \max(1, c_{\mathrm{S}}) |X|_{\mathcal{X}_k}, \quad x \in \mathcal{X}_{h,k}. \tag{3.2}$$

The discrete problem is therefore not stable with respect to the original norms, unless something more is assumed on c_{S}. In [And12] it was shown that a sufficient condition for the uniform boundedness of c_{S} is:

$$C_{\mathrm{CFL}} := k \sup_{v \in V_h} \frac{\|v\|_V}{\|v\|_{V^*}} < \infty, \qquad \text{for all } h \text{ and } k. \tag{3.3}$$

By quasi-uniformity and an inverse inequality, this reduces to a CFL condition $k \leq Ch^2$. Thus (3.3) ensures the stability of the discrete problem with respect to the original norms. More precisely, for a right-hand side of the form § 2.2.1 we have, similarly to (2.13):

$$\|\mathscr{F}\|_{(\mathcal{X}_{h,k},|\cdot|_{\mathcal{X}_k})^*} \leq \Big(c_{\mathrm{S}}^2 A_{\min}^{-1} \|f\|^2_{L^2((0,T);V^*)} + \|u_0\|_H^2 \Big)^{\frac{1}{2}}. \tag{3.4}$$

Within this setting, an analogue of Theorem 4 holds for $U \in (\mathcal{Y}_{h,k} \times V_h, \|\cdot\|_{\mathcal{Y}_H})$ with the bound modified as in (3.4).

In order to see that (3.1) amounts to a time-stepping scheme, we introduce the following notation:

$$F_i^L := \frac{2}{k_{i-1}} \int_{t_{i-1}}^{t_i} f\phi_i \,\mathrm{d}s, \qquad F_i^R := \frac{2}{k_i} \int_{t_i}^{t_{i+1}} f\phi_i \,\mathrm{d}s,$$

$$\big\langle A_i^L u, v \big\rangle := \frac{2}{k_{i-1}} \int_{t_{i-1}}^{t_i} a(u, v \otimes \phi_i) \,\mathrm{d}s, \quad \big\langle A_i^R u, v \big\rangle := \frac{2}{k_i} \int_{t_i}^{t_{i+1}} a(u, v \otimes \phi_i) \,\mathrm{d}s.$$

The discrete problem, on the pair of spaces $(\mathcal{Y}_{h,k} \times V_h, \mathcal{X}_{h,k})$, can be written explicitly as follows, for any $v \in V_h$:

$$\begin{aligned}
\big\langle U_1^{(0)} - u_0, v \big\rangle + \frac{k_0}{2} \big\langle A_0^R U_1^{(0)}, v \big\rangle &= \frac{k_0}{2} \big\langle F_0^R, v \big\rangle, \\
\big\langle U_1^{(i)} - U_1^{(i-1)}, v \big\rangle + \frac{1}{2} \big\langle k_i A_i^R U_1^{(i)} + k_{i-1} A_i^L U_1^{(i-1)}, v \big\rangle &= \frac{1}{2} \big\langle k_i F_i^R + k_{i-1} F_i^L, v \big\rangle, \\
\big\langle U_2^{(N)} - U_1^{(N-1)}, v \big\rangle + \frac{k_{N-1}}{2} \big\langle A_N^L U_1^{(N-1)}, v \big\rangle &= \frac{k_{N-1}}{2} \big\langle F_N^L, v \big\rangle.
\end{aligned}$$

Here the $U_1^{(i)} \in V_h$ denote the coefficients of $U_1 = \sum_{i=0}^{N-1} U^{(i)} \psi_i$, and $U_2^{(N)} \in V_h$ is the approximation of $u_2(t_N)$. The scheme is a combination of one step of backward Euler, several steps of Crank–Nicolson, and a final step of forward Euler.

From the discrete counterpart to equation (2.14) and from (3.4), it follows that the norm of the numerical solution is bounded as follows:

$$A_{\min} \|U_1\|^2_{\mathcal{Y}} + \|U_2^{(N)}\|_H^2 \leq c_{\mathrm{S}}^2 A_{\min}^{-1} \|f\|^2_{L^2((0,T);V^*)} + \|u_0\|_H^2. \tag{3.5}$$

3.2. **Decomposition of the scheme.** By noticing that in the case of a partition with a single element, the scheme reduces to

$$\begin{aligned}
\big\langle U_1^{(0)} - u_0, v \big\rangle + \frac{k_0}{2} \big\langle A_0^R U_1^{(0)}, v \big\rangle &= \frac{k_0}{2} \big\langle F_0^R, v \big\rangle, \\
\big\langle U_2^{(1)} - U_1^{(0)}, v \big\rangle + \frac{k_0}{2} \big\langle A_1^L U_1^{(0)}, v \big\rangle &= \frac{k_0}{2} \big\langle F_1^L, v \big\rangle,
\end{aligned}$$

we can think of iterating such a decomposition over each time interval I_i, thus obtaining the extra values that approximate $u_2(t_i)$ at each grid point t_i.

The scheme becomes, for $i = 0, \dots, N-1$ and $U^{(0)} = u_0$:

$$\begin{aligned} \langle U_1^{(i)} - U_2^{(i)}, v\rangle + \frac{k_i}{2}\langle A_i^R U_1^{(i)}, v\rangle &= \frac{k_i}{2}\langle F_i^R, v\rangle, \\ \langle U_2^{(i+1)} - U_1^{(i)}, v\rangle + \frac{k_i}{2}\langle A_{i+1}^L U_1^{(i)}, v\rangle &= \frac{k_i}{2}\langle F_{i+1}^L, v\rangle. \end{aligned} \tag{3.6}$$

It follows from a suitable variant of Theorem 4 that the following holds:

$$A_{\min}\|U_1\|_{\mathcal{Y}}^2 + \max_{i=1,\dots,N} \|U_2^{(i)}\|_H^2 \leq c_{\mathrm{S}}^2 A_{\min}^{-1}\|f\|_{L^2((0,T);V^*)}^2 + \|u_0\|_H^2.$$

Remark 7. *An important thing to notice is that $U_2^{(n)}$ can be constructed from $U_1\big|_{(0,t_n)}$ even if one does not want to introduce the splitting proposed above. The second equation in* (3.6) *can indeed by used at any time, as long as we have the values of U_1.*

3.3. Temporal discretization with polynomials of higher degree.

The results in this section can be generalized to polynomials of arbitrary degree with respect to time. We denote by $S_{k,q+1}$ the space of continuous functions that are piecewise polynomials of degree at most $q+1$, with respect to the partition $\mathcal{T}_k$, and by $Q_{k,q}$ the space of discontinuous functions which are piecewise polynomials of degree at most q, for the same partition. We adopt the same convention and notation as before and define the finite-dimensional subspaces $\mathcal{Y}_{h,k,q} := Q_{k,q} \otimes V_h$, and $\mathcal{X}_{h,k,q+1} := S_{k,q+1} \otimes V_h$, for some finite-dimensional subspace $V_h \subset V$.

The discretized problem can be written in variational form as

$$U \in \mathcal{Y}_{h,k,q} \times V_h \colon \mathscr{B}^*(U, X) = \mathscr{F}(X), \quad \forall X \in \mathcal{X}_{h,k,q+1}. \tag{3.7}$$

Results of existence and uniqueness follow from a minor modification of the argument used in the case $q = 0$, that is, by modifying the norm on the space $\mathcal{X}_{h,k,q+1}$ as follows:

$$\begin{aligned} \|X\|_{\mathcal{X}_{k,q+1}}^2 &:= \sum_{i=0}^{N-1}\int_{I_i}\left(\|\dot{X}\|_{V^*}^2 + \|\Pi_i^{(q)}X\|_V^2\right)\mathrm{d}s + \|X(0)\|_H^2, \\ |X|_{\mathcal{X}_{k,q+1}}^2 &:= \sum_{i=0}^{N-1}\int_{I_i}\left(\|\dot{X}\|_{\dot{H}^{-1}}^2 + \|\Pi_i^{(q)}X\|_{\dot{H}^1}^2\right)\mathrm{d}s + \|X(0)\|_H^2, \end{aligned} \tag{3.8}$$

where now $\Pi^{(q)}$ is locally defined on each I_i as the orthogonal L^2-projection onto the space of polynomials of degree at most q. In particular, the splitting introduced in § 3.2 still holds.

3.4. The roles of U_1 and U_2.

In this section we state a result that relates the two components of U by means of a discretization based on the first space-time formulation. We start by considering the original problem (1.1). The first space-time formulation (2.1) leads to the following discretization:

$$W \in \mathcal{X}_{h,k,q+1} : \mathscr{B}(W, Y) = \mathscr{F}(Y), \quad \forall Y \in \mathcal{Y}_{h,k,q} \times V_h,$$

while the weak space-time formulation is given in (3.7). The next theorem states that the discrete solutions to the first and to the weak formulations of (1.1) differ only up to a term proportional to the interpolation error of the right-hand side. Since this result is not central in this paper, we omit the proof.

Theorem 8. *If $f^{(\gamma)} \in L^2((0,T);V)$ for some $\gamma \in \mathbf{N}$, then*

$$\|U_1 - \Pi^{(q)} W\|_{L^2((0,T);V)} + \|U_2^{(N)} - W(t_N)\|_H \leq C k^{\theta+1} \|f^{(\theta)}\|_{L^2((0,T);V)},$$

where $\theta := \min\{q+1, \gamma\}$.

4. A priori error estimates

In order to obtain error estimates for our scheme, we first rely on the quasi-optimality theory, thus achieving an error estimate consistent with the natural norm of the solution in (3.5). However, numerical experiments (see Figures 1b and 2b) and Theorem 12 suggest that the second component of the solution converges faster, with a rate proportional to k^2. This is consistent with the fact that our method is a modification of the standard Crank–Nicolson method. By means of a duality argument we give a rigorous proof of this fact in Theorem 21 in Section 5.

4.1. Quasi-optimality.

We consider the subspaces $\mathcal{Y}_{h,k} \times V_h \subset \mathcal{Y}_H$ and $\mathcal{X}_{h,k} \subset \mathcal{X}$ previously introduced, endowed with the norms $|\cdot|_{\mathcal{Y}_H}$ and $|\cdot|_{\mathcal{X}_k}$, respectively. The following result of quasi-optimality holds:

Theorem 9. *If u and U are solutions to* (2.5) *and* (3.1)*, respectively, the error $u - U$ satisfies the following bound:*

$$\begin{aligned} &A_{\min} \|u_1 - U_1\|^2_{L^2((0,t_n);V)} + \|u_2(t_n) - U_2^{(n)}\|^2_H \\ &\quad \leq \max\{1, c_{\mathrm{S}}\}^2 \Big(A_{\max} \|u_1 - Y_1\|^2_{L^2((0,t_n);V)} + \|u_2(t_n) - Y_2^{(n)}\|^2_H \Big), \end{aligned} \tag{4.1}$$

for arbitrary $Y_1 \in \mathcal{Y}_{h,k}$ and $Y_2^{(n)} \in V_h$ and for any n. In particular, it follows that

$$\begin{aligned} &A_{\min} \|u_1 - U_1\|^2_{L^2((0,T);V)} + \max_{i=1,\dots,N} \|u_2(t_i) - U_2^{(i)}\|^2_H \\ &\quad \leq \max\{1, c_{\mathrm{S}}\}^2 \Big(A_{\max} \|u_1 - Y_1\|^2_{L^2((0,T);V)} + \max_{i=1,\dots,N} \|u_2(t_i) - Y_2^{(i)}\|^2_H \Big). \end{aligned} \tag{4.2}$$

Proof. We consider the problem on $(0, t_n)$ with arbitrary t_n and omit t_n in the notation for the spaces and bilinear form. We denote by $R\colon \mathcal{Y}_H \mapsto \mathcal{Y}_{h,k} \times V_h$ the Ritz projection, defined as $Ru = U$, that is,

$$\mathscr{B}^*(R\phi, X) = \mathscr{B}^*(\phi, X), \quad \forall X \in \mathcal{X}_{h,k}. \tag{4.3}$$

Since R is idempotent and $\mathcal{Y}_H$ is a Hilbert space, we have $\|I - R\|_{\mathscr{L}(\mathcal{Y}_H)} = \|R\|_{\mathscr{L}(\mathcal{Y}_H)}$ (see [XZ03]), so that, for any $Y \in \mathcal{Y}_{h,k} \times V_h$,

$$|u - U|_{\mathcal{Y}_H} = |(I - R)u|_{\mathcal{Y}_H} = |(I - R)(u - Y)|_{\mathcal{Y}_H} \leq \|R\|_{\mathscr{L}(\mathcal{Y}_H)} |u - Y|_{\mathcal{Y}_H}.$$

Here, we have

$$\begin{aligned} \|R\|_{\mathscr{L}(\mathcal{Y}_H)} &= \sup_{\phi \in \mathcal{Y}_H} \frac{|R\phi|_{\mathcal{Y}_H}}{|\phi|_{\mathcal{Y}_H}} \leq \frac{1}{c_B} \sup_{\phi \in \mathcal{Y}_H} \sup_{X \in \mathcal{X}_{h,k}} \frac{\mathscr{B}^*(R\phi, X)}{|\phi|_{\mathcal{Y}_H} |X|_{\mathcal{X}_k}} \\ &= \frac{1}{c_B} \sup_{\phi \in \mathcal{Y}_H} \sup_{X \in \mathcal{X}_{h,k}} \frac{\mathscr{B}^*(\phi, X)}{|\phi|_{\mathcal{Y}_H} |X|_{\mathcal{X}_k}} \leq \frac{C_B}{c_B} \sup_{\phi \in \mathcal{Y}_H} \sup_{X \in \mathcal{X}_{h,k}} \frac{|\phi|_{\mathcal{Y}_H} |X|_{\mathcal{X}}}{|\phi|_{\mathcal{Y}_H} |X|_{\mathcal{X}_k}}, \end{aligned}$$

where we first used the discrete counterpart of (2.11) with respect to $|\cdot|_{\mathcal{Y}_H}$ and $|\cdot|_{\mathcal{X}_k}$, then (4.3), and (2.10). Finally, by means of (3.2) we obtain that

$$\frac{C_B}{c_B} \sup_{\phi \in \mathcal{Y}_H} \sup_{X \in \mathcal{X}_{h,k}} \frac{|\phi|_{\mathcal{Y}_H} |X|_{\mathcal{X}}}{|\phi|_{\mathcal{Y}_H} |X|_{\mathcal{X}_k}} \leq \frac{C_B}{c_B} \max\{1, c_{\mathrm{S}}\} = \max\{1, c_{\mathrm{S}}\},$$

since $C_B = c_B = 1$. Since $Y \in \mathcal{Y}_{h,k} \times V_h$ is arbitrary, (4.1) follows by using the equivalence between the norms $\|\cdot\|_{L^2((0,t_n);V)}$ and $\|\cdot\|_{L^2((0,t_n);\dot{H}^1)}$. Since t_n is arbitrary, the second bound (4.2) follows as well. □

4.2. **Convergence.** We first show convergence of the method under minimal assumptions, namely a right-hand side $\mathscr{F} \in \mathcal{X}^*$ and no further regularity.

Theorem 10. *Let u and U be solutions to* (2.5) *and* (3.1)*, respectively. If we assume the validity of* (3.3)*, and if $\mathscr{F} \in \mathcal{X}^*$, then $\|u - U\|_{\mathcal{Y}_H} \to 0$ as $k, h \to 0$.*

Proof. From the quasi-optimality theorem we have

$$\|u - U\|_{\mathcal{Y}_H} \leq C\|u - Y\|_{\mathcal{Y}_H}, \quad \text{for any } Y \in \mathcal{Y}_{h,k} \times V_h,$$

where C depends on $A_{\min}$, $A_{\max}$, and c_{S}, hence independent of h and k due to (3.3). We choose $\mathscr{V}$ to be a space of sufficiently smooth functions, dense in $\mathcal{Y}_H$, for example $\mathscr{V} := H^1((0,T);V) \times V$. For arbitrary ϵ, we choose $v_\epsilon \in \mathscr{V}$ such that, by density,

$$\|u - v_\epsilon\|_{\mathcal{Y}_H} \leq \epsilon/2.$$

We then choose $h = h(\epsilon)$ and $k = k(\epsilon)$ such that $\tilde{v}_\epsilon \in \mathcal{Y}_{h,k} \times V_h$, which denotes the interpolant of v_ϵ, satisfies

$$\|v_\epsilon - \tilde{v}_\epsilon\|_{\mathcal{Y}_H} \leq C(h + k)\|v_\epsilon\|_{\mathscr{V}} \leq \epsilon/2.$$

We conclude

$$\|u - U\|_{\mathcal{Y}_H} \leq C\|u - \tilde{v}_\epsilon\|_{\mathcal{Y}_H} \leq C\Big(\|u - v_\epsilon\|_{\mathcal{Y}_H} + \|v_\epsilon - \tilde{v}_\epsilon\|_{\mathcal{Y}_H}\Big) \leq C\epsilon.$$

Since ϵ is arbitrary, the claim follows. □

4.3. **Convergence of first order in time.** In order to prove the next results we assume that the spatial discretization is done by using a polynomial space of sufficiently high degree, so that all the quantities we use make sense and are not trivial. This choice is not strictly necessary but it is motivated by the fact that condition (3.3) becomes $k \lesssim h^2$ in the case, for example, of spatial discretization with Lagrange elements. Thus, in order to have consistency between the spatial and the temporal rate of convergence, we need to have order 2 in the spatial H^1-norm in the following theorem (polynomials of degree $p = 2$), and similarly order 4 in the one after.

We make once again use of the spaces $\dot{H}^\beta$ as in § 2.5. The right-hand side of the expression in (4.2) can be further estimated by means of standard interpolation estimates, thus we obtain the following theorem:

Theorem 11. *Let u and U be solutions to* (2.5) *and* (3.1)*, respectively. For sufficiently smooth data f and u_0, and assuming the validity of* (3.3)*, we have:*

$$\begin{aligned}\|u_1 - U_1\|_{L^2((0,T);V)} &+ \max_{i=1,\dots,N} \|u_2(t_i) - U_2^{(i)}\|_H \\ &\leq C(k + h^2)\Big(\|f\|_{L^2((0,T);\dot{H}^1)} + \|u_0\|_{\dot{H}^2}\Big).\end{aligned}$$

Proof. Quasi-optimality (4.2) and interpolation error estimates give us that

$$\begin{aligned}\|u_1 - U_1\|_{L^2((0,T);V)} &+ \max_{i=1,\dots,N} \|u_2(t_i) - U_2^{(i)}\|_H \\ &\leq C\Big(k\|\dot{u}\|_{L^2((0,T);\dot{H}^1)} + h^2\big(\|u\|_{L^2((0,T);\dot{H}^3)} + \max_{i=1,\dots,N} \|u_2(t_i)\|_{\dot{H}^2}\big)\Big),\end{aligned}$$

for u sufficiently smooth, with C depending on c_{S}, $A_{\min}$, and $A_{\max}$. In particular, if the right-hand side is of the form defined in (2.6), we can rely on Theorem 6 with $\beta = 2$ to prove the claim. □

4.4. **Convergence of second order in time.** By means of the connection between first and second discrete space-time formulation and by using the fact that the first space-time formulation seen as a time stepping coincides with the traditional Crank–Nicolson scheme, we can obtain the following result:

Theorem 12. *The scheme in* (3.6) *converges with a rate proportional to* k^2 *at the grid points* $\{t_i\}_{i=1,\dots,N}$ *for sufficiently smooth data.*

Proof. We take W as in Theorem 8, and notice that for every t_i we have:

$$\|U_2^{(i)} - u_2(t_i)\|_H \le \|U_2^{(i)} - W(t_i)\|_H + \|u_2(t_i) - W(t_i)\|_H.$$

We can bound the first term by Ck^2 according to Theorem 8. The primal formulation produces exactly the Crank–Nicolson time stepping, so that the second term is also bounded by Ck^2. □

5. Temporal semidiscretization

We provide in Theorem 21 a direct proof of the result of Theorem 12, that does not rely on a comparison with the Crank–Nicolson method and that extends to arbitrary degree. Following [Tho06, Theorem 12.3] we present only the temporally semidiscrete part of the error, since our main focus is the time discretization. The proof is based on a duality argument but first we need to develop a substitute for the quasi-optimality theory in the semidiscrete case.

5.1. **Existence and uniqueness.** We introduce the following notation for the temporally semidiscrete spaces:

$$\mathcal{Y}_{k,q} := \{Y \in \mathcal{Y} : Y\big|_{I_i} \in \mathbb{P}^q[t] \otimes \dot{H}^1\}, \quad \mathcal{X}_{k,q+1} := \{X \in \mathcal{X} : X\big|_{I_i} \in \mathbb{P}^{q+1}[t] \otimes \dot{H}^1\},$$

and we endow $\mathcal{X}_{k,q+1}$ with the norm $|\cdot|_{\mathcal{X}_{k,q+1}}$ which we introduced in (3.8). The semidiscrete problem reads:

$$\hat{U} \in \mathcal{Y}_{k,q} \times H \colon \mathscr{B}^*(\hat{U}, X) = \mathscr{F}(X), \quad \forall X \in \mathcal{X}_{k,q+1}. \tag{5.1}$$

In particular, we can split the scheme as in (3.6) in order to produce pointwise values of $\hat{U}_2^{(i)}$ at each t_i.

Our main concern is to avoid the use of (3.2), because c_{S} would not be finite in the semidiscrete case. It turns out that a consistent theory of existence and uniqueness based on the Banach–Nečas–Babuška can be derived even in this case, although more regularity on f must be assumed. We start by presenting a semidiscrete version of Lemma 2:

Lemma 13. *The norm* $|\!|\!|\cdot|\!|\!|_{\mathcal{X}_{k,q+1}}$, *defined by*

$$|\!|\!|X|\!|\!|^2_{\mathcal{X}_{k,q+1}} := \|X(t)\|_H^2 + \sum_{i=1}^{N-1} \int_{I_i} \|A^{\frac{1}{2}}\Pi^{(q)}X(s) - A^{-\frac{1}{2}}\dot{X}(s)\|_H^2 \, \mathrm{d}s$$

is equal on $\mathcal{X}_{k,q+1}$ *to the norm* $|\cdot|_{\mathcal{X}_{k,q+1}}$.

Proof. Similarly to the proof of Lemma 2, we have

$$\begin{aligned}
|||X|||^2_{\mathcal{X}_{k,q+1}} &= \|X(t)\|^2_H + \sum_{i=1}^{N-1} \int_{I_i} \Big(\|A^{\frac{1}{2}}\Pi_i^{(q)} X\|^2_H + \|A^{-\frac{1}{2}}\dot{X}\|^2_H \\
&\qquad\qquad - 2\,{}_V\langle \Pi_i^{(q)} X, \dot{X}\rangle_{V^*} \Big)\,\mathrm{d}s \\
&= \|X(t)\|^2_H + \sum_{i=1}^{N-1} \int_{I_i} \Big(\|A^{\frac{1}{2}}\Pi_i^{(q)} X\|^2_H + \|A^{-\frac{1}{2}}\dot{X}\|^2_H \\
&\qquad\qquad - \|X(t_{i+1})\|^2_H + \|X(t_i)\|^2_H \Big)\,\mathrm{d}s \\
&= \|X(0)\|^2_H + \sum_{i=0}^{N-1} \int_{I_i} \Big(\|A^{\frac{1}{2}}\Pi_i^{(q)} X\|^2_H + \|A^{-\frac{1}{2}}\dot{X}\|^2_H \Big)\,\mathrm{d}s = |x|^2_{\mathcal{X}_{k,q+1}},
\end{aligned}$$

since $\int_{I_i} {}_V\langle \Pi_i^{(q)} X, \dot{X}\rangle_{V^*}\,\mathrm{d}s = \int_{I_i} {}_V\langle X, \dot{X}\rangle_{V^*}\,\mathrm{d}s$. This is the desired result. □

Theorem 14. *The bilinear form* (2.3) *satisfies the following:*

$$C_B := \sup_{0\neq Y\in\mathcal{Y}_{k,q}\times H}\ \sup_{0\neq X\in\mathcal{X}_{k,q+1}} \frac{\mathscr{B}^*(Y,X)}{|Y|_{\mathcal{Y}_H}|X|_{\mathcal{X}_{k,q+1}}} = 1, \tag{5.2}$$

$$c_B := \inf_{0\neq Y\in\mathcal{Y}_{k,q}\times H}\ \sup_{0\neq X\in\mathcal{X}_{k,q+1}} \frac{\mathscr{B}^*(Y,X)}{|Y|_{\mathcal{Y}_H}|X|_{\mathcal{X}_{k,q+1}}} = 1. \tag{5.3}$$

Proof. We first notice that, on each I_i,

$$\int_{I_i} {}_V\langle Y(s), -\dot{X}(s) + A^* X(s)\rangle_{V^*}\,\mathrm{d}s = \int_{I_i} {}_V\langle Y(s), -\dot{X}(s) + A^*\Pi_i^{(q)} X(s)\rangle_{V^*}\,\mathrm{d}s,$$

so that we can use Hölder's inequality as in the proof of Theorem 3 and obtain (5.2). The proof of (5.3) follows by choosing, for $X \in \mathcal{X}_{k,q+1}$,

$$Y_X = \Big(\Pi^{(q)} X - A^{-1}\dot{X}, X(t)\Big),$$

and proceeding in the same way as in the continuous case. □

Since we are in a semidiscrete case, the conditions (BNB1) and (BNB2) are not equivalent, and one should prove also the latter. We refrain from doing so and refer to [Tan13, Proposition 4.2], where a complete proof for the case $q = 0$ can be found. The case of $q > 0$ follows similarly. In order to have solvability of (5.1) it now only remains to bound $\mathscr{F}$ with respect to the norm $|\cdot|_{\mathcal{X}_{k,q+1}}$.

Lemma 15. *If $f \in L^2((0,T);\dot{H}^1)$ and $u_0 \in H$, then we have for $X \in \mathcal{X}_{k,q+1}$ the following inequality:*

$$\Big|\mathscr{F}(X)\Big| \leq \Big[\sum_{i=0}^{N-1} \Big(\int_{I_i} \|f(s)\|^2_{\dot{H}^{-1}}\,\mathrm{d}s + k_i^2 \int_{I_i} \|f(s)\|^2_{\dot{H}^1}\,\mathrm{d}s \Big) + \|u_0\|^2_H \Big]^{\frac{1}{2}} |X|_{\mathcal{X}_{k,q+1}}.$$

Proof. We use the fact that, for $X \in \mathcal{X}_{k,q+1}$ and for every subinterval I_i, we have

$$\|X - \Pi_i^{(q)} X\|^2_{L^2(I_i;\dot{H}^{-1})} \leq \|X - \Pi_i^{(0)} X\|^2_{L^2(I_i;\dot{H}^{-1})} \leq k_i^2 \|\dot{X}\|^2_{L^2(I_i;\dot{H}^{-1})}. \tag{5.4}$$

By adding and subtracting $\Pi^{(q)}X$, we have

$$\begin{aligned}\mathscr{F}(X) = \sum_{i=0}^{N-1} \Big(\int_{I_i} \langle f(s), \Pi_i^{(q)} X(s)\rangle_H \,\mathrm{d}s + \int_{I_i} \langle f(s), X(s) - \Pi_i^{(q)} X(s)\rangle_H \,\mathrm{d}s\Big) \\ + \langle u_0, X(0)\rangle_H,\end{aligned}$$

so that

$$\begin{aligned}\Big|\mathscr{F}(X)\Big| &\leq \sum_{i=0}^{N-1} \Big(\|f\|_{L^2(I_i;\dot H^{-1})} \|\Pi_i^{(q)} X\|_{L^2(I_i;\dot H^1)} + \|f\|_{L^2(I_i;\dot H^1)} k_i \|\dot X\|_{L^2(I_i;\dot H^{-1})}\Big) \\ &\quad + \|u_0\|_H \|X(0)\|_H \\ &\leq \Big[\sum_{i=0}^{N-1} \Big(\|f\|^2_{L^2(I_i;\dot H^{-1})} + k_i^2 \|f\|^2_{L^2(I_i;\dot H^1)}\Big) + \|u_0\|_H^2 \Big]^{\frac12} |X|_{\mathcal{X}_{k,q+1}},\end{aligned}$$

which proves the claim. □

The previous lemma shows in particular that

$$\|\mathscr{F}\|_{(\mathcal{X}_{k,q+1}, |\cdot|_{\mathcal{X}_{k,q+1}})^*} \leq \Big(\|f\|^2_{L^2((0,T);\dot H^{-1})} + k^2 \|f\|^2_{L^2((0,T);\dot H^1)} + \|u_0\|_H^2 \Big)^{\frac12},$$

so that the next theorem follows:

Theorem 16. *If $f \in L^2((0,T);\dot H^1)$ and $u_0 \in H$, there exists a unique solution $\hat U \in \mathcal{Y}_{k,q} \times H$ to the semidiscrete problem, and its norm is such that*

$$|\hat U|_{\mathcal{Y}_H} \leq \Big(\|f\|^2_{L^2((0,T);\dot H^{-1})} + k^2 \|f\|^2_{L^2((0,T);\dot H^1)} + \|u_0\|_H^2 \Big)^{\frac12}.$$

5.2. **A priori error estimate.** In the proof of Theorem 9 we relied on the boundedness of $\mathscr{B}^*$ with respect to $|\cdot|_{\mathcal{X}}$ and $|\cdot|_{\mathcal{Y}_H}$, together with the norm equivalence (3.2) between $|\cdot|_{\mathcal{X}}$ and $|\cdot|_{\mathcal{X}_k}$, to show its boundedness with respect to $|\cdot|_{\mathcal{X}_k}$ and $|\cdot|_{\mathcal{Y}_H}$. This does not work here due to the fact that the constant c_{S}, that would appear, is not finite in the semidiscrete case. We solve this problem by bounding the bilinear form with respect to $|\cdot|_{\mathcal{X}_{k,q+1}}$ and a stronger norm on $\mathcal{Y}$.

Lemma 17. *The following boundedness estimate holds for any $X \in \mathcal{X}_{k,q+1}$ and $y \in L^2((0,t_n);\dot H^3) \times H$ such that $y_2 = 0$:*

$$|\mathscr{B}^*(y,X)| \leq C \Big[\sum_{i=0}^{N-1} \Big(\int_{I_i} \|y\|^2_{\dot H^1} \,\mathrm{d}s + k_i^2 \int_{I_i} \|y\|^2_{\dot H^3} \,\mathrm{d}s \Big)\Big]^{\frac12} |X|_{\mathcal{X}_{k,q+1}}.$$

Proof. The term we need to modify in order to achieve the $|\cdot|_{\mathcal{X}_{k,q+1}}$-norm, is the one not involving the time derivative. For this term we have

$$\int_{I_i} \langle y, A^* X\rangle_H \,\mathrm{d}s = \int_{I_i} \langle Ay, \Pi_i^{(q)} X\rangle_H \,\mathrm{d}s + \int_{I_i} \langle Ay, X - \Pi_i^{(q)} X\rangle_H \,\mathrm{d}s.$$

If we now take norms and use (5.4), we get

$$\Big| \int_{I_i} \langle y, A^* X\rangle_H \,\mathrm{d}s \Big| \leq \|y\|_{L^2(I_i;\dot H^1)} \|\Pi_i^{(q)} X\|_{L^2(I_i;\dot H^1)} + k_i \|y\|_{L^2(I_i;\dot H^3)} \|\dot X\|_{L^2(I_i;\dot H^{-1})}.$$

This proves the claim. □

We can now prove a substitute for a quasi-optimality theorem for the semidiscrete case.

Theorem 18. *If u and $\hat{U}$ are solutions to* (2.2) *and* (5.1)*, respectively, then the error $u - \hat{U}$ satisfies the following bound:*

$$|u - \hat{U}|_{\mathcal{Y}_H} \leq C\Big[\sum_{i=0}^{N-1}\Big(\int_{I_i}\|u_1 - Y_1\|^2_{\dot{H}^1}\,\mathrm{d}s + k_i^2\int_{I_i}\|u_1 - Y_1\|^2_{\dot{H}^3}\,\mathrm{d}s\Big)\Big]^{\frac{1}{2}},$$

for any $Y_1 \in \mathcal{Y}_{k,q,3} := \{Y \in \mathcal{Y}\colon Y\big|_{I_i} \in \mathbb{P}^q[t] \otimes \dot{H}^3\}$.

Proof. We notice that we have the orthogonality

$$\mathscr{B}^*(u - \hat{U}, X) = 0, \quad \forall X \in \mathcal{X}_{k,q+1},$$

so that, for any $Y \in \mathcal{Y}_{k,q} \times H$,

$$\begin{aligned}
|u - \hat{U}|_{\mathcal{Y}_H} &\leq |u - Y|_{\mathcal{Y}_H} + |\hat{U} - Y|_{\mathcal{Y}_H}\\
&\leq |u - Y|_{\mathcal{Y}_H} + \sup_{X \in \mathcal{X}_{k,q+1}} \frac{\mathscr{B}^*(\hat{U} - Y, X)}{|X|_{\mathcal{X}_{k,q+1}}}\\
&= |u - Y|_{\mathcal{Y}_H} + \sup_{X \in \mathcal{X}_{k,q+1}} \frac{\mathscr{B}^*(u - Y, X) + \mathscr{B}^*(\hat{U} - u, X)}{|X|_{\mathcal{X}_{k,q+1}}}\\
&= |u - Y|_{\mathcal{Y}_H} + \sup_{X \in \mathcal{X}_{k,q+1}} \frac{\mathscr{B}^*(u - Y, X)}{|X|_{\mathcal{X}_{k,q+1}}},
\end{aligned}$$

where the first inequality comes from (5.3), while the last equality comes from orthogonality. If we choose Y such that its second component is equal to u_2, which is possible in the semidiscrete case, then we have $Y_2 - u_2 = 0$, so that Lemma 17 applies, giving:

$$|u - \hat{U}|_{\mathcal{Y}_H} \leq C\Big[\sum_{i=0}^{N-1}\Big(\int_{I_i}\|u_1 - Y_1\|^2_{\dot{H}^1}\,\mathrm{d}s + k_i^2\int_{I_i}\|u_1 - Y_1\|^2_{\dot{H}^3}\,\mathrm{d}s\Big)\Big]^{\frac{1}{2}},$$

for any arbitrary $Y_1 \in \mathcal{Y}_{k,q,3} := \{Y \in \mathcal{Y}\colon Y\big|_{I_i} \in \mathbb{P}^q[t] \otimes \dot{H}^3\}$. □

Note that in this proof we cannot use $\|I - R\| = \|R\|$, as in the proof of Theorem 9, because we use different norms on U and u in $U = Ru$.

Remark 19. *It is worth noticing that everything said so far still holds when we shift spatial regularity and work with a solution $u \in L^2((0,T); \dot{H}^{\beta+1})$; it is easy to see that this leads to the following modified inequality:*

$$\begin{aligned}
&\|u - \hat{U}\|_{L^2((0,T);\dot{H}^{\beta+1})\times\dot{H}^\beta}\\
&\qquad\leq C\Big[\sum_{i=0}^{N-1}\Big(\int_{I_i}\|u_1 - Y_1\|^2_{\dot{H}^{\beta+1}}\,\mathrm{d}s + k_i^2\int_{I_i}\|u_1 - Y_1\|^2_{\dot{H}^{\beta+3}}\,\mathrm{d}s\Big)\Big]^{\frac{1}{2}},
\end{aligned}$$

for any Y_1 in the space $\mathcal{Y}_{k,q,\beta+3} := \{Y \in \mathcal{Y}\colon Y\big|_{I_i} \in \mathbb{P}^q[t] \otimes \dot{H}^{\beta+3}\}$,

5.3. **Convergence of order** $q + 1$**.** Now that we have an abstract error estimate for the semidiscrete case, we can derive an analogue to Theorem 11.

Theorem 20. *For sufficiently smooth data, the error in the semidiscrete scheme* (5.1) *satisfies the following inequality, for* $\beta \geq 0$,

$$\begin{aligned}\|u_1 - \hat{U}_1\|_{L^2((0,T);\dot{H}^{\beta+1})} &+ \max_{i=1,\dots,N} \|u_2(t_i) - \hat{U}_2^{(i)}\|_{\dot{H}^{\beta}} \\ &\leq C\Big[\sum_{i=0}^{N-1} k_i^{2(q+1)} \Big(\|u_1^{(q)}\|^2_{L^2(I_i;\dot{H}^{\beta+3})} + \|u_1^{(q+1)}\|^2_{L^2(I_i;\dot{H}^{\beta+1})}\Big)\Big]^{\frac{1}{2}}.\end{aligned}$$

5.4. **Pointwise superconvergence of order** $2(q+1)$**.** We can now give a rigorous proof of Theorem 12 that does not rely on the explicit form of the scheme obtained by discretizing with the first space-time formulation. The advantage of an explicit proof is that it holds for any arbitrary q, while Theorem 12 relies on the fact that the particular time stepping obtained for the first space-time formulation of (1.1) is the Crank–Nicolson method.

Theorem 21. *For sufficiently smooth data, the numerical solution obtained by splitting* (5.1) *is superconvergent at the grid points, that is,*

$$\begin{aligned}&\max_{n=1,\dots,N} \|u_2(t_n) - U_2^{(n)}\|_H \\ &\quad\leq Ck^{q+1}\Big[\sum_{i=0}^{N-1} k_i^{2(q+1)} \Big(\|u_1^{(q)}\|^2_{L^2(I_i;\dot{H}^{2q+5})} + \|u_1^{(q+1)}\|^2_{L^2(I_i;\dot{H}^{2q+3})}\Big)\Big]^{\frac{1}{2}},\end{aligned} \tag{5.5}$$

or, in terms of the data,

$$\begin{aligned}&\max_{n=1,\dots,N} \|u_2(t_n) - U_2^{(n)}\|_H \\ &\quad\leq Ck^{2(q+1)}\Big(\|f^{(q)}\|_{L^2((0,T);\dot{H}^{2q+3})} + \|u_{q,0}\|_{\dot{H}^{2q+4}}\Big),\end{aligned} \tag{5.6}$$

where $u_{q,0}$ *is defined as:*

$$u_{q,0} := \sum_{k=0}^{q-1} (-A)^k f^{(q-1-k)}(0) + (-A)^q u_0.$$

Proof. We consider the problem on $(0, t_n)$ with arbitrary t_n and omit t_n in the notation for the spaces and bilinear form. The following orthogonality relation is satisfied, for $e = u - \hat{U}$:

$$\mathscr{B}^*(e, X) = 0, \quad \forall X \in \mathcal{X}_{k,q+1}. \tag{5.7}$$

We now consider the adjoint problem given by

$$\begin{aligned}-\dot{z}(s) + Az(s) &= 0, \quad \text{in } V^*,\ s \in (0, t_n),\\ z(t_n) &= \phi, \quad \text{in } H,\end{aligned}$$

where ϕ is an arbitrary element of H. The first space-time formulation of this problem is given in the continuous case by

$$z \in \mathcal{X}\colon \mathscr{B}^*(y, z) = \langle y_2, \phi\rangle_H, \quad \forall y = (y_1, y_2) \in \mathcal{Y}_H. \tag{5.8}$$

In particular, if we choose $y = (0, e_2)$ in (5.8) and use the orthogonality relation (5.7), we have that for any $X \in \mathcal{X}_{k,q+1}$:

$$\langle e_2, \phi\rangle_H = \mathscr{B}^*(e, z) = \mathscr{B}^*(e, z - X).$$

If we assume that we have sufficient smoothness for the next quantities to make sense, we have:

$$|\langle e_2, \phi\rangle_H| \le \|e\|_{L^2((0,t_n);\dot H^{\beta+1})\times \dot H^{\beta}} \|z - X\|_{L^2((0,t_n);\dot H^{1-\beta})\cap H^1((0,t_n);\dot H^{-1-\beta})}.$$

For the second term we choose $X \in \mathcal{X}_{k,q+1}$ to be a standard interpolant of z:

$$\begin{aligned}
&\|z - X\|_{L^2((0,t_n);\dot H^{1-\beta})\cap H^1((0,t_n);\dot H^{-1-\beta})}\\
&\qquad \le Ck^{q+1}\Big(\|z^{(q+1)}\|_{L^2((0,t_n);\dot H^{1-\beta})} + \|z^{(q+2)}\|_{L^2((0,t_n);\dot H^{-1-\beta})}\Big)\\
&\qquad = Ck^{q+1}\Big(\|A^{q+1}z\|_{L^2((0,t_n);\dot H^{1-\beta})} + \|A^{q+1}\dot z\|_{L^2((0,t_n);\dot H^{-1-\beta})}\Big)\\
&\qquad = Ck^{q+1}\Big(\|z\|_{L^2((0,t_n);\dot H^{1-\beta+(2q+2)})} + \|\dot z\|_{L^2((0,t_n);\dot H^{-1-\beta+(2q+2)})}\Big)\\
&\qquad = Ck^{q+1}\Big(\|z\|_{L^2((0,t_n);\dot H^{1})} + \|\dot z\|_{L^2((0,t_n);\dot H^{-1})}\Big) \le Ck^{q+1}\|\phi\|_H,
\end{aligned}$$

where we chose $\beta = 2(q+1)$ and used a standard bound for z. Hence,

$$\|e_2\|_H \le Ck^{q+1}\|e\|_{L^2((0,t_n);\dot H^{2(q+1)+1})\times \dot H^{2(q+1)}},$$

and (5.5) follows by Theorem 20 and recalling that n is arbitrary.

In order to show (5.6), we notice that (5.5) implies the non-localized bound

$$\max_{n=1,\ldots,N}\|e_2^{(n)}\|_H \le Ck^{2(q+1)}\Big(\|u_1^{(q)}\|_{L^2((0,T);\dot H^{2q+5})} + \|u_1^{(q+1)}\|_{L^2((0,T);\dot H^{2q+3})}\Big).$$

The final step is achieved by bounding the norm of the solution in terms of the norm of its data. By using the notation $u_q := u^{(q)}$, and noticing that u_q is the solution to the primal formulation of

$$\dot u_q + Au_q = f^{(q)},\ t \in (0,T); \quad u_q(0) = u_{q,0},$$

we can see that the boundedness of $\|u^{(q)}\|_{L^2((0,T);\dot H^{2q+5})} + \|u^{(q+1)}\|_{L^2((0,T);\dot H^{2q+3})}$, is equivalent to $u_q \in L^2((0,T);\dot H^{2q+5}) \cap H^1((0,T);\dot H^{2q+3})$. According to Theorem 6 a sufficient condition for this is given by $f^{(q)} \in L^2((0,T);\dot H^{2q+3})$ and $u_{q,0} \in \dot H^{2q+4}$, which gives

$$\|u_q\|^2_{L^2((0,T);\dot H^{2q+5})} + \|\dot u_q\|^2_{L^2((0,T);\dot H^{2q+3})} \le \|f^{(q)}\|^2_{L^2((0,T);\dot H^{2q+3})} + \|u_{q,0}\|^2_{\dot H^{2q+4}}.$$

We thus achieve the final estimate

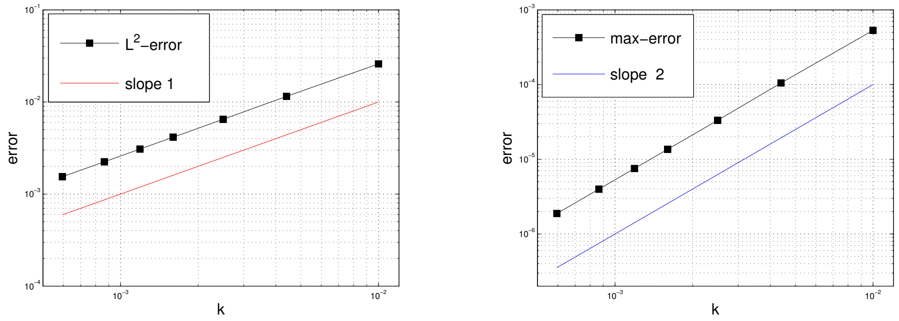
$$\max_{n=1,\ldots,N}\|e_2^{(n)}\|_H \le Ck^{2(q+1)}\Big(\|f^{(q)}\|_{L^2((0,T);\dot H^{2q+3})} + \|u_{q,0}\|_{\dot H^{2q+4}}\Big),$$

which completes the proof. □

Remark 22. *Theorem 21 shows a gain of an extra factor k^{q+1}, which comes from the duality argument and interpolation of degree $q+1$ in the $H^1(I_i;\dot H^s)$-norm (Aubin–Nitsche trick). A similar argument in* [Tho06, Theorem 12.3] *for the* dG(q)*-method yields only a factor k^q because the test functions are of degree q.*

6. Numerical experiments

Since our main concern is about the temporal evolution of the problem, we restrict the numerical tests to the case of one and two spatial dimensions, discretized by means of Lagrangian elements of sufficiently high degree so that the dominating term in the error is given by the temporal part. We test for two different problems



(a) Decay of the error.

(b) Superconvergence.

Figure 1. Numerical tests for Problem (6.1).

the validity of our *a priori* estimates. In both cases we impose the validity of condition (3.3) by taking $k = h^2$.

6.1. **One-dimensional test.**

We test our scheme for the following problem on the space-time domain $(0,1) \times (0,1]$:

$$\begin{aligned} &\dot{u}(\xi,t) - u''(\xi,t) = 2\,\pi\,\sin(2\,\pi\,\xi)\Big(\cos(2\,\pi\,t) + 2\,\pi\,\sin(2\,\pi\,t)\Big), \\ &u(0,t) = u(1,t) = 0, \quad t \in [0,1], \\ &u(\xi,0) = 0, \quad \xi \in [0,1], \end{aligned} \tag{6.1}$$

which has the solution $u(\xi,t) = \sin(2\,\pi\,\xi)\sin(\pi\,t)$.

In Figure 1a we report a log-log graph showing the decay of the error normalized by the norm of the right-hand side, for the numerical solution of Problem (6.1). In Figure 1b we show that the second component of the error satisfies the superconvergence bound stated in Theorem 21.

6.2. **Two-dimensional test.**

We test our scheme for the following problem on the space-time domain $(0,1)^2 \times (0,1]$:

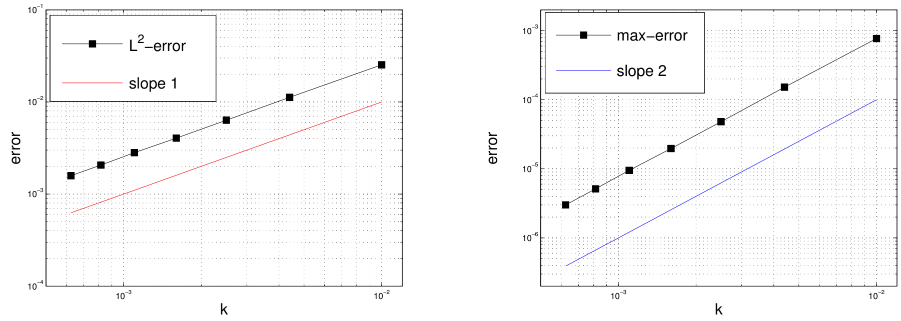
$$\begin{aligned} &\dot{u}(\xi,\eta,t) - \Delta u(\xi,\eta,t) = \pi\,\sin(\pi\,\xi)\sin(\pi\eta)\Big(\cos(\pi\,t) + 2\pi\,\sin(\pi\,t)\Big), \\ &u(0,\eta,t) = u(1,\eta,t) = 0, \quad t \in [0,1],\ \eta \in (0,1), \\ &u(\xi,0,t) = u(\xi,1,t) = 0, \quad t \in [0,1],\ \xi \in (0,1), \\ &u(\xi,\eta,0) = 0, \quad (\xi,\eta) \in [0,1]^2, \end{aligned} \tag{6.2}$$

which has the solution $u(\xi,\eta,t) = \sin(\pi\,\xi)\sin(\pi\,\eta)\sin(\pi\,t)$.

In Figures 2a and 2b we report the analogous results to the ones presented in the one-dimensional case.

6.3. **One-dimensional test, $q = 1$.**

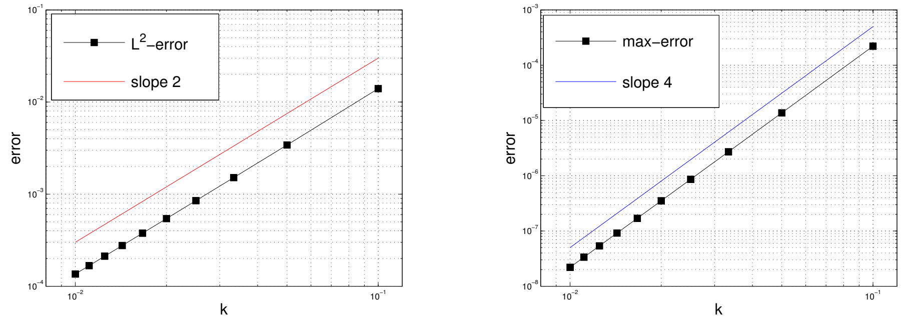
In Figures 3a and 3b we can see the results of convergence and superconvergence when this scheme is used to solve Problem (6.1). The convergence rate is optimal and consistent with our predictions.



(a) Decay of the error.

(b) Superconvergence.

Figure 2. Numerical tests for Problem (6.2).



(a) Decay of the error.

(b) Superconvergence.

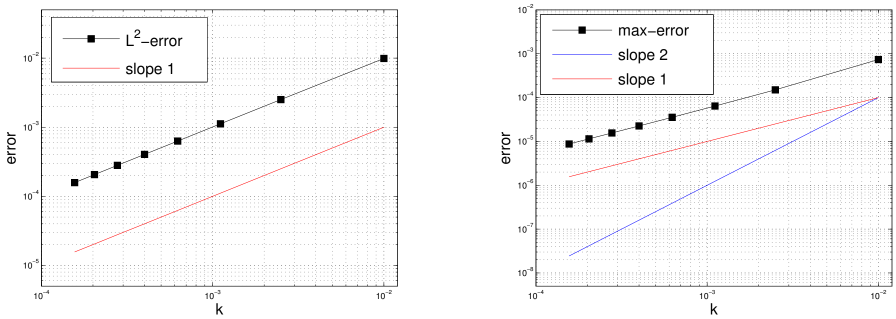
Figure 3. Numerical tests for Problem (6.1), $q = 1$.

6.4. **One-dimensional test, low-regularity.** We investigate the behaviour of the error when the solution is not as smooth as we need to have superconvergence. We pick a problem such that u has the first time-derivative which is square integrable, but not the second one. More in detail, we choose u equal to $|t-0.5|^{\frac{3-\varepsilon}{2}}\sin(\pi\xi)$, where ε is taken equal to 0.1 in the case here investigated.

In Figures 4a and 4b we can see the results of convergence and superconvergence when this scheme is used to solve our problem. The convergence rate for the first component of the error is optimal and consistent with our predictions. In this case the second component of the error does not superconverge and its rate of convergence behaves as the rate of convergence of the first component.

7. Final remarks

In this article we have constructed a numerical scheme that produces a numerical solution under minimal regularity assumptions. The error of the solution has first been bounded in terms of the best possible approximation using the quasi-optimality theory, which does not require any further assumptions of regularity on the solution. The quasi-optimality constant that we obtain depends on the chosen discretization and requires the fulfilment of a certain CFL condition in order to



(a) Decay of the error.

(b) Non-superconvergence.

FIGURE 4. Numerical tests for a problem with low-regular right-hand side.

have stability, consistently with the results in [And12] and [Tan13]. We have shown that our scheme is of first order in time if we assume extra regularity, which means that the scheme is optimal with respect to the norm used to measure the error. Moreover, we have superconvergence at the points constituting the temporal grid, which means that the scheme is of second order in space and time. This further confirms the optimality of our method and its consistency with the known properties of the Crank–Nicolson scheme. Since we do not need extra regularity to prove existence and uniqueness of a discrete solution, our scheme is in particular usable in contexts in which a smooth solution does not exist in the first place, and this can, for example, constitute a novel approach for numerics to stochastic PDEs.

REFERENCES

[And12] R. Andreev. Stability of Space-Time Petrov-Galerkin Discretizations for Parabolic Evolution Equations. PhD thesis, ETH Zürich, Dissertation No. 20842, 2012.

[And13] R. Andreev. Stability of sparse space-time finite element discretizations of linear parabolic evolution equations. *IMA J. Numer. Anal.*, 33(1):242–260, 2013.

[And16] R. Andreev. On long time integration of the heat equation. *Calcolo*, 53(1):19–34, 2016.

[BA72] I. Babuška and A. K. Aziz. Survey lectures on the mathematical foundations of the finite element method. In *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations (Proc. Sympos., Univ. Maryland, Baltimore, Md., 1972)*, pages 1–359. Academic Press, New York, 1972.

[BJ89] I. Babuška and T. Janik. The *h-p* version of the finite element method for parabolic equations. I. The *p*-version in time. *Numer. Methods Partial Differential Equations*, 5(4):363–399, 1989.

[BJ90] I. Babuška and T. Janik. The *h-p* version of the finite element method for parabolic equations. II. The *h-p* version in time. *Numer. Methods Partial Differential Equations*, 6(4):343–369, 1990.

[CDD$^+$14] P. A. Cioica, S. Dahlke, N. Döhring, U. Friedrich, S. Kinzel, F. Lindner, T. Raasch, K. Ritter, and R. L. Schilling. Convergence analysis of spatially adaptive Rothe methods. *Found. Comput. Math.*, 14(5):863–912, 2014.

[CSt11] N. Chegini and R. Stevenson. Adaptive wavelet schemes for parabolic problems: sparse matrices and numerical results. *SIAM J. Numer. Anal.*, 49(1):182–212, 2011.

[EG04] A. Ern and J.L. Guermond. *Theory and Practice of Finite Elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.

[LM16] S. Larsson and M. Molteni. A weak space-time formulation for the linear stochastic heat equation. *Int. J. Appl. Comput. Math.*, 2016. electronic.

[Mol13] C. Mollet. Stability of Petrov-Galerkin discretizations: Application to the space-time weak formulation for parabolic evolution problems. *Comput. Methods. Appl. Math.*, 14(2):231–255, 2013.

[SS13] Ch. Schwab and E. Süli. Adaptive Galerkin approximation algorithms for Kolmogorov equations in infinite dimensions. *Stochastic Partial Differential Equations: Analysis and Computations*, 1(1):483–493, 2013.

[SSt09] Ch. Schwab and R. Stevenson. Space-time adaptive wavelet methods for parabolic evolution problems. *Math. Comp.*, 78(267):1293–1318, 2009.

[Tan13] F. Tantardini. Quasi-Optimality in the Backward Euler-Galerkin Method for Linear Parabolic Problems. Tesi di dottorato, Università degli Studi di Milano, 2013.

[Tho06] V. Thomée. *Galerkin Finite Element Methods for Parabolic Problems*, volume 25 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 2006.

[UP14] K. Urban and A. T. Patera. An improved error bound for reduced basis approximation of linear parabolic problems. *Math. Comp.*, 83(288):1599–1615, 2014.

[XZ03] J. Xu and L. Zikatanov. Some observations on Babuška and Brezzi theories. *Numer. Math.*, 94(1):195–202, 2003.

Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, SE–412 96 Gothenburg, Sweden
E-mail address: stig@chalmers.se

Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, SE–412 96 Gothenburg, Sweden
E-mail address: molteni@chalmers.se