

ON THE INSTABILITIES OF NAIVE FEM DISCRETIZATIONS FOR PDES WITH SIGN-CHANGING COEFFICIENTS

MARTIN HALLA¹  AND FLORIAN OBERENDER^{2,*} 

Abstract. We consider a scalar diffusion equation with a sign-changing coefficient in its principle part. The well-posedness of such problems has already been studied extensively provided that the contrast of the coefficient is non-critical. Furthermore, many different approaches have been proposed to construct stable discretizations thereof, because naive finite element discretizations are expected to be non-reliable in general. However, no explicit example proving the actual instability is known and numerical experiments often do not manifest instabilities in a conclusive manner. To this end we construct an explicit example with a broad family of meshes for which we prove that the corresponding naive finite element discretizations are unstable. On the other hand, we also provide a broad family of (non-symmetric) meshes for which we prove that the discretizations are stable. Together, these two findings explain the results observed in numerical experiments.

Mathematics Subject Classification. 65N12, 65N30, 78M10.

Received April 11, 2025. Accepted December 21, 2025.

1. INTRODUCTION

In this article we consider diffusion equations $-\operatorname{div}(\sigma \nabla u) = f$ with a sign-changing coefficient σ , *i.e.*, the domain Ω admits a decomposition in Ω_{\pm} for which $\pm\sigma|_{\Omega_{\pm}} > 0$. Such equations occur, *e.g.*, for fully homogenized meta materials and their reliable simulation is essential for the development of technical devices, *e.g.*, to control sound [12] and for cloaking [14]. The well-posedness of problems with sign-changing coefficients has been studied extensively by means of the T-coercivity technique [5–7] and is known to depend on the contrast of σ and the smoothness/geometry of the interface $\overline{\Omega_+} \cap \overline{\Omega_-}$. An alternative approach to analyze such PDEs has been investigated in [17, 18] by means of the limiting absorption principle. The stability of standard H^1 -conforming finite element discretizations is known only for sufficiently large contrasts [4] and therefore a variety of approaches to construct stable approximations have been explored, including locally symmetric meshes [8, 15], optimization based methods [1, 2, 10, 11], boundary element methods [20], weakly coercive reformulations [16] and primal-dual stabilizations [9]. However, in contrast to this extensive research on the development of stable discretizations the question if those specialized methods are actually necessary has received much less attention. Indeed, for reasonably small contrasts the error curves (for decreasing mesh sizes) of naive FEMs generally do not look

Keywords and phrases. sign-changing coefficients, meta materials, finite element method, stability analysis.

¹ Institut für Angewandte und Numerische Mathematik, Karlsruher Institut für Technologie, Karlsruhe, Germany.

² Institut für Numerische und Angewandte Mathematik, Georg-August Universität Göttingen, Göttingen, Germany.

*Corresponding author: j.oberender@math.uni-goettingen.de

reliable, but still decrease often with a saw tooth like profile [9]. At other test cases, it can even be hard to trigger some anomalies at all [16]. Actually, the analysis of [8] suggest that meshes being “almost locally symmetric” can be expected to yield stable results, and a mesh generator might produce such meshes without further due. However, without any quantification it is hard to obtain decisive conclusions from this observation.

To study such questions we construct in this article an explicit example with a piece-wise constant coefficient ($\sigma_{\pm} := \sigma|_{\Omega_{\pm}}$) and a discretization by nodal finite elements with uniform rectangular grids in Ω_{\pm} . Thereby the ratio r of the mesh sizes in Ω_{\pm} will play an important role in the analysis and acts in an inverse manner to the contrast $\kappa = \sigma_+/\sigma_-$, *i.e.*, $r\kappa$ will be a crucial quantity. We prove that depending on the parameter range that all considered discretizations are either stable or unstable. For an unstructured mesh we expect that either case can be dominant, which explains the inconclusive observations in numerical experiments.

The remainder of the manuscript is structured as follows. In Section 2 we specify the two considered problems and their discretization. In particular, we consider one problem on an unbounded domain and a second problem on a bounded domain, where the first can be seen as a preparational step to the second. In Section 3 we conduct our stability analysis with Theorems 3.17 and 3.33 as our main results. In Section 4 we present computational examples to confirm our theoretical results.

2. NOTATION AND SETTING

We consider all vector spaces over \mathbb{R} and denote scalar and vectorial L^2 -scalar products over a domain $D \subset \mathbb{R}^l$, $l = 1, 2$ as $\langle \cdot, \cdot \rangle_D$. Let $\mathbb{N} := \{1, 2, \dots\}$, $L > 0$ and consider the domains

$$\begin{aligned} \Omega &:= (-\infty, \infty) \times (0, \pi), & \Omega_- &:= (-\infty, 0) \times (0, \pi), & \Omega_+ &:= (0, \infty) \times (0, \pi), \\ \tilde{\Omega} &:= (-L, L) \times (0, \pi), & \tilde{\Omega}_- &:= (-L, 0) \times (0, \pi), & \tilde{\Omega}_+ &:= (0, L) \times (0, \pi). \end{aligned}$$

On $H_0^1(D)$, $D = \Omega, \tilde{\Omega}$ we work with the scalar product $\langle u, u^\dagger \rangle_{H_0^1(D)} := \langle \nabla u, \nabla u^\dagger \rangle_D$. For the bounded domain $\tilde{\Omega}$ the equivalence of $\langle \cdot, \cdot \rangle_{H_0^1(\tilde{\Omega})}$ to the standard $H^1(\tilde{\Omega})$ -scalar product is well known. For the unbounded domain Ω this equivalence requires a short discussion: Let

$$\theta_m(y) := \sqrt{\frac{2}{\pi}} \sin(my), \quad m \in \mathbb{N}$$

and recall that each $u \in H_0^1(\Omega)$ and $u \in H_0^1(\tilde{\Omega})$ admits a (Fourier series) representation

$$u(x, y) = \sum_{m \in \mathbb{N}} u_m(x) \theta_m(y), \quad u_m(x) := \langle u(x, \cdot), \theta_m \rangle_{(0, \pi)} \quad (1)$$

with

$$\begin{aligned} \|u\|_{H^1(\Omega)}^2 &= \sum_{m \in \mathbb{N}} \|\partial_x u_m\|_{L^2(\mathbb{R})}^2 + (\lambda_m^2 + 1) \|u_m\|_{L^2(\mathbb{R})}^2 & \text{and} \\ \|u\|_{H^1(\tilde{\Omega})}^2 &= \sum_{m \in \mathbb{N}} \|\partial_x u_m\|_{L^2(-L, L)}^2 + (\lambda_m^2 + 1) \|u_m\|_{L^2(-L, L)}^2 \end{aligned}$$

respectively, where

$$\lambda_m := m, \quad m \in \mathbb{N}.$$

It follows that $\|u\|_{H_0^1(\Omega)}^2 \geq \frac{1}{2} \|u\|_{H^1(\Omega)}^2$. As usual, we consider any subspaces of $H_0^1(\Omega)$ and $H_0^1(\tilde{\Omega})$ to be equipped with their inherited scalar product.

Let σ be constant on Ω_{\pm} with values $\sigma_- := \sigma|_{\Omega_-} < 0$ and $\sigma_+ := \sigma|_{\Omega_+} > 0$. Let $f \in L^2(\tilde{\Omega})$ and identify f with its continuation by zero to Ω . We consider the following two model problems:

$$\text{Find } u \in H_0^1(\Omega) \text{ such that } -\operatorname{div}(\sigma \nabla u) = f \text{ in } \Omega; \quad (2a)$$

$$\text{Find } u \in H_0^1(\tilde{\Omega}) \text{ such that } -\operatorname{div}(\sigma \nabla u) = f \text{ in } \tilde{\Omega}, \quad (2b)$$

and their variational formulations:

$$\text{Find } u \in H_0^1(\Omega) \text{ such that } a_{\Omega}(u, u^{\dagger}) = \langle f, u^{\dagger} \rangle_{\Omega} \text{ for all } u^{\dagger} \in H_0^1(\Omega); \quad (3a)$$

$$\text{Find } u \in H_0^1(\tilde{\Omega}) \text{ such that } a_{\tilde{\Omega}}(u, u^{\dagger}) = \langle f, u^{\dagger} \rangle_{\tilde{\Omega}} \text{ for all } u^{\dagger} \in H_0^1(\tilde{\Omega}), \quad (3b)$$

with the corresponding sesquilinear forms

$$a_D(u, u^{\dagger}) := \langle \sigma \nabla u, \nabla u^{\dagger} \rangle_D, \quad D = \Omega, \tilde{\Omega}.$$

Furthermore, let $\mathcal{A} \in \mathcal{L}(H_0^1(\Omega))$, $\tilde{\mathcal{A}} \in \mathcal{L}(H_0^1(\tilde{\Omega}))$ be the associated operators defined by

$$\langle \mathcal{A}u, u^{\dagger} \rangle_{H_0^1(\Omega)} = a_{\Omega}(u, u^{\dagger}) \quad \text{for all } u, u^{\dagger} \in H_0^1(\Omega), \quad (4a)$$

$$\langle \tilde{\mathcal{A}}u, u^{\dagger} \rangle_{H_0^1(\tilde{\Omega})} = a_{\tilde{\Omega}}(u, u^{\dagger}) \quad \text{for all } u, u^{\dagger} \in H_0^1(\tilde{\Omega}). \quad (4b)$$

To specify the approximations of the former problems let P_1 be the space of polynomials in one variable of order lower equal than one. Thence let

$$M \in \mathbb{N}, \quad h_y := \pi/M, \quad y_m := h_y m \quad \text{for } m = 0, \dots, M,$$

and

$$W_{h_y} := \{w \in H_0^1(0, \pi) : w|_{(y_m, y_{m+1})} \in P_1 \text{ for all } m = 0, \dots, M-1\}.$$

To discretize (3a) consider

$$h_{\pm} > 0, \quad x_n := h_+ n, \text{ for } n = 0, 1, \dots; \quad x_n := h_- n, \text{ for } n = -1, -2, \dots,$$

and

$$V_{h_{\pm}} := \{v \in H^1(\mathbb{R}) : v|_{(x_n, x_{n+1})} \in P_1 \text{ for all } n = \dots, -1, 0, 1, \dots\},$$

and to discretize (3b) let

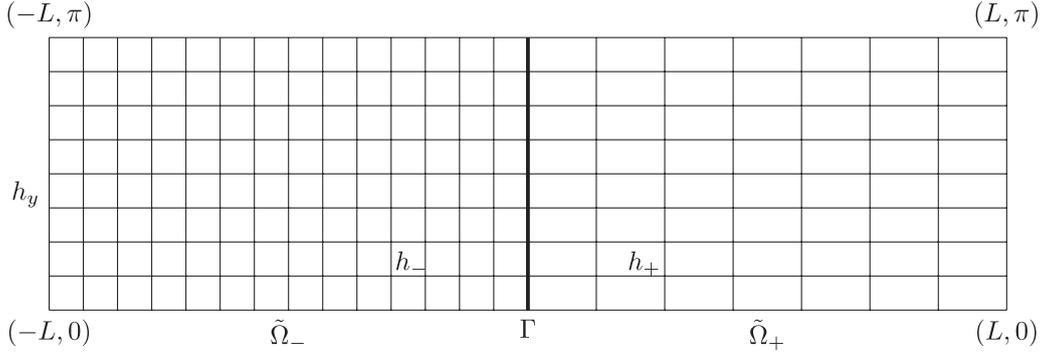
$$N_{\pm} \in \mathbb{N}, \quad h_{\pm} := L/N_{\pm}, \quad \tilde{x}_n := h_+ n, \text{ for } n = 0, \dots, N_+; \quad \tilde{x}_n := h_- n, \text{ for } n = -1, \dots, -N_-,$$

and

$$\tilde{V}_{h_{\pm}} := \{v \in H_0^1(-L, L) : v|_{(\tilde{x}_n, \tilde{x}_{n+1})} \in P_1 \text{ for all } n = -N_-, \dots, N_+ - 1\}.$$

A sketch of the domain with a mesh and mesh parameters is shown in Figure 1. Note that $V_{h_{\pm}}$ is a *single* space which depends on *both* parameters h_+ and h_- . The same applies to $\tilde{V}_{h_{\pm}}$. Consequently we consider the Galerkin approximations of (3a) and (3b) with discrete tensor product spaces $V_{h_{\pm}} \otimes W_{h_y} \subset H_0^1(\Omega)$, $\tilde{V}_{h_{\pm}} \otimes W_{h_y} \subset H_0^1(\tilde{\Omega})$:

$$\text{Find } u \in V_{h_{\pm}} \otimes W_{h_y} \text{ such that } a_{\Omega}(u, u^{\dagger}) = \langle f, u^{\dagger} \rangle_{\Omega} \text{ for all } u^{\dagger} \in V_{h_{\pm}} \otimes W_{h_y}; \quad (5a)$$

FIGURE 1. Sketch of the domain $\tilde{\Omega}$ and an asymmetric mesh.

$$\text{Find } u \in \tilde{V}_{h_{\pm}} \otimes W_{h_y} \text{ such that } a_{\tilde{\Omega}}(u, u^{\dagger}) = \langle f, u^{\dagger} \rangle_{\tilde{\Omega}} \quad \text{for all } u^{\dagger} \in \tilde{V}_{h_{\pm}} \otimes W_{h_y}. \quad (5b)$$

Let $\mathcal{A}_{h_{\pm}, h_y} \in \mathcal{L}(V_{h_{\pm}} \otimes W_{h_y})$ and $\tilde{\mathcal{A}}_{h_{\pm}, h_y} \in \mathcal{L}(\tilde{V}_{h_{\pm}} \otimes W_{h_y})$ be the associated operators defined by

$$\langle \mathcal{A}_{h_{\pm}, h_y} u, u^{\dagger} \rangle_{H_0^1(\Omega)} = a_{\Omega}(u, u^{\dagger}) \quad \text{for all } u, u^{\dagger} \in V_{h_{\pm}} \otimes W_{h_y}, \quad (6a)$$

$$\langle \tilde{\mathcal{A}}_{h_{\pm}, h_y} u, u^{\dagger} \rangle_{H_0^1(\tilde{\Omega})} = a_{\tilde{\Omega}}(u, u^{\dagger}) \quad \text{for all } u, u^{\dagger} \in \tilde{V}_{h_{\pm}} \otimes W_{h_y}. \quad (6b)$$

We will see that the contrast κ (of σ) and the ratios of the meshes sizes

$$\kappa := \frac{\sigma_+}{\sigma_-}, \quad r := \frac{h_+}{h_-}, \quad r_y := \frac{h_y}{h_-},$$

will play a crucial role in the stability analysis. Since problem (3a) is posed on an unbounded domain its discretization (5a) is rather theoretical, but it allows us to perform a very explicit analysis. On the other hand, problem (3b) is posed on a bounded domain and hence its discretization (5b) is computationally feasible, but its analysis is a bit more technical. Indeed, the second setting (3b)/(5b) can be considered as an approximation of (3a)/(5a) by a truncation of the domain Ω to $\tilde{\Omega}$.

During the course of our analysis we will repeatedly use tensor product functions for which we apply the following notation in general:

$$u(x, y) = v(x)w(y).$$

In addition, let $\phi_n \in V_{h_{\pm}}, \tilde{\phi}_n \in \tilde{V}_{h_{\pm}}, \psi_m \in W_{h_y}$ be the nodal basis functions defined by

$$\begin{aligned} \phi_n(x_l) &= \delta_{nl}, & n, l \in \mathbb{Z}, \\ \tilde{\phi}_n(\tilde{x}_l) &= \delta_{nl}, & n, l = -N_- + 1, \dots, N_+ - 1, \\ \psi_m(y_l) &= \delta_{ml}, & m, l = 1, \dots, M - 1. \end{aligned}$$

3. STABILITY ANALYSIS

In this section we investigate the discretizations of (3a) and (3b). Our general strategy for this is as follows. First, we derive orthogonal decompositions of $V_{h_{\pm}} \otimes W_{h_y}$ and $\tilde{V}_{h_{\pm}} \otimes W_{h_y}$ which each consists of two spaces of functions that are supported on either side of the interface and the orthogonal complement of their sum. The stability of the operator only depends on how it acts on this orthogonal complement. We derive explicit

formulas for the eigenvalues of the operator depending on the mesh sizes h_{\pm} and h_y . Finally, we analyze their behaviour depending on the contrast and the ratios r, r_y and show in which cases they are zero, have limit zero or are bounded away from zero. This directly corresponds to (in-)stability of the problem. We do all of this first without discretization, then with discretization only in the x -direction and finally in the x - and y -direction to gradually increase the complexity of the calculations and results.

3.1. Unbounded domain

To study the discretization of (3a) we first analyze its well-posedness and subsequently analyze a semi-discretization before treating the full discretization.

3.1.1. Well-posedness analysis

We start by discussing the well-posed of (2a) to ensure that we have chosen a meaningful problem. Secondly, our analysis will serve as recipe for the forthcoming analysis dealing with the discretizations of (2a). Let

$$X_{\pm} := \{u \in H_0^1(\Omega) : u|_{\Omega_{\mp}} = 0\} \text{ and } X_0 := (X_- \oplus X_+)^{\perp}.$$

Lemma 3.1. *The space $H_0^1(\Omega)$ admits an orthogonal decomposition*

$$H_0^1(\Omega) = X_- \oplus^{\perp} X_0 \oplus^{\perp} X_+$$

where X_0 is spanned by the orthonormal basis $(\frac{1}{\sqrt{2\lambda_m}} e^{-\lambda_m|x|} \otimes \theta_m(y))_{m \in \mathbb{N}}$.

Proof. Let $u \in X_0$. By means of (1) we can write $u = \sum_{m \in \mathbb{N}} u_m \otimes \theta_m$ and it follows that u_m solves $-\partial_x \partial_x u_m + m^2 u_m = 0$ in \mathbb{R}^{\pm} . Thus $u_m|_{\mathbb{R}^+}(x) = c_1^+ e^{-\lambda_m x} + c_2^+ e^{\lambda_m x}$ and $u_m|_{\mathbb{R}^-}(x) = c_1^- e^{\lambda_m x} + c_2^- e^{-\lambda_m x}$ with constants $c_1^{\pm}, c_2^{\pm} \in \mathbb{R}$. Since $u_m \in H^1(\mathbb{R})$ it follows that $c_2^+ = c_2^- = 0$ and the continuity at the origin demands $c_1^+ = c_1^- =: c$, i.e., $u_m(x) = ce^{-\lambda_m|x|}$. The equality $c = \frac{1}{\sqrt{2\lambda_m}}$ follows from a simple computation. \square

Lemma 3.2. *The operator \mathcal{A} is block diagonal with respect to the decomposition of Lemma 3.1. The blocks corresponding to X_- , X_+ and X_0 equal the identity times σ_- , σ_+ and $\frac{\sigma_+ + \sigma_-}{2} = \sigma_- \frac{1+\kappa}{2}$ respectively. Non-verbally: For each $u_-, u_-^{\dagger} \in X_-$, $u_0, u_0^{\dagger} \in X_0$, $u_+, u_+^{\dagger} \in X_+$ it holds that*

$$a_{\Omega}(u_- + u_0 + u_+, u_-^{\dagger} + u_0^{\dagger} + u_+^{\dagger}) = \sigma_- \langle u_-, u_-^{\dagger} \rangle_{H_0^1(\Omega)} + \sigma_- \frac{1+\kappa}{2} \langle u_0, u_0^{\dagger} \rangle_{H_0^1(\Omega)} + \sigma_+ \langle u_+, u_+^{\dagger} \rangle_{H_0^1(\Omega)}.$$

Proof. Since the supports of functions in X_- and X_+ are disjoint it holds that $a_{\Omega}(u_+, u_-) = a_{\Omega}(u_-, u_+) = 0$ for all $u_{\pm} \in X_{\pm}$. We note that

$$a_{\Omega}(u_0, u_{\pm}) = \langle \sigma \nabla u_0, \nabla u_{\pm} \rangle_{\Omega} = \langle \sigma_{\pm} \nabla u_0, \nabla u_{\pm} \rangle_{\Omega_{\pm}} = \sigma_{\pm} \langle \nabla u_0, \nabla u_{\pm} \rangle_{\Omega} = \sigma_{\pm} \langle u_0, u_{\pm} \rangle_{H_0^1(\Omega)} = 0$$

and likewise $a_{\Omega}(u_{\pm}, u_0) = 0$ for all $u_0 \in X_0, u_{\pm} \in X_{\pm}$. It remains to prove $a_{\Omega}(u_0, u_0^{\dagger}) = \sigma_- \frac{1+\kappa}{2} \langle u_0, u_0^{\dagger} \rangle_{H_0^1(\Omega)}$, which follows from plugging in the orthonormal basis functions given in Lemma 3.1. \square

Corollary 3.3. *If $\kappa \neq -1$, then \mathcal{A} is bijective and $\|\mathcal{A}^{-1}\|_{\mathcal{L}(H_0^1(\Omega))} \leq \frac{1}{|\sigma_-| \min\{1, \kappa, \frac{1+\kappa}{2}\}}$.*

Proof. From Lemma 3.2 we know that on each of the three subspaces \mathcal{A} is the identity operator multiplied by the constants σ_- , σ_+ and $\sigma_- \frac{1+\kappa}{2}$ respectively. This implies that these are the only eigenvalues of the operator. It is therefore bijective as long as all of the three constants are nonzero, which is the case if $\kappa \neq -1$, and the norm of the inverse is the reciprocal of the smallest of the three eigenvalues. \square

Henceforth we will only discuss the case $b_{\pm}^{(m)} \neq 0$ and just note that the statements of all lemmas and theorems also hold for the case $b_{\pm}^{(m)} = 0$. Indeed $\mu_{m,1,\pm}, \mu_{m,2,\pm}$ are the roots of the polynomial $\mu \mapsto b_{\pm}^{(m)} \mu^2 + 2a_{\pm}^{(m)} \mu + b_{\pm}^{(m)}$.

Note that it follows from $\mu_{m,1,\pm} \mu_{m,2,\pm} = 1$ and $(a_{\pm}^{(m)})^2 - (b_{\pm}^{(m)})^2 > 0$ that $|\mu_{m,1,\pm}| < |\mu_{m,2,\pm}|$ and therefore

$$|\mu_{m,1,\pm}| < 1 \quad \text{and} \quad |\mu_{m,2,\pm}| > 1. \quad (11)$$

We introduce the abbreviation

$$\mu_{m,\pm} := \mu_{m,1,\pm} \quad (12)$$

and note that per definition $|\mu_{m,\pm}| < 1$. As the next step we further exploit the former representations in the following two lemmas which are in analogy to Section 3.1.1. To avoid misconceptions we emphasize that the n in $\mu_{m,\pm}^{\pm n}$ appearing in Lemma 3.4 is an actual power and not an index.

Lemma 3.4. *The space $\overline{V_{h_{\pm}} \otimes H_0^1(0, \pi)}^{\text{cl}_{H_0^1(\Omega)}}$ admits an orthogonal decomposition*

$$\overline{V_{h_{\pm}} \otimes H_0^1(0, \pi)}^{\text{cl}_{H_0^1(\Omega)}} = \left(\overline{V_{h_{\pm}}^- \otimes H_0^1(0, \pi)}^{\text{cl}_{H_0^1(\Omega)}} \right) \oplus^{\perp} V_{h_{\pm}}^0 \oplus^{\perp} \left(\overline{V_{h_{\pm}}^+ \otimes H_0^1(0, \pi)}^{\text{cl}_{H_0^1(\Omega)}} \right),$$

where $V_{h_{\pm}}^- := \{v \in V_{h_{\pm}} : v|_{\mathbb{R}_+} = 0\}$ and $V_{h_{\pm}}^+ := \{v \in V_{h_{\pm}} : v|_{\mathbb{R}_-} = 0\}$. The subspace $V_{h_{\pm}}^0$ is spanned by the orthonormal basis $(v_m(x) \otimes \theta_m(y))_{m \in \mathbb{N}}$, where

$$v_m(x) := \frac{1}{\sqrt{b_-^{(m)} \mu_{m,-} + a_-^{(m)} + a_+^{(m)} + b_+^{(m)} \mu_{m,+}}} \left(\phi_0(x) + \sum_{n \in \mathbb{N}} \mu_{m,+}^n \phi_n(x) + \sum_{n \in \mathbb{N}} \mu_{m,-}^n \phi_{-n}(x) \right).$$

Proof. Let $v \in \left(\overline{V_{h_{\pm}}^- \otimes H_0^1(0, \pi)}^{\text{cl}_{H_0^1(\Omega)}} \oplus \overline{V_{h_{\pm}}^+ \otimes H_0^1(0, \pi)}^{\text{cl}_{H_0^1(\Omega)}} \right)^{\perp}$. By means of (1) we can write $v = \sum_{m \in \mathbb{N}} v_m \otimes \theta_m$. We can write $v_m = \sum_{n \in \mathbb{Z}} \beta_n^{(m)} \phi_n$. The span of the functions $\phi_n \otimes \theta_m$ for $m, n \in \mathbb{Z}$, $n \neq 0$ is dense in $\left(\overline{V_{h_{\pm}}^- \otimes H_0^1(0, \pi)}^{\text{cl}_{H_0^1(\Omega)}} \oplus \overline{V_{h_{\pm}}^+ \otimes H_0^1(0, \pi)}^{\text{cl}_{H_0^1(\Omega)}} \right)$. By orthogonality we then have

$$\begin{aligned} 0 &= \langle v, \phi_{\pm n} \otimes \theta_m \rangle_{H_0^1(\Omega)} = \langle \partial_x v_m, \partial_x \phi_{\pm n} \rangle_{\mathbb{R}} + \lambda_m^2 \langle v_m, \phi_{\pm n} \rangle_{\mathbb{R}} \\ &= \left\langle \beta_{\pm n-1}^{(m)} \partial_x \phi_{\pm n-1}, \partial_x \phi_{\pm n} \right\rangle_{\mathbb{R}} + \left\langle \beta_{\pm n}^{(m)} \partial_x \phi_{\pm n}, \partial_x \phi_{\pm n} \right\rangle_{\mathbb{R}} + \left\langle \beta_{\pm n+1}^{(m)} \partial_x \phi_{\pm n+1}, \partial_x \phi_{\pm n} \right\rangle_{\mathbb{R}} \\ &\quad + \lambda_m^2 \left(\left\langle \beta_{\pm n-1}^{(m)} \phi_{\pm n-1}, \phi_{\pm n} \right\rangle_{\mathbb{R}} + \left\langle \beta_{\pm n}^{(m)} \phi_{\pm n}, \phi_{\pm n} \right\rangle_{\mathbb{R}} + \left\langle \beta_{\pm n+1}^{(m)} \phi_{\pm n+1}, \phi_{\pm n} \right\rangle_{\mathbb{R}} \right) \\ &= \sigma_{\pm} \left(\beta_{\pm n-1}^{(m)} b_{\pm}^{(m)} + 2\beta_{\pm n}^{(m)} a_{\pm}^{(m)} + \beta_{\pm n+1}^{(m)} b_{\pm}^{(m)} \right). \end{aligned}$$

Solving this three-term recurrence relation and recalling (12), (11) we obtain that

$$\beta_n^{(m)} = \beta_0^{(m)} \mu_{m,+}^n \quad \text{and} \quad \beta_{-n}^{(m)} = \beta_0^{(m)} \mu_{m,-}^n \quad \forall n \in \mathbb{N}.$$

Finally we compute the normalization constant by

$$\begin{aligned} 1 &= \langle \partial_x v_m, \partial_x v_m \rangle_{\mathbb{R}} + \lambda_m^2 \langle v_m, v_m \rangle_{\mathbb{R}} = \sum_{n \in \mathbb{Z}} \beta_n^{(m)} (\langle \partial_x \phi_n, \partial_x v_m \rangle_{\mathbb{R}} + \lambda_m^2 \langle \phi_n, v_m \rangle_{\mathbb{R}}) \\ &= \beta_0^{(m)} (\langle \partial_x \phi_0, \partial_x v_m \rangle_{\mathbb{R}} + \lambda_m^2 \langle \phi_0, v_m \rangle_{\mathbb{R}}) \\ &= \left(\beta_0^{(m)} \right)^2 \left(\mu_{m,-} b_-^{(m)} + a_-^{(m)} + a_+^{(m)} + \mu_{m,+} b_+^{(m)} \right), \end{aligned}$$

i.e.,

$$\beta_0^{(m)} = \frac{1}{\sqrt{\mu_{m,-} b_-^{(m)} + a_-^{(m)} + a_+^{(m)} + \mu_{m,+} b_+^{(m)}}}.$$

Note that this calculation also ensures that $v_m \otimes \theta_m$ has finite $H_0^1(\Omega)$ -norm, i.e., $v_m \otimes \theta_m \in H_0^1(\Omega)$ is well defined. \square

Lemma 3.5. *The operator \mathcal{A}_{h_\pm} is block diagonal with respect to the orthogonal decomposition given in Lemma 3.4. The blocks corresponding to $\overline{V_{h_\pm}^- \otimes H_0^1(0, \pi)}^{\text{cl}_{H_0^1(\Omega)}}$ and $\overline{V_{h_\pm}^+ \otimes H_0^1(0, \pi)}^{\text{cl}_{H_0^1(\Omega)}}$ equal the identity times σ_- and σ_+ respectively. The block corresponding to $V_{h_\pm}^0$ is diagonal with respect to the basis given in Lemma 3.4 and the diagonal entries are given by*

$$d_m := \frac{\sigma_- b_-^{(m)} \mu_{m,-} + \sigma_- a_-^{(m)} + \sigma_+ a_+^{(m)} + \sigma_+ b_+^{(m)} \mu_{m,+}}{b_-^{(m)} \mu_{m,-} + a_-^{(m)} + a_+^{(m)} + b_+^{(m)} \mu_{m,+}}, \quad m \in \mathbb{N}. \quad (13)$$

Non-verbally: For each $u_-, u_-^\dagger \in \overline{V_{h_\pm}^- \otimes H_0^1(0, \pi)}^{\text{cl}_{H_0^1(\Omega)}}$, $u_0, u_0^\dagger \in V_{h_\pm}^0$, $u_+, u_+^\dagger \in \overline{V_{h_\pm}^+ \otimes H_0^1(0, \pi)}^{\text{cl}_{H_0^1(\Omega)}}$ and $u_0 = \sum_{m \in \mathbb{N}} \beta_m v_m \otimes \theta_m$, $u_0^\dagger = \sum_{m \in \mathbb{N}} \beta_m^\dagger v_m \otimes \theta_m$, $(\beta_m)_{m \in \mathbb{N}}, (\beta_m^\dagger)_{m \in \mathbb{N}} \in \ell^2(\mathbb{N})$ it holds that

$$a_\Omega(u_- + u_0 + u_+, u_-^\dagger + u_0^\dagger + u_+^\dagger) = \sigma_- \langle u_-, u_-^\dagger \rangle_{H_0^1(\Omega)} + \sum_{m \in \mathbb{N}} d_m \beta_m \beta_m^\dagger + \sigma_+ \langle u_+, u_+^\dagger \rangle_{H_0^1(\Omega)}.$$

Proof. The arguments are essentially the same as in the continuous case (see Lem. 3.2) where we now use Lemma 3.4 instead of Lemma 3.1. With this we directly get $a_\Omega(u_-, u_-^\dagger) = a_\Omega(u_+, u_+^\dagger) = 0$ and that \mathcal{A}_{h_\pm} is the identity times σ_+ and σ_- on $\overline{V_{h_\pm}^- \otimes H_0^1(0, \pi)}^{\text{cl}_{H_0^1(\Omega)}}$ and $\overline{V_{h_\pm}^+ \otimes H_0^1(0, \pi)}^{\text{cl}_{H_0^1(\Omega)}}$ respectively. By the orthogonality of the decomposition we also get $a_\Omega(u_0, u_0^\dagger) = \sigma_\pm \langle u_0, u_0^\dagger \rangle_{H_0^1(\Omega)} = 0$ as in the continuous case. It remains to show, that operator is diagonal on $V_{h_\pm}^0$ with the claimed values. That it is indeed diagonal follows directly from the decomposition in (9). To compute the values we use the same calculation as for the normalization constant and get

$$\begin{aligned} a_\Omega(v_m \otimes \theta_m, v_m \otimes \theta_m) &= \frac{\sum_{n \in \mathbb{N}} a_\Omega(\mu_{m,-}^n \phi_{-n} \otimes \theta_m, v_m \otimes \theta_m) + \sum_{n \in \mathbb{N}} a_\Omega(\mu_{m,+}^n \phi_n \otimes \theta_m, v_m \otimes \theta_m)}{\sqrt{b_-^{(m)} \mu_{m,-} + a_-^{(m)} + a_+^{(m)} + b_+^{(m)} \mu_{m,+}}} \\ &\quad + \frac{a_\Omega(\phi_0 \otimes \theta_m, v_m \otimes \theta_m)}{\sqrt{b_-^{(m)} \mu_{m,-} + a_-^{(m)} + a_+^{(m)} + b_+^{(m)} \mu_{m,+}}} \\ &= \frac{a_\Omega(\phi_0 \otimes \theta_m, v_m \otimes \theta_m)}{\sqrt{b_-^{(m)} \mu_{m,-} + a_-^{(m)} + a_+^{(m)} + b_+^{(m)} \mu_{m,+}}} \\ &= \frac{\sigma_- b_-^{(m)} \mu_{m,-} + \sigma_- a_-^{(m)} + \sigma_+ a_+^{(m)} + \sigma_+ b_+^{(m)} \mu_{m,+}}{b_-^{(m)} \mu_{m,-} + a_-^{(m)} + a_+^{(m)} + b_+^{(m)} \mu_{m,+}}, \end{aligned}$$

where we exploited the orthogonality properties of v_m . \square

We observe that in contrast to Lemma 3.2 the block corresponding to $V_{h_\pm}^0$ is not a multiple of the identity, but still diagonal.

Up to this point very similar calculations are possible for higher order finite element discretizations. By first eliminating higher order terms a matrix like $\mathbf{A}^{(m)}$ with a tri-diagonal structure can be derived. For higher order

discretizations the entries $b_{\pm}^{(m)}$, $a_{\pm}^{(m)}$ then depend on the mesh size in increasingly complicated ways which transfers to d_m and make the following more explicit computations impossible.

To analyze the diagonal entries d_m we introduce the function

$$f_{\kappa,r}(t) := \frac{1 + \frac{\kappa\sqrt{r^2t^2+12}}{\sqrt{t^2+12}}}{1 + \frac{\sqrt{r^2t^2+12}}{\sqrt{t^2+12}}}.$$

Lemma 3.6. *The diagonal entries d_m defined in (13) satisfy $d_m = \sigma_- f_{\kappa,r}(\lambda_m h_-)$.*

Proof. To start with, plugging in the definitions (12), (10) of $\mu_{m,\pm}$ and $b_{\pm}^{(m)}$, $a_{\pm}^{(m)}$ respectively yields that

$$b_{\pm}^{(m)} \mu_{m,\pm} + a_{\pm}^{(m)} = \sqrt{\left(a_{\pm}^{(m)}\right)^2 - \left(b_{\pm}^{(m)}\right)^2} = \lambda_m \sqrt{1 + \frac{1}{12} \lambda_m^2 h_{\pm}^2}.$$

Inserting this into the definition (13) of d_m we obtain that

$$\begin{aligned} d_m &= \frac{\sigma_- \left(b_-^{(m)} \mu_{m,-} + a_-^{(m)}\right) + \sigma_+ \left(a_+^{(m)} + b_+^{(m)} \mu_{m,+}\right)}{\left(b_-^{(m)} \mu_{m,-} + a_-^{(m)}\right) + \left(a_+^{(m)} + b_+^{(m)} \mu_{m,+}\right)} = \frac{\sigma_- \sqrt{1 + \frac{1}{12} \lambda_m^2 h_-^2} + \sigma_+ \sqrt{1 + \frac{1}{12} \lambda_m^2 h_+^2}}{\sqrt{1 + \frac{1}{12} \lambda_m^2 h_-^2} + \sqrt{1 + \frac{1}{12} \lambda_m^2 h_+^2}} \\ &= \sigma_- \frac{1 + \kappa \frac{\sqrt{12 + \lambda_m^2 h_+^2}}{\sqrt{12 + \lambda_m^2 h_-^2}}}{1 + \frac{\sqrt{12 + \lambda_m^2 h_+^2}}{\sqrt{12 + \lambda_m^2 h_-^2}}} = \sigma_- \frac{1 + \kappa \frac{\sqrt{12 + \lambda_m^2 h_-^2 r^2}}{\sqrt{12 + \lambda_m^2 h_-^2}}}{1 + \frac{\sqrt{12 + \lambda_m^2 h_-^2 r^2}}{\sqrt{12 + \lambda_m^2 h_-^2}}} = \sigma_- f_{\kappa,r}(\lambda_m h_-), \end{aligned}$$

where we recall that $r = h_+/h_-$. □

In the following lemmas we analyze the function $f_{\kappa,r}$.

Lemma 3.7. *If one of the following two conditions*

$$\underbrace{|\kappa| < 1 \text{ and } r|\kappa| > 1}_{(14a)} \quad \text{or} \quad \underbrace{|\kappa| > 1 \text{ and } r|\kappa| < 1}_{(14b)} \tag{14}$$

is satisfied, then the only root of $f_{\kappa,r}$ in $[0, +\infty)$ is

$$t_{\kappa,r} := \sqrt{\frac{12(1 - \kappa^2)}{\kappa^2 r^2 - 1}}.$$

Proof. Since the denominator of $f_{\kappa,r}$ ranges for $t \geq 0$ between 2 and $1 + r$, it suffices to analyze its nominator. Because we only consider non-negative t and negative κ we get

$$\begin{aligned} 0 = 1 + \frac{\kappa\sqrt{r^2t^2+12}}{\sqrt{t^2+12}} &\Leftrightarrow \sqrt{t^2+12} = |\kappa|\sqrt{r^2t^2+12} \Leftrightarrow t^2+12 = \kappa^2(r^2t^2+12) \\ &\Leftrightarrow t^2(1 - \kappa^2 r^2) = 12(\kappa^2 - 1). \end{aligned}$$

The condition (14) now guarantees that $\kappa^2 - 1$ and $1 - \kappa^2 r^2$ have the same sign so the only solution is $t_{\kappa,r} = \sqrt{\frac{12(1 - \kappa^2)}{\kappa^2 r^2 - 1}}$. □

Lemma 3.8. *If (14) is satisfied, then $\lim_{h_- \rightarrow 0^+} \inf_{m \in \mathbb{N}} |f_{\kappa,r}(\lambda_m h_-)| = 0$.*

Proof. Write $h_- = \frac{1}{l+\epsilon} t_{\kappa,r}$ with $l \in \mathbb{N}_0$ and $\epsilon \in [0, 1)$. Choose $m = l$, recall that $\lambda_m = m$ and exploit that $\lim_{l \rightarrow +\infty} \frac{l}{l+\epsilon} \rightarrow 1$ uniformly in $\epsilon \in [0, 1)$. Apply Lemma 3.7 and the continuity of $f_{\kappa,r}$. \square

Lemma 3.9. *If one of the following two conditions*

$$\underbrace{|\kappa| < 1 \text{ and } r|\kappa| < 1}_{(15a)} \quad \text{or} \quad \underbrace{|\kappa| > 1 \text{ and } r|\kappa| > 1}_{(15b)} \quad (15)$$

is satisfied, then $\inf_{t \geq 0} |f_{\kappa,r}(t)| \geq \min\{|\frac{1+\kappa}{2}|, |\frac{1+r\kappa}{1+r}|\}$.

Proof. We note that $f'_{\kappa,r}(t) = \frac{12(\kappa-1)(r^2-1)t}{\sqrt{t^2+12}\sqrt{r^2t^2+12}(\sqrt{r^2t^2+12}+\sqrt{t^2+12})^2}$ and hence $f_{\kappa,r}(t)$ is a monotone function. This implies

$$f_{\kappa,r}(t) \in \left[\min\left\{f_{\kappa,r}(0), \lim_{s \rightarrow +\infty} f_{\kappa,r}(s)\right\}, \max\left\{f_{\kappa,r}(0), \lim_{s \rightarrow +\infty} f_{\kappa,r}(s)\right\} \right] \quad \forall t \in [0, +\infty).$$

Computing these values we get $f_{\kappa,r}(t) \in [\min\{\frac{1+\kappa}{2}, \frac{1+r\kappa}{1+r}\}, \max\{\frac{1+\kappa}{2}, \frac{1+r\kappa}{1+r}\}]$. The conditions now ensure that both values have the same sign so the absolute value of $f_{\kappa,r}(t)$ is always bigger than the minimal absolute value occurring in one of the bounds. \square

Now we are in the position to conclude our analysis of the semi discretization (7) in the following theorem.

Theorem 3.10. *If (15) is satisfied, then $\mathcal{A}_{h_{\pm}}^{-1}$ exists and satisfies*

$$\left\| \mathcal{A}_{h_{\pm}}^{-1} \right\|_{\mathcal{L}(\overline{V_{h_{\pm}} \otimes H_0^1(0,\pi)}^{c1} H_0^1(\Omega))} \leq \frac{1}{|\sigma_-| \min\left\{1, |\kappa|, \left|\frac{1+\kappa}{2}\right|, \left|\frac{1+r\kappa}{1+r}\right|\right\}}$$

for all $h_-, h_+ = rh_- > 0$. Contrary, if (14) is satisfied, then

$$\lim_{\substack{h_- \rightarrow 0+ \\ h_+ = rh_-}} \left\| \mathcal{A}_{h_{\pm}}^{-1} \right\|_{\mathcal{L}(\overline{V_{h_{\pm}} \otimes H_0^1(0,\pi)}^{c1} H_0^1(\Omega))} = +\infty$$

(where we define $\|\mathcal{A}_{h_{\pm}}^{-1}\|_{\mathcal{L}(\overline{V_{h_{\pm}} \otimes H_0^1(0,\pi)}^{c1} H_0^1(\Omega))} := +\infty$, if $\mathcal{A}_{h_{\pm}}^{-1}$ does not exist), and in particular: $\mathcal{A}_{h_{\pm}}$ admits a nontrivial kernel for each

$$h_- = \frac{1}{\lambda_m} t_{\kappa,r} = \frac{1}{m} \sqrt{\frac{12(1-\kappa^2)}{\kappa^2 r^2 - 1}}, \quad h_+ = rh_-, \quad m \in \mathbb{N}.$$

Proof. By Lemma 3.5 we know that the eigenvalues of $\mathcal{A}_{h_{\pm}}$ are σ_+ , σ_- and d_m for $m \in \mathbb{N}$. From Lemma 3.8 we get a lower bound on the absolute values of d_m which implies the first statement of the theorem. Then from Lemma 3.9 it follows that such a bound does not exist in the other case which implies the second statement and finally Lemma 3.7 gives us the precise values where we have a zero eigenvalue which implies a nontrivial kernel. \square

3.1.3. Full discretization

Next we consider the full discretization (5a) by means of the Galerkin spaces $V_{h_{\pm}} \otimes W_{h_y}$. To this end much of the analysis of Section 3.1.2 can be repeated, but we have to replace the orthogonal basis $(\theta_m)_{m \in \mathbb{N}}$ of $H_0^1(0, \pi)$ by a suitable discrete orthogonal basis $(\hat{\theta}_m)_{m=1, \dots, M-1}$ of W_{h_y} , and as a general rule we denote respective modified quantities by the same symbol as previously used but with an additional hat. Hence we consider the following eigenvalue problem:

$$\text{Find } (\tau, w) \in \mathbb{R}^+ \times W_{h_y} \setminus \{0\} \text{ such that } \langle \partial_y w, \partial_y w^\dagger \rangle_{(0,2\pi)} = \tau^2 \langle w, w^\dagger \rangle_{(0,\pi)} \quad \forall w^\dagger \in W_{h_y}. \quad (16)$$

To solve this problem we define

$$\left(\mathbf{B}^{(\tau)}\right)_{m',m} := \langle \partial_y \psi_m, \partial_y \psi_{m'} \rangle_{(0,\pi)} - \tau^2 \langle \psi_m, \psi_{m'} \rangle_{(0,\pi)}, \quad m, m' = 1, \dots, M-1.$$

Then we use that τ is an eigenvalue if and only if $\mathbf{B}^{(\tau)}$ has a zero eigenvalue. It holds that

$$\mathbf{B}^{(\tau)} = \begin{pmatrix} 2a & b & & & \\ b & 2a & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 2a & b \\ & & & b & 2a \end{pmatrix}, \quad a := \frac{1}{h_y} - \tau^2 \frac{1}{3} h_y, \quad b := -\frac{1}{h_y} - \tau^2 \frac{1}{6} h_y.$$

It follows that [3, 13]

$$\tau_m^2 = \frac{6}{h_y^2} \frac{1 - \cos(h_y m)}{2 + \cos(h_y m)}, \quad m = 1, \dots, M-1$$

with respective $L^2(0, \pi)$ -normalized eigenfunctions

$$\hat{\theta}_m(y) := c \sum_{l=1}^{M-1} \sin(m h_y l) \psi_l(y), \quad c := \left\| \sum_{l=1}^{M-1} \sin(m h_y l) \psi_l \right\|_{L^2(0,\pi)}^{-1}, \quad m = 1, \dots, M-1.$$

Note that these explicit calculations are only possible for the case of first order finite elements as higher orders require the analytical solution of generalized eigenvalue problems for more general Toeplitz matrices, which leads to severe complications.

We introduce

$$\hat{\lambda}_{m,r_y,h_-} := \frac{\sqrt{6}}{r_y h_-} \sqrt{\frac{1 - \cos(r_y h_- m)}{2 + \cos(r_y h_- m)}} = \frac{\sqrt{6}}{h_y} \sqrt{\frac{1 - \cos(h_y m)}{2 + \cos(h_y m)}}, \quad m = 1, \dots, M-1.$$

Hence we define respective modified quantities

$$\begin{aligned} \hat{a}_{\pm}^{(m)} &:= \frac{1}{h_{\pm}} + \hat{\lambda}_{m,r_y,h_-}^2 \frac{1}{3} h_{\pm}, \\ \hat{b}_{\pm}^{(m)} &:= -\frac{1}{h_{\pm}} + \hat{\lambda}_{m,r_y,h_-}^2 \frac{1}{6} h_{\pm}, \\ \hat{\mu}_{m,1,\pm} &:= \frac{1}{\hat{b}_{\pm}^{(m)}} \left(-\hat{a}_{\pm}^{(m)} + \sqrt{\left(\hat{a}_{\pm}^{(m)}\right)^2 - \left(\hat{b}_{\pm}^{(m)}\right)^2} \right), \\ \hat{\mu}_{m,2,\pm} &:= \frac{1}{\hat{b}_{\pm}^{(m)}} \left(-\hat{a}_{\pm}^{(m)} - \sqrt{\left(\hat{a}_{\pm}^{(m)}\right)^2 - \left(\hat{b}_{\pm}^{(m)}\right)^2} \right), \\ \hat{\mu}_{m,\pm} &:= \begin{cases} \hat{\mu}_{m,1,\pm}, & \hat{b}_{\pm}^{(m)} \neq 0, \\ 0, & \hat{b}_{\pm}^{(m)} = 0, \end{cases} \\ \hat{d}_m &:= \frac{\sigma_- \hat{b}_-^{(m)} \hat{\mu}_{m,-} + \sigma_- \hat{a}_-^{(m)} + \sigma_+ \hat{a}_+^{(m)} + \sigma_+ \hat{b}_+^{(m)} \hat{\mu}_{m,+}}{\hat{b}_-^{(m)} \hat{\mu}_{m,-} + \hat{a}_-^{(m)} + \hat{a}_+^{(m)} + \hat{b}_+^{(m)} \hat{\mu}_{m,+}}, \\ \hat{v}_m(x) &:= \frac{1}{\sqrt{\hat{b}_-^{(m)} \hat{\mu}_{m,-} + \hat{a}_-^{(m)} + \hat{a}_+^{(m)} + \hat{b}_+^{(m)} \hat{\mu}_{m,+}}} \left(\phi_0(x) + \sum_{n \in \mathbb{Z}^+} \hat{\mu}_{m,+}^n \phi_n(x) + \sum_{n \in \mathbb{Z}^-} \hat{\mu}_{m,-}^{-n} \phi_n(x) \right) \end{aligned}$$

for $m = 1, \dots, M - 1$.

The forthcoming two lemmas follow in analogy to Section 3.1.2.

Lemma 3.11. *The space $V_{h_\pm} \otimes W_{h_y}$ admits an orthogonal decomposition*

$$V_{h_\pm} \otimes W_{h_y} = \left(V_{h_\pm}^- \otimes W_{h_y} \right) \oplus^\perp \hat{V}_{h_\pm}^0 \oplus^\perp \left(V_{h_\pm}^+ \otimes W_{h_y} \right),$$

where $V_{h_\pm}^- = \{v \in V_{h_\pm} : v|_{\mathbb{R}_+} = 0\}$ and $V_{h_\pm}^+ = \{v \in V_{h_\pm} : v|_{\mathbb{R}_-} = 0\}$. The subspace $\hat{V}_{h_\pm}^0$ is spanned by the orthonormal basis $(\hat{v}_m(x) \otimes \hat{\theta}_m(y))_{m=1, \dots, M-1}$.

Proof. The proof can be obtained by following the steps of the proof of Lemma 3.4 one-to-one. \square

Lemma 3.12. *The operator \mathcal{A}_{h_\pm, h_y} is block diagonal with respect to the orthogonal decomposition given in Lemma 3.11. The blocks corresponding to $V_{h_\pm}^- \otimes W_{h_y}$ and $V_{h_\pm}^+ \otimes W_{h_y}$ equal the identity times σ_- and σ_+ respectively. The block corresponding to $\hat{V}_{h_\pm}^0$ is diagonal with respect to the basis given in Lemma 3.11 and the diagonal entries are given by \hat{d}_m . Non-verbally: For each $u_-, u_-^\dagger \in V_{h_\pm}^- \otimes W_{h_y}$, $u_0, u_0^\dagger \in \hat{V}_{h_\pm}^0$, $u_+, u_+^\dagger \in V_{h_\pm}^+ \otimes W_{h_y}$ and $u_0 = \sum_{m=1}^{M-1} \beta_m \hat{v}_m \otimes \hat{\theta}_m$, $u_0^\dagger = \sum_{m=1}^{M-1} \beta_m^\dagger \hat{v}_m \otimes \hat{\theta}_m$, $(\beta_m)_{m=1}^{M-1}, (\beta_m^\dagger)_{m=1}^M \in \mathbb{R}^{M-1}$ it holds that*

$$a_\Omega \left(u_- + u_0 + u_+, u_-^\dagger + u_0^\dagger + u_+^\dagger \right) = \sigma_- \left\langle u_-, u_-^\dagger \right\rangle_{H_0^1(\Omega)} + \sum_{m=1}^{M-1} \hat{d}_m \beta_m \beta_m^\dagger + \sigma_+ \left\langle u_+, u_+^\dagger \right\rangle_{H_0^1(\Omega)}.$$

Proof. The proof can be obtained by following the steps of the proof of Lemma 3.5 one-to-one. \square

To analyze the diagonal entries \hat{d}_m we define

$$\mathfrak{h}_{r_y}(s) := \sqrt{\frac{6}{r_y^2} \frac{1 - \cos(s)}{2 + \cos(s)}}$$

and introduce the following lemmas.

Lemma 3.13. *It holds that $\hat{\lambda}_{m, r_y, h_-} h_- = \mathfrak{h}_{r_y}(r_y m h_-)$ and $\hat{d}_m = \sigma_- \mathfrak{f}_{\kappa, r}(\mathfrak{h}_{r_y}(r_y m h_-))$ for $m = 1, \dots, M - 1$.*

Proof. An elementary computation shows

$$\hat{\lambda}_{m, r_y, h_-}^2 h_-^2 = \frac{6}{r_y^2 h_-^2} \frac{1 - \cos(r_y h_- m)}{2 + \cos(r_y h_- m)} h_-^2 = \mathfrak{h}_{r_y}(r_y m h_-)^2.$$

Then as in Lemma 3.6 it follows that $\hat{d}_m = \sigma_- \mathfrak{f}_{\kappa, r}(\hat{\lambda}_{m, r_y, h_-} h_-)$ and combining these two identities yields that $\hat{d}_m = \sigma_- \mathfrak{f}_{\kappa, r}(\mathfrak{h}_{r_y}(r_y m h_-))$. \square

Lemma 3.14. *If one of the following two conditions*

$$\underbrace{|\kappa| < 1 \text{ and } r^2 \kappa^2 > 1 + r_y^2 (1 - \kappa^2)}_{(17a)} \quad \text{or} \quad \underbrace{|\kappa| > 1 \text{ and } r^2 \kappa^2 < 1 + r_y^2 (1 - \kappa^2)}_{(17b)} \quad (17)$$

is satisfied, then the problem to find $s \in (0, \pi]$ such that $\mathfrak{h}_{r_y}(s) = t_{\kappa,r}$ admits the unique solution

$$s_{\kappa,r,r_y} := \arccos\left(\frac{1 - \frac{r_y^2 t_{\kappa,r}^2}{3}}{1 + \frac{r_y^2 t_{\kappa,r}^2}{6}}\right) = \arccos\left(1 + \frac{6r_y^2(1 - \kappa^2)}{(1 - \kappa^2 r^2) - 2r_y^2(1 - \kappa^2)}\right).$$

Proof. First, one can check, that the inequalities guarantee that $\frac{6r_y^2(1 - \kappa^2)}{(1 - \kappa^2 r^2) - 2r_y^2(1 - \kappa^2)} \in (-2, 0)$ so s_{κ,r,r_y} is well defined. Now we compute

$$\mathfrak{h}_{r_y}(s_{\kappa,r,r_y})^2 = \frac{6}{r_y^2} \frac{1 - \cos(s_{\kappa,r,r_y})}{2 + \cos(s_{\kappa,r,r_y})} = \frac{6}{r_y^2} \frac{1 - \frac{1 - \frac{r_y^2 t_{\kappa,r}^2}{3}}{1 + \frac{r_y^2 t_{\kappa,r}^2}{6}}}{2 + \frac{1 - \frac{r_y^2 t_{\kappa,r}^2}{3}}{1 + \frac{r_y^2 t_{\kappa,r}^2}{6}}} = \frac{6}{r_y^2} \frac{3 \frac{r_y^2 t_{\kappa,r}^2}{6}}{3} = t_{\kappa,r}^2.$$

□

Before we formulate the next Lemma 3.15, let us recall that $h_- = \frac{h_y}{r_y} = \frac{\pi}{r_y M}$.

Lemma 3.15. *If (17) is satisfied, then $\lim_{M \rightarrow +\infty} \inf_{m \in \{1, \dots, M-1\}} |\mathfrak{f}_{\kappa,r}(\mathfrak{h}_{r_y}(\frac{m\pi}{M}))| = 0$.*

Proof. Because the rational numbers are dense in the real numbers, for each $\epsilon > 0$ there exists $M_\epsilon \in \mathbb{N}$ such that for all $M \in \mathbb{N}, M > M_\epsilon$ there exists $m \in \{1, \dots, M-1\}$ such that $|\frac{m}{M}\pi - s_{\kappa,r,r_y}| < \epsilon$. The theorem now follows from the continuity of $\mathfrak{f}_{\kappa,r} \circ \mathfrak{h}_{r_y}$ and $\mathfrak{f}_{\kappa,r}(\mathfrak{h}_{r_y}(s_{\kappa,r,r_y})) = 0$. □

Lemma 3.16. *If one of the following two conditions*

$$\underbrace{|\kappa| < 1 \text{ and } r^2 \kappa^2 < 1 + r_y^2(1 - \kappa^2)}_{(18a)} \quad \text{or} \quad \underbrace{|\kappa| > 1 \text{ and } r^2 \kappa^2 > 1 + r_y^2(1 - \kappa^2)}_{(18b)} \quad (18)$$

is satisfied, then $\inf_{s \in \mathbb{R}} |\mathfrak{f}_{\kappa,r}(\mathfrak{h}_{r_y}(s))| \geq \min\left\{\left|\frac{1+\kappa}{2}\right|, \left|\frac{\sqrt{1+r_y^2} + \kappa \sqrt{r^2+r_y^2}}{\sqrt{1+r_y^2} + \sqrt{r^2+r_y^2}}\right|\right\} > 0$.

Proof. The reasoning is the same as in the proof of Lemma 3.9. The only difference in this case is, that $\mathfrak{h}_{r_y}(s) \in [0, \frac{\sqrt{12}}{r_y}]$ so we do not consider the limit at infinity and get

$$\mathfrak{f}_{\kappa,r}\left(\frac{\sqrt{12}}{r_y}\right) = \frac{1 + \kappa \sqrt{\frac{r^2+r_y^2}{1+r_y^2}}}{1 + \sqrt{\frac{r^2+r_y^2}{1+r_y^2}}} = \frac{\sqrt{1+r_y^2} + \kappa \sqrt{r^2+r_y^2}}{\sqrt{1+r_y^2} + \sqrt{r^2+r_y^2}}$$

instead. Again, the condition (18) ensures, that both bounds have the same sign, so we can safely take the minimum of their absolute values. □

Now we are in the position to conclude our analysis of the full discretization (7) in the following theorem.

Theorem 3.17. *If (18) is satisfied, then $\mathcal{A}_{h_\pm, h_y}^{-1}$ exists and satisfies*

$$\left\| \mathcal{A}_{h_\pm, h_y}^{-1} \right\|_{\mathcal{L}(V_{h_\pm} \otimes W_{h_y})} \leq \frac{1}{|\sigma_-| \min\left\{1, |\kappa|, \left|\frac{1+\kappa}{2}\right|, \left|\frac{\sqrt{1+r_y^2} + \kappa \sqrt{r^2+r_y^2}}{\sqrt{1+r_y^2} + \sqrt{r^2+r_y^2}}\right|\right\}}.$$

Contrary, if (17) is satisfied, then

$$\lim_{h_{\pm} \rightarrow 0} \left\| \mathcal{A}_{h_{\pm}, h_y}^{-1} \right\|_{\mathcal{L}(V_{h_{\pm}} \otimes W_{h_y})} = +\infty$$

(where we define $\|\mathcal{A}_{h_{\pm}, h_y}^{-1}\|_{\mathcal{L}(V_{h_{\pm}} \otimes W_{h_y})} := +\infty$, if $\mathcal{A}_{h_{\pm}, h_y}^{-1}$ does not exist), and in particular $\mathcal{A}_{h_{\pm}, h_y}$ admits a nontrivial kernel for each

$$h_{\pm} = \frac{1}{m} \frac{1}{r_y} s_{\kappa, r, r_y} = \frac{1}{m} \frac{1}{r_y} \arccos \left(1 + \frac{6r_y^2(1-\kappa^2)}{(1-\kappa^2r^2) - 2r_y^2(1-\kappa^2)} \right), \quad m \in \mathbb{N}.$$

Note that to simultaneously satisfy $\frac{\pi}{h_y} \in \mathbb{N}$ we can choose a particular r_y or r such that $s_{\kappa, r, r_y} = \pi l/k$, $l, k \in \mathbb{N}$, which yields that $\frac{\pi}{h_y} \in \mathbb{N}$ for $m \in k\mathbb{N}$.

Proof. The proof can be obtained by following the steps of the proof of Theorem 3.10 one-to-one and replacing needed lemmas by the corresponding lemmas from this section. \square

3.2. Bounded domain

Now we consider a bounded domain for which actual numerical computations are possible. Since the well-posedness analysis follows along the lines of Section 3.1.1 we suffice ourselves with stating the following lemma without proof.

Lemma 3.18. *The space $H_0^1(\tilde{\Omega})$ admits an orthogonal decomposition $H_0^1(\tilde{\Omega}) = \tilde{X}_- \oplus \tilde{X}_0 \oplus \tilde{X}_+$ where $\tilde{X}_{\pm} := \{u \in H_0^1(\tilde{\Omega}) : u|_{\tilde{\Omega}_{\mp}} = 0\}$ and \tilde{X}_0 is spanned by the orthonormal basis*

$$\left(\frac{1}{\sqrt{2\lambda_m(e^{4\lambda_m L} - 1)}} \left(e^{2\lambda_m L} e^{-\lambda_m |x|} - e^{\lambda_m |x|} \right) \otimes \theta_m(y) \right)_{m \in \mathbb{N}}.$$

The operator $\tilde{\mathcal{A}}$ is block diagonal and the statements of Lemma 3.2 and Corollary 3.3 apply with $\mathcal{A}, X_{\pm}, X_0$ being replaced by $\tilde{\mathcal{A}}, \tilde{X}_{\pm}, \tilde{X}_0$.

3.2.1. Semi discretization

We consider a semi discretization of problem (2b) by means of Galerkin spaces $\tilde{V}_{h_{\pm}} \otimes H_0^1(0, \pi)$:

$$\text{Find } u \in \tilde{V}_{h_{\pm}} \otimes H_0^1(0, \pi) \text{ such that } a_{\tilde{\Omega}}(u, u^{\dagger}) = \langle f, u^{\dagger} \rangle_{\tilde{\Omega}} \text{ for all } u^{\dagger} \in \tilde{V}_{h_{\pm}} \otimes H_0^1(0, \pi). \quad (19)$$

Let $\tilde{\mathcal{A}}_{h_{\pm}} \in \mathcal{L}(\tilde{V}_{h_{\pm}} \otimes H_0^1(0, \pi))$ be the associated operator defined by

$$\left\langle \tilde{\mathcal{A}}_{h_{\pm}} u, u^{\dagger} \right\rangle_{H_0^1(\tilde{\Omega})} = a_{\tilde{\Omega}}(u, u^{\dagger}) \text{ for all } u, u^{\dagger} \in \tilde{V}_{h_{\pm}} \otimes H_0^1(0, \pi). \quad (20)$$

As in the unbounded case we can use (1) and derive a decomposition into Fourier modes as in (9). Similarly we then also get that for $u_m = \sum_{n=-N_-}^{N_+} \alpha_n \phi_n$, $u_m^{\dagger} = \sum_{n=-N_-}^{N_+} \alpha_n^{\dagger} \phi_n$, $\alpha_n, \alpha_n^{\dagger} \in \mathbb{R}$ it follows that

$$\langle \sigma \partial_x u_m, \partial_x u_m^{\dagger} \rangle_{\mathbb{R}} + \lambda_m^2 \langle \sigma u_m, u_m^{\dagger} \rangle_{\mathbb{R}} = \sum_{n=-N_-}^{N_+} \sum_{n'=-N_-}^{N_+} \alpha_{n'}^{\dagger} \mathbf{A}_{n', n}^{(m)} \alpha_n,$$

$$c_{1,+}\mu_{m,1,+}^{N_+} + c_{2,+}\mu_{m,2,+}^{N_+} = 0,$$

with the solution

$$c_{1,+} = \frac{\tilde{\beta}_0^{(m)}}{1 - \nu_{m,+}^{N_+}}, \quad c_{2,+} = -\frac{\tilde{\beta}_0^{(m)}}{1 - \nu_{m,+}^{N_+}}\nu_{m,+}^{N_+}, \quad c_{1,-} = \frac{\tilde{\beta}_0^{(m)}}{1 - \nu_{m,-}^{N_-}}, \quad c_{2,-} = -\frac{\tilde{\beta}_0^{(m)}}{1 - \nu_{m,-}^{N_-}}\nu_{m,-}^{N_-}.$$

Here $\tilde{\beta}_0^{(m)}$ still has to be determined. This means, that

$$\begin{aligned} \tilde{\beta}_n^{(m)} &= \tilde{\beta}_0^{(m)} \frac{\mu_{m,1,-}^{-n} - \nu_{m,-}^{N_-} \mu_{m,2,-}^{-n}}{1 - \nu_{m,-}^{N_-}}, & -N_- \leq n \leq 0 \\ \tilde{\beta}_n^{(m)} &= \tilde{\beta}_0^{(m)} \frac{\mu_{m,1,+}^n - \nu_{m,+}^{N_+} \mu_{m,2,+}^n}{1 - \nu_{m,+}^{N_+}}, & 0 \leq n \leq N_+. \end{aligned}$$

One can check that these formulas still hold in the case where $b_-^{(m)}$ or $b_+^{(m)}$ is zero if we then set $\mu_{m,1,\pm}$ to zero as we have done before and set $\mu_{m,2,\pm}$ to an arbitrary positive number. Finally we compute $\tilde{\beta}_0^{(m)}$ to normalize the solution. For this we only have to consider the scalar product with ϕ_0 because of orthogonality. This leads to

$$\begin{aligned} 1 &= \langle \partial_x \tilde{v}_m, \partial_x \tilde{v}_m \rangle_{\mathbb{R}} + \lambda_m^2 \langle \tilde{v}_m, \tilde{v}_m \rangle_{\mathbb{R}} \\ &= \sum_{n=N_-}^{N_+} \tilde{\beta}_n^{(m)} (\langle \partial_x \phi_n, \partial_x \tilde{v}_m \rangle_{\mathbb{R}} + \lambda_m^2 \langle \phi_n, \tilde{v}_m \rangle_{\mathbb{R}}) \\ &= \tilde{\beta}_0^{(m)} (\langle \partial_x \phi_0, \partial_x \tilde{v}_m \rangle_{\mathbb{R}} + \lambda_m^2 \langle \phi_0, \tilde{v}_m \rangle_{\mathbb{R}}) \\ &= \left(\tilde{\beta}_0^{(m)} \right)^2 \left(\frac{\mu_{m,1,-} - \nu_{m,-}^{N_-} \mu_{m,2,-}}{1 - \nu_{m,-}^{N_-}} b_-^{(m)} + a_-^{(m)} + a_+^{(m)} + \frac{\mu_{m,1,+} - \nu_{m,+}^{N_+} \mu_{m,2,+}}{1 - \nu_{m,+}^{N_+}} b_+^{(m)} \right) \\ &= \left(\tilde{\beta}_0^{(m)} \right)^2 \left(\frac{(\mu_{m,1,-} b_-^{(m)} + a_-^{(m)}) - \nu_{m,-}^{N_-} (\mu_{m,2,-} b_-^{(m)} + a_-^{(m)})}{1 - \nu_{m,-}^{N_-}} \right. \\ &\quad \left. + \frac{(\mu_{m,1,+} b_+^{(m)} + a_+^{(m)}) - \nu_{m,+}^{N_+} (\mu_{m,2,+} b_+^{(m)} + a_+^{(m)})}{1 - \nu_{m,+}^{N_+}} \right) \\ &= \left(\tilde{\beta}_0^{(m)} \right)^2 \left(\frac{\sqrt{(a_-^{(m)})^2 - (b_-^{(m)})^2} + \nu_{m,-}^{N_-} \sqrt{(a_-^{(m)})^2 - (b_-^{(m)})^2}}{1 - \nu_{m,-}^{N_-}} \right. \\ &\quad \left. + \frac{\sqrt{(a_+^{(m)})^2 - (b_+^{(m)})^2} + \nu_{m,+}^{N_+} \sqrt{(a_+^{(m)})^2 - (b_+^{(m)})^2}}{1 - \nu_{m,+}^{N_+}} \right) \\ &= \left(\tilde{\beta}_0^{(m)} \right)^2 \left(\frac{1 + \nu_{m,-}^{N_-}}{1 - \nu_{m,-}^{N_-}} \sqrt{(a_-^{(m)})^2 - (b_-^{(m)})^2} + \frac{1 + \nu_{m,+}^{N_+}}{1 - \nu_{m,+}^{N_+}} \sqrt{(a_+^{(m)})^2 - (b_+^{(m)})^2} \right), \end{aligned}$$

i.e.,

$$\tilde{\beta}_0^{(m)} = \frac{1}{\sqrt{\frac{1+\nu_{m,-}^{N_-}}{1-\nu_{m,-}^{N_-}} \sqrt{(a_-^{(m)})^2 - (b_-^{(m)})^2} + \frac{1+\nu_{m,+}^{N_+}}{1-\nu_{m,+}^{N_+}} \sqrt{(a_+^{(m)})^2 - (b_+^{(m)})^2}}}.$$

□

Lemma 3.20. *The operator $\tilde{\mathcal{A}}_{h_\pm}$ is block diagonal with respect to the orthogonal decomposition given in Lemma 3.19. The blocks corresponding to $\tilde{V}_{h_\pm}^- \otimes H_0^1(0, \pi)$ and $\tilde{V}_{h_\pm}^+ \otimes H_0^1(0, \pi)$ equal the identity times σ_- and σ_+ respectively. The block corresponding to $\tilde{V}_{h_\pm}^0$ is diagonal with respect to the basis given in Lemma 3.19 and the diagonal entries are given by*

$$\tilde{d}_m := \frac{\sigma_- \frac{1+\nu_{m,-}^{N_-}}{1-\nu_{m,-}^{N_-}} \sqrt{(a_-^{(m)})^2 - (b_-^{(m)})^2} + \sigma_+ \frac{1+\nu_{m,+}^{N_+}}{1-\nu_{m,+}^{N_+}} \sqrt{(a_+^{(m)})^2 - (b_+^{(m)})^2}}{\frac{1+\nu_{m,-}^{N_-}}{1-\nu_{m,-}^{N_-}} \sqrt{(a_-^{(m)})^2 - (b_-^{(m)})^2} + \frac{1+\nu_{m,+}^{N_+}}{1-\nu_{m,+}^{N_+}} \sqrt{(a_+^{(m)})^2 - (b_+^{(m)})^2}}, \quad m \in \mathbb{N}. \quad (21)$$

Non-verbally: For each $u_-, u_-^\dagger \in \tilde{V}_{h_\pm}^- \otimes H_0^1(0, \pi)$, $u_0, u_0^\dagger \in \tilde{V}_{h_\pm}^0$, $u_+, u_+^\dagger \in \tilde{V}_{h_\pm}^+ \otimes H_0^1(0, \pi)$ and $u_0 = \sum_{m \in \mathbb{N}} \beta_m \tilde{v}_m \otimes \theta_m$, $u_0^\dagger = \sum_{m \in \mathbb{N}} \beta_m^\dagger \tilde{v}_m \otimes \theta_m$, $(\beta_m)_{m \in \mathbb{N}}, (\beta_m^\dagger)_{m \in \mathbb{N}} \in \ell^2(\mathbb{N})$ it holds that

$$a_{\tilde{\Omega}}(u_- + u_0 + u_+, u_-^\dagger + u_0^\dagger + u_+^\dagger) = \sigma_- \langle u_-, u_-^\dagger \rangle_{H_0^1(\tilde{\Omega})} + \sum_{m \in \mathbb{N}} \tilde{d}_m \beta_m \beta_m^\dagger + \sigma_+ \langle u_+, u_+^\dagger \rangle_{H_0^1(\tilde{\Omega})}.$$

Proof. The statement can be obtained by repeating the steps from Lemma 3.5 one-to-one. □

We will now analyze the diagonal entries \tilde{d}_m and to do this we introduce the following functions:

$$\begin{aligned} \tilde{f}_{\kappa,r,\lambda_m}(t) &:= \frac{1 + \frac{\kappa\sqrt{r^2 t^2 + 12}}{\sqrt{t^2 + 12}} \mathfrak{z}_{r,\lambda_m}(t)}{1 + \frac{\sqrt{r^2 t^2 + 12}}{\sqrt{t^2 + 12}} \mathfrak{z}_{r,\lambda_m}(t)}, & \mathfrak{z}_{r,\lambda_m}(t) &:= \frac{j_{\lambda_m L}(\mathfrak{q}(t))}{j_{\lambda_m L}(\mathfrak{q}(rt))}, \\ \mathfrak{q}(t) &:= \left(\frac{1 + \frac{1}{3} t^2 - t \sqrt{1 + \frac{1}{12} t^2}}{1 + \frac{1}{3} t^2 + t \sqrt{1 + \frac{1}{12} t^2}} \right)^{\frac{1}{t}}, & j_n(q) &:= \frac{1 - q^n}{1 + q^n}. \end{aligned}$$

Lemma 3.21. *The diagonal entries \tilde{d}_m defined in (21) satisfy $\tilde{d}_m = \sigma_- \tilde{f}_{\kappa,r,\lambda_m}(\lambda_m h_-)$.*

Proof. We first compute

$$\nu_{m,-}^{N_-} = \left(\frac{\mu_{m,1,-}}{\mu_{m,2,-}} \right)^{\frac{L_-}{h_-}} = \left(\frac{1 + \frac{1}{3} \lambda_m^2 h_-^2 - h_- \lambda_m \sqrt{1 + \frac{1}{12} \lambda_m^2 h_-^2}}{1 + \frac{1}{3} \lambda_m^2 h_-^2 + h_- \lambda_m \sqrt{1 + \frac{1}{12} \lambda_m^2 h_-^2}} \right)^{\frac{\lambda_m L_-}{\lambda_m h_-}} = \mathfrak{q}(h_- \lambda_m)^{\lambda_m L_-}.$$

In the same way, we get $\nu_{m,+}^{N_+} = \mathfrak{q}(r h_- \lambda_m)^{\lambda_m L_-}$. Now we calculate

$$\tilde{d}_m = \frac{\sigma_- \frac{1+\nu_{m,-}^{N_-}}{1-\nu_{m,-}^{N_-}} \sqrt{(a_-^{(m)})^2 - (b_-^{(m)})^2} + \sigma_+ \frac{1+\nu_{m,+}^{N_+}}{1-\nu_{m,+}^{N_+}} \sqrt{(a_+^{(m)})^2 - (b_+^{(m)})^2}}{\frac{1+\nu_{m,-}^{N_-}}{1-\nu_{m,-}^{N_-}} \sqrt{(a_-^{(m)})^2 - (b_-^{(m)})^2} + \frac{1+\nu_{m,+}^{N_+}}{1-\nu_{m,+}^{N_+}} \sqrt{(a_+^{(m)})^2 - (b_+^{(m)})^2}}$$

$$\begin{aligned}
&= \frac{\frac{\sigma_- \lambda_m}{j_{\lambda_m L}(\mathfrak{q}(\lambda_m h_-))} \sqrt{1 + \frac{1}{12} \lambda_m^2 h_-^2} + \frac{\sigma_+ \lambda_m}{j_{\lambda_m L}(\mathfrak{q}(r \lambda_m h_-))} \sqrt{1 + \frac{1}{12} r^2 \lambda_m^2 h_-^2}}{\frac{\lambda_m}{j_{\lambda_m L}(\mathfrak{q}(\lambda_m h_-))} \sqrt{1 + \frac{1}{12} \lambda_m^2 h_-^2} + \frac{\lambda_m}{j_{\lambda_m L}(\mathfrak{q}(r \lambda_m h_-))} \sqrt{1 + \frac{1}{12} r^2 \lambda_m^2 h_-^2}} \\
&= \sigma_- \frac{\sqrt{1 + \frac{1}{12} \lambda_m^2 h_-^2} + \frac{\kappa j_{\lambda_m L}(\mathfrak{q}(\lambda_m h_-))}{j_{\lambda_m L}(\mathfrak{q}(r \lambda_m h_-))} \sqrt{1 + \frac{1}{12} r^2 \lambda_m^2 h_-^2}}{\sqrt{1 + \frac{1}{12} \lambda_m^2 h_-^2} + \frac{j_{\lambda_m L}(\mathfrak{q}(\lambda_m h_-))}{j_{\lambda_m L}(\mathfrak{q}(r \lambda_m h_-))} \sqrt{1 + \frac{1}{12} r^2 \lambda_m^2 h_-^2}} \\
&= \sigma_- \frac{1 + \kappa \mathfrak{z}_{r, \lambda_m}(\lambda_m h_-) \frac{\sqrt{12 + r^2 \lambda_m^2 h_-^2}}{\sqrt{12 + \lambda_m^2 h_-^2}}}{1 + \mathfrak{z}_{r, \lambda_m}(\lambda_m h_-) \frac{\sqrt{12 + r^2 \lambda_m^2 h_-^2}}{\sqrt{12 + \lambda_m^2 h_-^2}}} = \sigma_- \tilde{\mathfrak{f}}_{\kappa, r, \lambda_m}(\lambda_m h_-).
\end{aligned}$$

□

In preparation of Lemma 3.23 we formulate the following lemma.

Lemma 3.22. *For each $c > 0$ it holds that $\lim_{h_- \rightarrow 0^+} \sup_{\lambda > 0, \lambda h_- \geq c} \mathfrak{z}_{r, \lambda}(\lambda h_-) = 1$.*

Proof. Note that the auxiliary function j_n in the definition of $\mathfrak{z}_{r, \lambda_m}$ was chosen this way to deal with a particular technicality in proof Lemma 3.26. Here however, we exploit the more explicit representation

$$\mathfrak{z}_{r, \lambda}(\lambda h_-) = \frac{1 - \left(\mathfrak{q}(\lambda h_-)^{\lambda h_-}\right)^{\frac{L}{h_-}}}{1 + \left(\mathfrak{q}(\lambda h_-)^{\lambda h_-}\right)^{\frac{L}{h_-}}} \frac{1 + \left(\mathfrak{q}(r \lambda h_-)^{r \lambda h_-}\right)^{\frac{L}{r h_-}}}{1 - \left(\mathfrak{q}(r \lambda h_-)^{r \lambda h_-}\right)^{\frac{L}{r h_-}}}.$$

The claim follows now from $\sup_{t \geq \frac{c}{\max(1, r)}} \mathfrak{q}(t)^t < 1$. □

The following lemma is the pendant to Lemma 3.8.

Lemma 3.23. *If (14) is satisfied, then $\lim_{h_- \rightarrow 0^+} \inf_{m \in \mathbb{N}} |\tilde{\mathfrak{f}}_{\kappa, r, \lambda_m}(\lambda_m h_-)| = 0$.*

Proof. We proceed as in the proof of Lemma 3.8:

Let $h_- = \frac{1}{l+\epsilon} t_{\kappa, r}$ with $l \in \mathbb{N}_0$ and $\epsilon \in [0, 1)$. We choose $m = l$ and exploit that $\lim_{l \rightarrow +\infty} \frac{l}{l+\epsilon} \rightarrow 1$ uniformly in $\epsilon \in [0, 1)$. Thus also $\lambda_m h_- = \frac{l}{l+\epsilon} t_{\kappa, r} \xrightarrow{l \rightarrow +\infty} t_{\kappa, r}$ uniformly in $\epsilon \in [0, 1)$.

Lemma 3.22 yields that $\lim_{l \rightarrow +\infty} \mathfrak{z}_{r, \lambda_l}(\frac{l}{l+\epsilon} t_{\kappa, r}) = 1$, which in combination with the continuity of $\mathfrak{f}_{\kappa, r}$ and $\mathfrak{f}_{\kappa, r}(t_{\kappa, r}) = 0$ provides the claim. □

In preparation of Lemma 3.26 we introduce the following Lemmas 3.24 and 3.25, where Lemma 3.24 is itself an auxiliary result for Lemma 3.25.

Lemma 3.24. *Let I be a closed subintervall of $(0, 1)$ and $c > 0$. Then the following statements hold:*

- (1) *The family $\{j_n : n \in [c, \infty)\}$ and the family of their derivatives are both uniformly bounded on I .*
- (2) *The family $\{\frac{1}{j_n} : n \in [c, \infty)\}$ and the family of their derivatives are both uniformly bounded on I .*
- (3) *\mathfrak{q} is continuously differentiable on $[0, \infty)$.*
- (4) *$\lim_{t \rightarrow 0^+} \mathfrak{q}(t) = e^{-2}$.*

Proof. Let $I = [a, b]$, $a, b \in (0, 1)$. Thence $q^n \in (0, b^c] \subset (0, 1)$ for $q \in [a, b]$. Thus $j_n(q) \leq 1$ and $1/j_n(q) \leq \sup_{s \in (0, b^c]} \frac{1+s}{1-s} = \frac{1+b^c}{1-b^c}$. Furthermore, $j'_n(q) = \frac{2nq^{n-1}}{(1-q^n)^2}$ and $(1/j_n)'(q) = \frac{-2nq^{n-1}}{(1+q^n)^2}$ from which we can deduce the first two claims. The differentiability of \mathfrak{q} follows in a straightforward fashion. The last claim follows by applying l'Hospital's rule to $\log \mathfrak{q}$. □

Lemma 3.25. For $r, c > 0$ the family $\{\mathfrak{z}_{r,\lambda} : \lambda \in [c, \infty)\}$ is equicontinuous from the right at 0 and $\lim_{t \rightarrow 0^+} \mathfrak{z}_{r,\lambda}(t) = 1$.

Proof. Due to the second half of Lemma 3.24 there exist $\delta > 0$ and $q_1, q_2 \in (0, 1)$, $q_1 < q_2$ such that $\mathbf{q}(t), \mathbf{q}(rt) \in (q_1, q_2)$ for all $t \in [0, \delta)$. Then Lemma 3.24 and the product rule yield that $t \mapsto \mathfrak{z}_{r,\lambda}(t)$ is uniformly equicontinuous on $t \in [0, \delta)$, $\lambda \geq c$. At last we obtain $\lim_{t \rightarrow 0^+} \mathfrak{z}_{r,\lambda}(t) = 1$ from $\lim_{t \rightarrow 0^+} \mathbf{q}(t) = e^{-2} = \lim_{t \rightarrow 0^+} \mathbf{q}(rt)$. \square

Lemma 3.26. If $\epsilon \in (0, 1)$ and one of the following two conditions

$$\underbrace{|\kappa|(1+\epsilon) < 1 \quad \text{and} \quad r|\kappa|(1+\epsilon) < 1}_{(22a)} \quad \text{or} \quad \underbrace{|\kappa|(1-\epsilon) > 1 \quad \text{and} \quad r|\kappa|(1-\epsilon) > 1}_{(22b)} \quad (22)$$

is satisfied, then there exists $\delta > 0$ such that

$$\inf_{\substack{h_- \in (0, \delta) \\ \lambda \geq 1}} \left| \tilde{\mathfrak{f}}_{\kappa,r,\lambda}(\lambda h_-) \right| \geq \min_{p \in \{\pm 1\}} \min \left\{ \frac{|1 + (1+p\epsilon)\kappa|}{2+\epsilon}, \frac{|1 + (1+p\epsilon)\kappa r|}{1 + (1+\epsilon)r} \right\}.$$

Proof. Due to Lemma 3.25 we can find $\tau > 0$ such that $\mathfrak{z}_{r,\lambda}(\lambda h_-) \in [1-\epsilon, 1+\epsilon]$ for all $\lambda h_- \in [0, \tau]$, $h_- > 0$, $\lambda \geq 1$. On the other hand, Lemma 3.22 yields the existence of $\delta > 0$ such that $\mathfrak{z}_{r,\lambda}(\lambda h_-) \in [1-\epsilon, 1+\epsilon]$ for all $\lambda h_- \geq \tau$, $h_- \in (0, \delta)$, $\lambda \geq 1$. Thus $\mathfrak{z}_{r,\lambda}(t) \in [1-\epsilon, 1+\epsilon]$ for all $t = \lambda h_-$, $\lambda \geq 1$, $h_- \in (0, \delta)$. Thence

$$\inf_{\substack{h_- \in (0, \delta) \\ \lambda \geq 1}} \left| \tilde{\mathfrak{f}}_{\kappa,r,\lambda}(\lambda h_-) \right| \geq \inf_{t > 0, p \in \{\pm 1\}} \frac{|1 + \kappa(1+p\epsilon)\sqrt{\frac{12+r^2t^2}{12+t^2}}|}{1 + (1+\epsilon)\sqrt{\frac{12+r^2t^2}{12+t^2}}},$$

from which the claim follows. \square

Having analyzed $\tilde{\mathfrak{f}}_{\kappa,r,\lambda_m}$ we can now prove the following theorem about the stability of $\tilde{\mathcal{A}}_{h_\pm}$.

Theorem 3.27. If for some $\epsilon \in (0, 1)$ (22) is satisfied, then $\tilde{\mathcal{A}}_{h_\pm}^{-1}$ exists and satisfies

$$\left\| \tilde{\mathcal{A}}_{h_\pm}^{-1} \right\|_{\mathcal{L}(\tilde{V}_{h_\pm} \otimes H_0^1(0,\pi))} \leq \frac{1}{|\sigma_-| \min \left\{ 1, |\kappa|, \frac{|1+(1+\epsilon)\kappa|}{2+\epsilon}, \frac{|1+(1-\epsilon)\kappa|}{2+\epsilon}, \frac{|1+(1+\epsilon)\kappa r|}{1+(1+\epsilon)r}, \frac{|1+(1-\epsilon)\kappa r|}{1+(1+\epsilon)r} \right\}}$$

for all $h_- \in (0, \delta)$ with $\delta > 0$ as in Lemma 3.26. Contrary, if (14) is satisfied then

$$\lim_{h_- \rightarrow 0} \left\| \tilde{\mathcal{A}}_{h_\pm}^{-1} \right\|_{\mathcal{L}(\tilde{V}_{h_\pm} \otimes H_0^1(0,\pi))} = +\infty$$

(where we define $\left\| \tilde{\mathcal{A}}_{h_\pm}^{-1} \right\|_{\mathcal{L}(\tilde{V}_{h_\pm} \otimes H_0^1(0,\pi))} := +\infty$, if $\tilde{\mathcal{A}}_{h_\pm}^{-1}$ does not exist).

Proof. As in the previous sections the theorem follows directly from the properties of $\tilde{\mathfrak{f}}_{\kappa,r,\lambda_m}$ that were shown in the Lemmas 3.23 and 3.26. \square

3.2.2. Full discretization

As for the unbounded domain, the only difference is that we now consider $\hat{\lambda}_{m,r_y,h_-}$ that also depend on h_- and where in variables that depend on λ_m we replace it by $\hat{\lambda}_{m,r_y,h_-}$ and indicate this by adding a hat. In addition, we define $\hat{\nu}_{m,\pm} := \frac{\hat{\mu}_{m,1,\pm}}{\hat{\mu}_{m,2,\pm}}$. The following three lemmas can then be derived correspondingly to Lemmas 3.19–3.21.

Lemma 3.28. *The space $\tilde{V}_{h_{\pm}} \otimes W_{h_y}$ admits an orthogonal decomposition*

$$\tilde{V}_{h_{\pm}} \otimes W_{h_y} = \left(\tilde{V}_{h_{\pm}}^- \otimes W_{h_y} \right) \oplus^{\perp} \hat{V}_{h_{\pm}}^0 \oplus^{\perp} \left(\tilde{V}_{h_{\pm}}^+ \otimes W_{h_y} \right),$$

where $\tilde{V}_{h_{\pm}}^- := \{v \in \tilde{V}_{h_{\pm}} : v|_{\mathbb{R}_+} = 0\}$ and $\tilde{V}_{h_{\pm}}^+ := \{v \in \tilde{V}_{h_{\pm}} : v|_{\mathbb{R}_-} = 0\}$. The subspace $\hat{V}_{h_{\pm}}^0$ is spanned by the orthonormal basis $(\hat{v}_m(x) \otimes \hat{\theta}_m(y))_{m=1, \dots, M-1}$, where

$$\begin{aligned} \hat{v}_m(x) := & \frac{1}{\sqrt{\frac{1+\hat{\nu}_{m,-}^{N_-}}{1-\hat{\nu}_{m,-}^{N_-}} \sqrt{(\hat{a}_-^{(m)})^2 - (\hat{b}_-^{(m)})^2} + \frac{1+\hat{\nu}_{m,+}^{N_+}}{1-\hat{\nu}_{m,+}^{N_+}} \sqrt{(\hat{a}_+^{(m)})^2 - (\hat{b}_+^{(m)})^2}}} \\ & \times \left(\phi_0(x) + \sum_{n=-N_-}^{-1} \frac{\hat{\mu}_{m,1,-}^{-n} - \hat{\nu}_{m,-}^{N_-} \hat{\mu}_{m,2,-}^{-n}}{1 - \hat{\nu}_{m,-}^{N_-}} \phi_n(x) + \sum_{n=1}^{N_+} \frac{\hat{\mu}_{m,1,+}^n - \hat{\nu}_{m,+}^{N_+} \hat{\mu}_{m,2,+}^n}{1 - \hat{\nu}_{m,+}^{N_+}} \phi_n(x) \right). \end{aligned}$$

Lemma 3.29. *The operator $\tilde{\mathcal{A}}_{h_{\pm}, h_y}$ is block diagonal with respect to the orthogonal decomposition given in Lemma 3.28. The blocks corresponding to $\tilde{V}_{h_{\pm}}^- \otimes W_{h_y}$ and $\tilde{V}_{h_{\pm}}^+ \otimes W_{h_y}$ equal the identity times σ_- and σ_+ respectively. The block corresponding to $\hat{V}_{h_{\pm}}^0$ is diagonal with respect to the basis given in Lemma 3.28 and the diagonal entries are given by*

$$\hat{d}_m := \frac{\sigma_- \frac{1+\hat{\nu}_{m,-}^{N_-}}{1-\hat{\nu}_{m,-}^{N_-}} \sqrt{(\hat{a}_-^{(m)})^2 - (\hat{b}_-^{(m)})^2} + \sigma_+ \frac{1+\hat{\nu}_{m,+}^{N_+}}{1-\hat{\nu}_{m,+}^{N_+}} \sqrt{(\hat{a}_+^{(m)})^2 - (\hat{b}_+^{(m)})^2}}{\frac{1+\hat{\nu}_{m,-}^{N_-}}{1-\hat{\nu}_{m,-}^{N_-}} \sqrt{(\hat{a}_-^{(m)})^2 - (\hat{b}_-^{(m)})^2} + \frac{1+\hat{\nu}_{m,+}^{N_+}}{1-\hat{\nu}_{m,+}^{N_+}} \sqrt{(\hat{a}_+^{(m)})^2 - (\hat{b}_+^{(m)})^2}}, \quad m = 1, \dots, M-1. \quad (23)$$

Non-verbally: For each $u_-, u_-^\dagger \in \tilde{V}_{h_{\pm}}^- \otimes W_{h_y}$, $u_0, u_0^\dagger \in \hat{V}_{h_{\pm}}^0$, $u_+, u_+^\dagger \in \tilde{V}_{h_{\pm}}^+ \otimes W_{h_y}$ and $u_0 = \sum_{m=1}^{M-1} \beta_m \hat{v}_m \otimes \hat{\theta}_m$, $u_0^\dagger = \sum_{m=1}^{M-1} \beta_m^\dagger \hat{v}_m \otimes \hat{\theta}_m$, $(\beta_m)_{m=1}^{M-1}, (\beta_m^\dagger)_{m=1}^M \in \mathbb{R}^{M-1}$ it holds that

$$a_{\tilde{\Omega}}(u_- + u_0 + u_+, u_-^\dagger + u_0^\dagger + u_+^\dagger) = \sigma_- \langle u_-, u_-^\dagger \rangle_{H_0^1(\tilde{\Omega})} + \sum_{m=1}^{M-1} \hat{d}_m \beta_m \beta_m^\dagger + \sigma_+ \langle u_+, u_+^\dagger \rangle_{H_0^1(\tilde{\Omega})}.$$

Lemma 3.30. *The diagonal entries \hat{d}_m satisfy $\hat{d}_m = \sigma_- \tilde{f}_{\kappa, r, \hat{\lambda}_{m, r_y, h_-}}(\mathfrak{h}_{r_y}(r_y m h_-))$ for $m = 1, \dots, M-1$.*

Now we investigate $\tilde{f}_{\kappa, r, \hat{\lambda}_{m, r_y, h_-}}(\mathfrak{h}_{r_y}(r_y m h_-))$. Since we have already shown, that $\tilde{f}_{\kappa, r, \lambda}$ is equicontinuous at zero in $\lambda \geq 1$ and because $\hat{\lambda}_{m, r_y, h_-} \geq \lambda_m \geq 1$, the only difference to the previous analysis concerning the unbounded domain is that we have to deal with the additional composition with the continuous function \mathfrak{h}_{r_y} . Before we formulate the next Lemma 3.15, let us recall that $h_- = \frac{h_y}{r_y} = \frac{\pi}{r_y M}$.

Lemma 3.31. *If (17) is satisfied, then $\lim_{M \rightarrow +\infty} \inf_{m \in \{1, \dots, M-1\}} |\tilde{f}_{\kappa, r, \hat{\lambda}_{m, r_y, h_-}}(\mathfrak{h}_{r_y}(\frac{m\pi}{r_y M}))| = 0$.*

Proof. The proof follows along the lines of the proof of Lemma 3.15, where in addition we apply Lemma 3.22 to cope with the replacement of $\tilde{f}_{\kappa, r}$ by $\tilde{f}_{\kappa, r, \hat{\lambda}_{m, r_y, h_-}}$. \square

Lemma 3.32. *If for $\epsilon \in (0, 1)$ one of the following two conditions*

$$|\kappa|(1 + \epsilon) < 1 \quad \text{and} \quad r^2 \kappa^2 (1 + \epsilon)^2 < 1 + r_y (1 - \kappa^2 (1 + \epsilon)^2) \quad (24a)$$

or

$$|\kappa|(1-\epsilon) > 1 \quad \text{and} \quad r^2\kappa^2(1-\epsilon)^2 > 1 + r_y^2(1-\kappa^2(1-\epsilon)^2) \quad (24b)$$

is satisfied, then there exists $\delta > 0$ such that

$$\begin{aligned} & \inf_{\substack{h_- \in (0, \delta) \\ m \in \{1, \dots, M-1\}}} \left| \tilde{\mathfrak{f}}_{\kappa, r, \hat{\lambda}_m, r_y, h_-}(\mathfrak{h}_{r_y}(r_y m h_-)) \right| \\ & \geq \min_{p \in \{\pm 1\}} \min \left\{ \frac{|1 + (1+p\epsilon)\kappa|}{2 + \epsilon}, \frac{|\sqrt{1+r_y^2} + (1+p\epsilon)\kappa\sqrt{r^2+r_y^2}|}{\sqrt{1+r_y^2} + (1+\epsilon)\sqrt{r^2+r_y^2}} \right\} > 0. \end{aligned}$$

Proof. It suffices to combine the techniques used for Lemmas 3.16 and 3.26. As in the proof of Lemma 3.26 we choose $\delta > 0$ such that $\mathfrak{z}_{r, \hat{\lambda}_m, r_y, h_-}(\hat{\lambda}_m, r_y, h_-) \in [1 - \epsilon, 1 + \epsilon]$ for all $\hat{\lambda}_m, r_y, h_- \geq 1, h_- \in (0, \delta)$. Thence

$$\inf_{\substack{h_- \in (0, \delta) \\ m \in \{1, \dots, M-1\}}} \left| \tilde{\mathfrak{f}}_{\kappa, r, \hat{\lambda}_m, r_y, h_-}(\mathfrak{h}_{r_y}(r_y m h_-)) \right| \geq \inf_{t \in [0, \sqrt{12}/r_y], p \in \{\pm 1\}} \frac{|1 + \kappa(1+p\epsilon)\sqrt{\frac{12+r^2t^2}{12+t^2}}|}{1 + (1+\epsilon)\sqrt{\frac{12+r^2t^2}{12+t^2}}},$$

from which the claim follows. \square

Now we are in the position to conclude our analysis of the full discretization of (2b) in the following theorem.

Theorem 3.33. *If (24) is satisfied for some $\epsilon \in (0, 1)$, then $\tilde{\mathcal{A}}_{h_\pm, h_y}^{-1}$ exists and satisfies*

$$\left\| \tilde{\mathcal{A}}_{h_\pm, h_y}^{-1} \right\|_{\mathcal{L}(\tilde{V}_{h_\pm} \otimes W_{h_y})} \leq \frac{1}{|\sigma_-| \min_{p \in \{\pm 1\}} \min \left\{ 1, |\kappa|, \frac{|1+(1+p\epsilon)\kappa|}{2+\epsilon}, \frac{|\sqrt{1+r_y^2} + (1+p\epsilon)\kappa\sqrt{r^2+r_y^2}|}{\sqrt{1+r_y^2} + (1+\epsilon)\sqrt{r^2+r_y^2}} \right\}}$$

for all $h_- \in (0, \delta)$ with $\delta > 0$ as in Lemma 3.16. Contrary, if (17) is satisfied, then

$$\lim_{h_- \rightarrow 0} \left\| \tilde{\mathcal{A}}_{h_\pm, h_y}^{-1} \right\|_{\mathcal{L}(\tilde{V}_{h_\pm} \otimes W_{h_y})} = +\infty$$

(where we define $\|\tilde{\mathcal{A}}_{h_\pm, h_y}^{-1}\|_{\mathcal{L}(\tilde{V}_{h_\pm} \otimes W_{h_y})} := +\infty$, if $\tilde{\mathcal{A}}_{h_\pm, h_y}^{-1}$ does not exist).

Proof. As for Theorems 3.10, 3.17 and 3.27 the claims follow directly from the respective Lemmas 3.29–3.32. \square

4. COMPUTATIONAL EXAMPLES

We will now confirm our theoretical results by testing them in explicit computational examples. The code to reproduce all of them is provided in [19]. We consider the following problem posed on a bounded domain:

$$\begin{aligned} & \text{Find } u \in H_0^1(\tilde{\Omega}) \quad \text{such that} \quad -\operatorname{div}(\sigma \nabla u) = f \quad \text{in } \tilde{\Omega} = (-L, L) \times (0, \pi), \\ & \text{with } f(x, y) = -\sigma \left[\frac{-2y(y-\pi)(y-2\pi)}{L^2} + 6 \left(1 - \frac{x^2}{L^2} \right) (y-\pi) \right], \end{aligned}$$

which has the solution

$$u(x, y) = \left(1 - \frac{x^2}{L^2} \right) y(y-\pi)(y-2\pi).$$

First we examine the case of unstable discretizations. To this end we consider parameters as follows:

$$\sigma_- = -1, \quad \sigma_+ = 1.2, \quad r = 0.5, \quad r_y = \frac{2}{\sqrt{11}}.$$

Hence the contrast at hand $\kappa = -1.2$ is rather moderate. Nevertheless, we are going to exhibit instabilities with this particular choice of parameters, whereas typical examples of instabilities often require a much more critical contrast $|\kappa - \kappa_{\text{crit}}| \approx 10^{-3}$ ([2], Fig. 1), ([9], Fig. 3). We choose a sufficiently large $L \approx 26$ such that our problem is adequately close to an unbounded domain, and hence we can expect the critical values

$$h_- := \frac{1}{m} \frac{1}{r_y} \arccos \left(1 + \frac{6r_y^2(1 - \kappa^2)}{(1 - \kappa^2 r^2) - 2r_y^2(1 - \kappa^2)} \right) = \frac{\sqrt{11}}{2} \frac{1}{m} \arccos(1 - 1) = \frac{\sqrt{11}\pi}{4m} \quad m \in \mathbb{N} \quad (25)$$

given in Theorem 3.17 to yield also sensible values for our example. In particular, we choose $L := 10 \frac{\sqrt{11}\pi}{4}$ such that

$$N_- = \frac{L}{h_-} = 10 \frac{\sqrt{11}\pi}{4} \frac{4m}{\sqrt{11}\pi} = 10m, \quad N_+ = \frac{L}{h_+} = 10 \frac{\sqrt{11}\pi}{4} \frac{8m}{\sqrt{11}\pi} = 20m,$$

$$M = \frac{\pi}{r_y h_-} = \pi \frac{4m}{\sqrt{11}\pi} \frac{\sqrt{11}}{2} = 2m$$

are natural numbers for each $m \in \mathbb{N}$. Even though we cannot expect our discretizations to have a non-trivial kernel at h_- we observe in Figure 2 (solid lines) exorbitant errors. Nevertheless, we recognize a decrease in the error which can be explained as follows: The drastic error is triggered by the basis function $\hat{v}_m \otimes \hat{\theta}_m$ for which the respective coefficient $\langle f, \hat{v}_m \otimes \hat{\theta}_m \rangle_{L^2}$ of f decreases w.r.t. $m \in \mathbb{N}$. In Figure 3 we see the numerical solution for a critical value of h_- and we observe the oscillating behavior of $\hat{v}_m \otimes \hat{\theta}_m$ that is corrupting the solution. As predicted in Lemma 3.19 the corruption is the strongest close to the interface and decreases exponentially towards the boundary.

To further explore the possible behaviours of different discretizations we keep all parameters apart from h_- unchanged and choose now $h_- = \frac{\sqrt{11}\pi}{4(m+\frac{1}{2})}$ to maximize the distance of h_- to the critical values (25) (while keeping $N_-, N_+, M \in \mathbb{N}$). In contrast to our previous results, we observe in Figure 2 (dashed lines) a distinct convergence of errors with convenient rates.

Finally, we consider meshes with the values of h_- and h_+ being exchanged. We observe in Figure 4 a convergence with convenient rates as predicted by Theorem 3.17. In Figure 5 we see that using the flipped mesh avoids any oscillating behaviour at the interface.

To conclude, for problems with more complicated geometries and discretizations with non-uniform meshes we expect a mixed behaviour, where at each new mesh refinement a stable or unstable setting is dominant in an unpredictable way, giving rise to the commonly observed zigzag error curves.

5. CONCLUSION

In this article we considered the inconclusive observations of instabilities in the numerical simulation of sign-changing materials. To study this question we took the most primitive geometry, *i.e.*, a flat interface, and discretization, *i.e.*, lowest order tensor product FEM, but allowed different mesh sizes in the two subdomains and in the horizontal direction. Our analysis reveals that depending on the ratios of those mesh sizes relative to the contrast the obtained discretizations are either unstable or stable. In applications we expect a mixture of those settings leading to the observed spikes in the otherwise well-behaved error *vs.* mesh size plots. Nevertheless, our numerical studies showed that the predicted instabilities generated by a constructed sequence of critical mesh sizes are only triggered for mesh sizes sufficiently close the critical values.

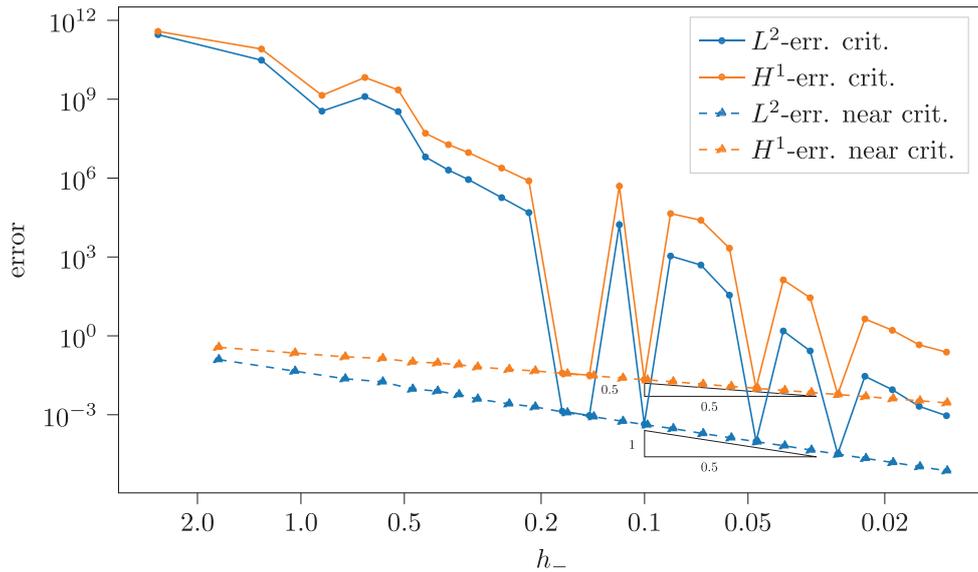


FIGURE 2. Relative errors for critical values of h_- (solid lines) and for nearly critical values of h_- (dashed lines).

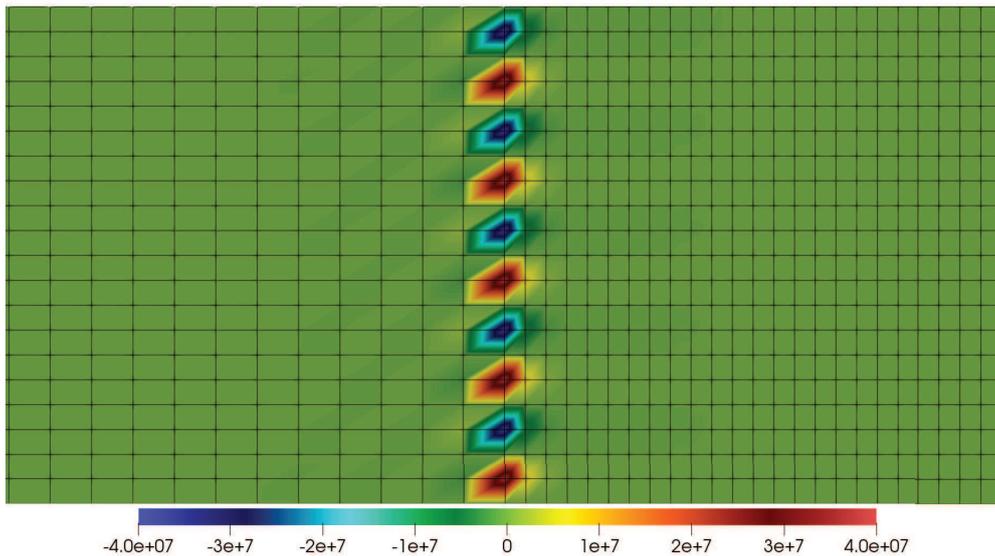


FIGURE 3. Computed solution for $h_- \approx 0.26048$ in a neighborhood of the interface $\{0\} \times (0, \pi)$.

To interpret the implications of our results we make the following comments. From a purely mathematical point of view we proved that in general standard FEMs for sign-changing PDEs are not stable, which justifies to research on specialized methods. On the other hand, we showed that the triggering of the instabilities can be elusive, which explains the “almost good” looking performance of standard FEMs when compared to specialized methods, and hence this reappearing phenomenon in computational studies is not a numerical artifact.

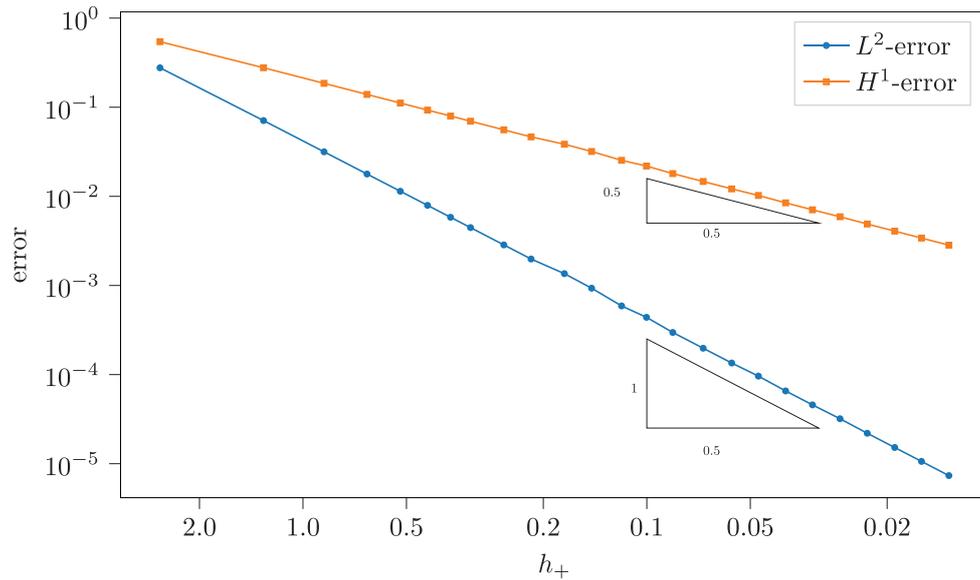


FIGURE 4. Relative errors for a mesh satisfying (24).

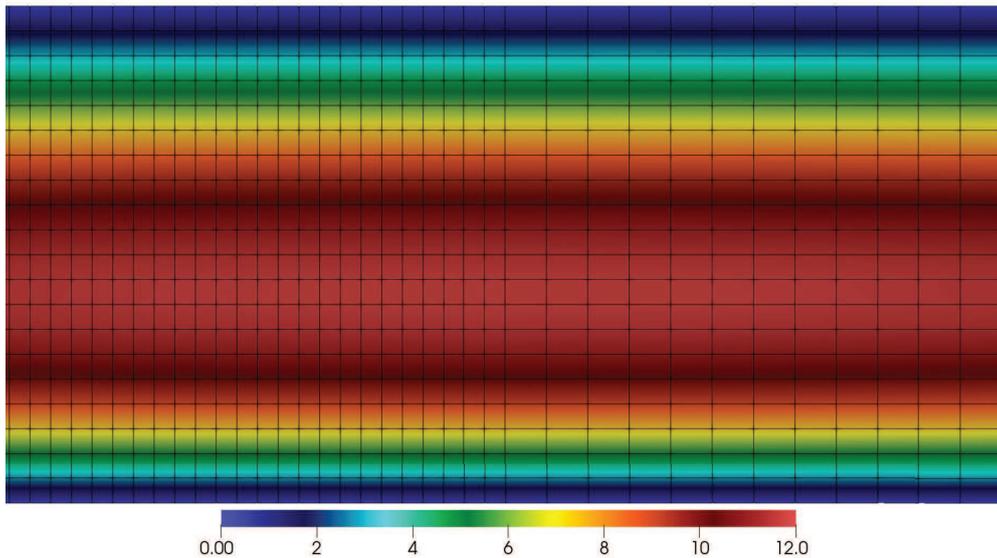


FIGURE 5. Computed solution for $h_+ \approx 0.26048$ on a flipped mesh in a neighborhood of the interface $\{0\} \times (0, \pi)$.

It is a natural question, if the presented analysis can be extended to more general setting such as (1) high order FEM, (2) triangular meshes or (3) interfaces with corners. Regarding (1) we anticipate that a result similar to Theorem 3.33 is valid. However, for higher polynomial degrees the analysis would become even more technical and we are not able to derive certain expressions explicitly any more. Regarding a generalization to triangular meshes we note that our analysis crucially relies on a tensor product structure of all spaces and

operators. Since triangular meshes would violate this structure a direct extension of our technique to this case seems impossible. Similar considerations apply to interfaces with corners: We cannot stretch our analysis to this geometry, because a tensor product structure appears only after a transformation to polar coordinates (which is not desired). Nevertheless, a result similar to Theorem 3.33 for corner interfaces would be of great practical interest, because it could potentially lead to simple meshing rules guaranteeing stability.

FUNDING

The first author acknowledges funding from Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), projects 541433971 and 258734477 – SFB 1173 and that part of this work was conducted at the Johann Radon Institute for Computational and Applied Mathematics. The second author acknowledges support from DFG, CRC 1456 project 432680300.

DATA AVAILABILITY STATEMENT

The code associated with this article is available in “GRO.data”, under the reference <https://doi.org/10.25625/BHLLP> [19].

REFERENCES

- [1] A. Abdulle and S. Lemaire, An optimization-based method for sign-changing elliptic PDEs. *ESAIM Math. Model. Numer. Anal.* **58** (2024) 2187–2223.
- [2] A. Abdulle, M.E. Huber and S. Lemaire, An optimization-based numerical method for diffusion problems with sign-changing coefficients. *C. R. Math.* **355** (2017) 472–478.
- [3] D. Boffi, Finite element approximation of eigenvalue problems. *Acta Numer.* **19** (2010) 1–120.
- [4] A.-S. Bonnet-Ben Dhia, P. Ciarlet Jr. and C.M. Zwölf, Time harmonic wave diffraction problems in materials with sign-shifting coefficients. *J. Comput. Appl. Math.* **234** (2010) 1912–1919.
- [5] A.-S. Bonnet-BenDhia, L. Chesnel and P. Ciarlet, T-coercivity for scalar interface problems between dielectrics and metamaterials. *Math. Mod. Num. Anal.* **46** (2012) 363–1387.
- [6] A.-S. Bonnet-BenDhia, L. Chesnel and P. Ciarlet, T-coercivity for the Maxwell problem with sign-changing coefficients. *Commun. Partial Differ. Equ.* **39** (2014) 1007–1031.
- [7] A.-S. Bonnet-Ben Dhia, L. Chesnel and P. Ciarlet, Jr., Two-dimensional Maxwell’s equations with sign-changing coefficients. *Appl. Numer. Math.* **79** (2014) 29–41.
- [8] A.-S. Bonnet-Ben Dhia, C. Carvalho and P. Ciarlet, Mesh requirements for the finite element approximation of problems with sign-changing coefficients. *Numer. Math.* **138** (2018) 801–838.
- [9] E. Burman, A. Ern, and J. Preuss. A stabilized hybridized Nitsche method for sign-changing elliptic PDEs. *Math. Models Methods Appl. Sci.*, **35** (2025) 2977–3009.
- [10] F. Chaaban, P. Ciarlet and M. Rihani, Solving numerically the two-dimensional time harmonic Maxwell problem with sign-changing coefficients (2025). <https://hal.science/hal-04909034>.
- [11] P. Ciarlet, Jr., D. Lassounon and M. Rihani, An optimal control-based numerical method for scalar transmission problems with sign-changing coefficients. *SIAM J. Numer. Anal.* **61** (2023) 1316–1339.
- [12] S.A. Cummer, J. Christensen and A. Alù, Controlling sound with acoustic metamaterials. *Nat. Rev. Mater.* **1** (2016) 1–13.
- [13] S.-E. Ekström and S. Serra-Capizzano, Eigenvalues and eigenvectors of banded Toeplitz matrices and the related symbols. *Numer. Linear Algebra Appl.* **25** (2018) e2137.
- [14] A. Greenleaf, Y. Kurylev, M. Lassas and G. Uhlmann, Cloaking devices, electromagnetic wormholes, and transformation optics. *SIAM Rev.* **51** (2009) 3–33.
- [15] M. Halla, On the approximation of dispersive electromagnetic eigenvalue problems in two dimensions. *IMA J. Numer. Anal.* **43** (2023) 535–559.
- [16] M. Halla, T. Hohage, and F. Oberender. A new numerical method for scalar eigenvalue problems in heterogeneous, dispersive, sign-changing materials. *Journal of Scientific Computing* (accepted), (2024). <https://arxiv.org/abs/2401.16368>.
- [17] H.-M. Nguyen, Limiting absorption principle and well-posedness for the Helmholtz equation with sign changing coefficients. *J. Math. Pures Appl.* **106** (2016) 342–374.

- [18] H.-M. Nguyen and S. Sil, Limiting absorption principle and well-posedness for the time-harmonic Maxwell equations with anisotropic sign-changing coefficients. *Commun. Math. Phys.* **379** (2020) 145–176.
- [19] F. Oberender, Replication data for: on the instabilities of naive FEM discretizations for PDEs with sign-changing coefficients (2025). DOI: [10.25625/BHLLP](https://doi.org/10.25625/BHLLP).
- [20] G. Unger, Convergence analysis of a Galerkin boundary element method for electromagnetic resonance problems. *Partial Differ. Equ. Appl.* **2** (2021) 39.



Please help to maintain this journal in open access!

This journal is currently published in open access under the Subscribe to Open model (S2O). We are thankful to our subscribers and supporters for making it possible to publish this journal in open access in the current year, free of charge for authors and readers.

Check with your library that it subscribes to the journal, or consider making a personal donation to the S2O programme by contacting subscribers@edpsciences.org.

More information, including a list of supporters and financial transparency reports, is available at <https://edpsciences.org/en/subscribe-to-open-s2o>.