Institut für Numerische Simulation
Rheinische Friedrich-Wilhelms-Universität Bonn

Wegelerstraße 6 • 53115 Bonn • Germany
phone +49 228 73-3427 • fax +49 228 73-7527
www.ins.uni-bonn.de

S. Beuchler, K. Hofer

Additive Schwarz solvers for hp-FEM discretizations of PDE-constrained optimzation problems

INS Preprint No. 1625

October 2016

Additive Schwarz solvers for hp-FEM discretizations of PDE-constrained optimzation problems

Sven Beuchler[*] Katharina Hofer[†]

October 21, 2016

Abstract

In this paper, we investigate the minimization of a quadratic functional subject to a boundary value problem of a second order linear elliptic partial differential equation. There are no inequality constraints for state and control. This problem is discretized by hp-finite elements. The main focus of this talk is the development of efficient solution methods for the corresponding system of linear algebraic equations. From the literature it is known that this system is symmetric and indefinite. This paper considers two different iterative solvers

- a conjugate gradient (CG) method in a special inner product, following Schöberl/Zulehner, and
- the minimal residual method.

In both methods, efficient preconditioners for finite element mass and stiffness matrix acclerate the convergence speed of the underlying iterative method. This contribution presents overlapping hp- FEM preconditioners for mass and stiffness matrix. Optimal condition number estimates are proved. Robustness with respect to the regularization parameter can be shown for the CG-method. Finally several numerical examples show the efficiency of the presented algorithm.

1 Introduction

This paper deals with the numerical solution to the following optimal control problem: Minimize the functional $J(y,u)$ given by

$$\min_{y,u} J(y,u) = \min_{y,u} \left(\frac{1}{2} \int_\Omega (y(x) - y_d(x))^2 \, \mathrm{d}x + \frac{\alpha}{2} \int_\Omega u^2(x) \, \mathrm{d}x \right) \tag{1}$$

subject to: Find $y \in H^1_{\Gamma_\mathcal{D}}(\Omega) := \{v \in H^1(\Omega) : v \,|_{\Gamma_\mathcal{D}} = 0\}$ such that

$$\mathfrak{a}(y,v) := \int_\Omega (D(x)\nabla y \cdot \nabla v + c(x)yv) \; \mathrm{d}x = \int_\Omega (u+f)v \, \mathrm{d}x \quad \forall v \in H^1_{\Gamma_\mathcal{D}}(\Omega). \tag{2}$$

Here, the control is denoted by u, while the solution y to (2) is the corresponding state. For the precise statement of the assumptions on the data of this problem we refer to Section 2. Under these assumptions, the optimal control problem is uniquely solvable, [50]. The numerical solution to the optimal control problem (1), (2) requires efficient discretization methods for the approximate solution to boundary value problems as in (2) for given u and f. Among other methods, the finite element method (FEM) of low order, i.e. the h-version of the FEM, is a very powerful method for the discretization of (2), see e.g. [13, 18].

[*] Institute for Numerical Simulation, Wegelerstraße 6, 53115 Bonn, Germany, beuchler@ins.uni-bonn.de

[†] Institute for Numerical Simulation, Wegelerstraße 6, 53115 Bonn, Germany, hofer@ins.uni-bonn.de

The discretization of optimal control problems subject to elliptic partial differential equations by means of FEM is a well investigated topic. Usually, (adaptive) h-FEM is applied, see [1,4,15,16,19,34,35] and the references therein.

For the discretization of boundary value problems of the form (2) with smooth solutions, spectral methods (see e.g. [27]) and finite elements of higher order (p-version, see e.g. [20, 21, 45, 47] and the references therein) have become more and more popular. In contrast to h-FEM, in the p-version of the FEM the polynomial degree p is increased and the mesh-size h is kept constant. Both ideas, mesh refinement and increasing the polynomial degree, are combined in the so-called hp-version of the FEM. The advantage of the p-version and hp-version in comparison to the h-version is that the discrete solution converges faster to the exact solution with respect to the number of unknowns N (of course provided that the solution is sufficiently smooth). However, additional singularities can occur in the case of inequality constraints. Applications of hp-FEM to optimal control problems are investigated in [7,8,51,52] with inequality constraints and in [17] without inequality constraints. In [7, 8, 51] a special hp-discretization, the boundary concentrated FEM (BC-FEM), [29] is applied.

After the discretization of the optimal control problem (1), (2) it remains to solve a huge system of algebraic finite element equations $\mathcal{A}\underline{x} = \underline{g}$. If there is no inequality constraint, it is possible to rewrite the problem in a saddle point formulation, see e.g., [32,42,44,46]. The system matrix $\mathcal{A}$ is symmetric, but indefinite. Possible iterative solvers are the preconditioned conjugate gradient method (PCG) in a special scalar product, see [44] or the method of minimal residuals, (MINRES), see [23]. The convergence speed of the iterative solvers depends strongly on the distribution of the eigenvalues of the matrix $\mathcal{A}$. Therefore, preconditioners are applied in order to acclerate the calculation of the solution to the saddle point system. In this special case, the two main ingredients of such a preconditioner are

- a preconditioner for the finite element mass matrix M,
- a preconditioner for a linear combination of the mass matrix M and the stiffness matrix of the discretization of (2).

The papers mentioned above use piecewise linear elements for the discretization. Alternative ideas using multigrid can be found in [24, 40, 43] and the references therein. We refer also to [11] concerning time dependent problems.

This paper is devoted to the efficient numerical solution of the linear systems of algebraic equations arising from the discretization of the optimal control problems by means of hp-FEM. Using the saddle point formulation of [44], the system $\mathcal{A}\underline{x} = \underline{g}$ is solved by the PCG method in a special scalar product, see [44] as well as MINRES. In case of hp-discretizations, it has to be considered, that not only good preconditioners for the stiffness matrix as in case of h-refinement, but also for the mass matrix are necessary, since the condition number of the mass matrix depends strongly on the polynomial degree p, [33]. Most preconditioners for hp-FEM use additive Schwarz methods as domain decomposition methods with inexact subproblem solvers, see e.g [48] and the references therein. Therefore, preconditioners for h-FEM (see e.g. [12,53,54]) have to be combined with preconditioners for p-FEM (see e.g. [3,37]). We refer also to [5,10,22,30,31,41]. Using overlapping additive Schwarz preconditioners as in [37], we are able to prove that the convergence speed of the underlying iterative solution method does not depend on the meshsize h and the polynomial degree p. In addition, we are able to prove robustness with respect to the regularization parameter α in the case of PCG method by Schöberl-Zulehner. To the knowledge of the authors, there are only a few publications for the efficient solution of systems of algebraic equations arising from the discretization of PDE-constrained optimization by means of hp-FEM. For BC-FEM in two space dimensions, direct solvers can be applied [8], see also [28]. However, the fast complexity goes lost in three space dimensions.

The outline of the paper is follows. The setting of the problem is described in Section 2. The discretization with hp-finite elements is described in section 3. Section 4 deals with the numerical solution to the system of linear algebraic equations. The main convergence results are proved. Several numerical experiments are presented in Section 5.

Throughout this paper, the integer p denotes the polynomial degree. For two real symmetric and positive definite $n \times n$ matrices A, B, the relation $A \preceq B$ means that $A - cB$ is negative definite, where $c > 0$ is a constant independent of h, or p. The relation $A \sim B$ means $A \preceq B$ and $B \preceq A$, i.e. the matrices A and B are spectrally equivalent. The isomorphism between a function $u = \sum_i u_i \psi_i \in L^2$ and the vector of coefficients $\underline{u} = [u_i]_i$ with respect to the basis $[\Psi] = [\psi_1, \psi_2, \ldots]$ is denoted as $u = [\Psi]\underline{u}$.

2 Setting of the problem

Considered is the distributed optimal control problem (1), (2), namely

$$\min_{y,u} J(y,u) = \min_{y,u} \left(\frac{1}{2} \int_\Omega (y(x) - y_d(x))^2 \, \mathrm{d}x + \frac{\alpha}{2} \int_\Omega u^2(x) \, \mathrm{d}x \right)$$

subject to: Find $y \in H^1_{\Gamma_\mathcal{D}}(\Omega) := \{v \in H^1(\Omega) : v \mid_{\Gamma_\mathcal{D}} = 0\}$ such that

$$\mathfrak{a}(y,v) := \int_\Omega D(x) \nabla y \cdot \nabla v + c(x) y v \, \mathrm{d}x = \int_\Omega (u + f) v \, \mathrm{d}x \quad \forall v \in H^1_{\Gamma_\mathcal{D}}(\Omega).$$

Concerning the data, the following assumptions hold:

Assumption 2.1. *The domain $\Omega \subset \mathbb{R}^d$ is open, bounded, Lipschitz and has a polygonal boundary $\partial\Omega := \Gamma = \overline{\Gamma_\mathcal{D}} \cup \overline{\Gamma_\mathcal{N}}$ and $\Gamma_\mathcal{N} \cap \Gamma_\mathcal{D} = \emptyset$. For the coefficients D and c, it holds that $D, c \in L_\infty(\Omega)$, $D > 0, c \geq 0$ a.e., and if $\mathrm{meas}(\Gamma_\mathcal{D}) = \emptyset$ it holds $c > 0$. Moreover, D and c are chosen such that the differential operator is bounded and uniformly elliptic in H^1. For the regularization parameter α, it holds $\alpha > 0$, the desired state satisfies $y_d \in L_2(\Omega)$ and the right hand side $f \in L_2(\Omega)$.*

To solve the optimal control problem the adjoint state q is introduced. The adjoint state $q \in H^1_{\Gamma_\mathcal{D}}(\Omega)$ is the unique solution to the variational equation

$$\mathfrak{a}(v,q) = \int_\Omega (y_d - y) v \, \mathrm{d}x \quad \forall v \in H^1_{\Gamma_\mathcal{D}}(\Omega). \tag{3}$$

Since there are no bounds on the control, the control is computed by the projection formula

$$u(x) = \frac{1}{\alpha} q(x) \qquad \text{in } \Omega, \tag{4}$$

see e.g. [50]. The goal is to write the problem (2) - (4) in a saddle point formulation. This paper follows the ansatz in [44]. Therefore, two variational equations are derived. For the first equation, the equations (3) with the test function $v = \tilde{q}$ and the weak formulation of (4) are added. The second equation is the variational formulation of (2). The spaces $\mathbb{Y} = H^1_{\Gamma_\mathcal{D}}(\Omega)$, $\mathbb{U} = L_2(\Omega)$ and $\mathbb{Q} = H^1_{\Gamma_\mathcal{D}}(\Omega)$ and $\mathbb{W} = \mathbb{Y} \times \mathbb{U}$ are introduced. With the Hilbert spaces $\mathbb{W}$ and $\mathbb{Q}$ and the functions $w = (y,u), \tilde{w} = (\tilde{y}, \tilde{u})$, the optimal control problem is reformulated as: Find $(w,q) \in \mathbb{W} \times \mathbb{Q}$ such that

$$\begin{aligned} a(w,\tilde{w}) + b(\tilde{w},q) &= \langle F, \tilde{w} \rangle_\Omega \qquad \text{for all } \tilde{w} \in \mathbb{W}, \\ b(w,\tilde{q}) &= \langle G, \tilde{q} \rangle_\Omega \qquad \text{for all } \tilde{q} \in \mathbb{Q} \end{aligned} \tag{5}$$

with the bilinear forms

$$a(w,\tilde{w}) = \int_\Omega y\tilde{y} \, \mathrm{d}x + \alpha \int_\Omega u\tilde{u} \, \mathrm{d}x \text{ and } b(w,\tilde{q}) = \mathfrak{a}(y,\tilde{q}) - \int_\Omega u\tilde{q} \, \mathrm{d}x, \tag{6}$$

and the linear forms

$$\langle F, \tilde{w} \rangle_\Omega = \int_\Omega y_d \tilde{y} \, \mathrm{d}x \text{ and } \langle G, \tilde{q} \rangle_\Omega = \int_\Omega f \tilde{q} \, \mathrm{d}x.$$

The existence and uniqueness of solutions to (5) is proved by using Brezzis theorem.

Theorem 2.2. *(Brezzis theorem, see e.g. [14]) Let $\mathbb{W}, \mathbb{Q}$ denote real Hilbert spaces, $a : \mathbb{W} \times \mathbb{W} \to \mathbb{R}$, $b : \mathbb{W} \times \mathbb{Q} \to \mathbb{R}$ are bilinear forms, $F : \mathbb{W} \to \mathbb{R}$, $G : \mathbb{Q} \to \mathbb{R}$ are continuous linear functionals. Furthermore, it is assumed that:*

1. *The bilinear form $a(\cdot,\cdot)$ is bounded, i.e. it exists a constant $a_0 < \infty$ such that*
$$a(w,\tilde{w}) \leq a_0 \|w\|_{\mathbb{W}} \|\tilde{w}\|_{\mathbb{W}} \qquad \forall\, w, \tilde{w} \in \mathbb{W}.$$
2. *The bilinear form $a(\cdot,\cdot)$ is coercive on $\ker B = \{\tilde{w} \in \mathbb{W} : b(\tilde{w},\tilde{q}) = 0 \quad \forall\, \tilde{q} \in \mathbb{Q}\}$, i.e. there exists a constant $a_1 > 0$ such that*
$$a(w,w) \geq a_1 \|w\|_{\mathbb{W}}^2 \qquad \forall\, w \in \ker B.$$
3. *The bilinear form $b(\cdot,\cdot)$ is bounded:*
$$\exists\, b_0 < \infty : \qquad \sup_{0 \neq w \in \mathbb{W}} \frac{b(w,q)}{\|w\|_{\mathbb{W}}} \leq b_0 \|q\|_{\mathbb{Q}} \qquad \forall\, q \in \mathbb{Q}.$$
4. *The bilinear form $b(\cdot,\cdot)$ satisfies the inf-sup condition: There exists a constant $b_1 > 0$ such that*
$$\sup_{0 \neq w \in \mathbb{W}} \frac{b(w,q)}{\|w\|_{\mathbb{W}}} \geq b_1 \|q\|_{\mathbb{Q}} \qquad \forall\, q \in \mathbb{Q}.$$

Then, problem (5) *admits a unique solution. Moreover, the a-priori estimates*

$$\|w\|_{\mathbb{W}} \leq \frac{1}{a_1}\|F\|_{\mathbb{W}^*} + \frac{1}{b_1}\left(1 + \frac{a_0}{a_1}\right)\|G\|_{\mathbb{Q}^*} \text{ and } \|q\|_{\mathbb{Q}} \quad \leq \frac{1}{b_1}\left(1 + \frac{a_0}{a_1}\right)\|F\|_{\mathbb{W}^*} + \frac{a_0}{b_1^2}\left(1 + \frac{a_0}{a_1}\right)\|G\|_{\mathbb{Q}^*}$$

hold.

Our aim is to check the assumptions of Brezzis theorem 2.2 with constants a_0, a_1, b_0 and b_1 independent of α. Following [44], non-standard scalar products are introduced. Therefore let

$$\begin{aligned}
\langle u, \tilde{u}\rangle_{\mathbb{U}} &= \alpha \langle u, \tilde{u}\rangle_{\Omega}, \\
\langle y, \tilde{y}\rangle_{\mathbb{Y}} &= \langle y, \tilde{y}\rangle_{\Omega} + \sqrt{\alpha}\, \mathfrak{a}(y,\tilde{y}), \\
\langle q, \tilde{q}\rangle_{\mathbb{Q}} &= \frac{1}{\alpha}\langle q, \tilde{q}\rangle_{\Omega} + \frac{1}{\sqrt{\alpha}}\, \mathfrak{a}(q,\tilde{q}) = \frac{1}{\alpha}\langle q, \tilde{q}\rangle_{\mathbb{Y}}
\end{aligned}$$

denote scalar products in $\mathbb{U}, \mathbb{Y}$ and $\mathbb{Q}$, respectively. The scalar product in the space $\mathbb{W}$ is given by

$$\langle w, \tilde{w}\rangle_{\mathbb{W}} = \langle u, \tilde{u}\rangle_{\mathbb{U}} + \langle y, \tilde{y}\rangle_{\mathbb{Y}} \qquad \text{with } w = (y,u), \tilde{w} = (\tilde{y}, \tilde{u}) \in \mathbb{W}.$$

Then, the energy norms are defined as

$$\|w\|_{\mathbb{W}}^2 = \langle w, w\rangle_{\mathbb{W}}, \text{ and } \|q\|_{\mathbb{Q}}^2 = \langle q, q\rangle_{\mathbb{Q}}.$$

Remark 2.3. *For a fixed $\alpha > 0$ the introduced norms on $\mathbb{Y}$ and $\mathbb{Q}$ are equivalent to the usual $H^1(\Omega)$ norm.*

Next, the assumptions of Brezzis theorem 2.2 are checked.

Lemma 2.4. *Let the bilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ be given by* (6)*, respectively. Then:*

1. *The bilinear form $a(\cdot,\cdot)$ is bounded, i.e. $a(w,\tilde{w}) \leq \|w\|_{\mathbb{W}}\|\tilde{w}\|_{\mathbb{W}} \qquad \forall\, w, \tilde{w} \in \mathbb{W}.$*

2. *The bilinear form* $a(\cdot,\cdot)$ *is coercive on* $\ker B$*, i.e.* $a(w,w) \geq \frac{2}{3}\|w\|_{\mathbb{W}}^2 \qquad \forall\, w \in$ *ker* B.
3. *The bilinear form* $b(\cdot,\cdot)$ *is bounded , i.e.* $\sup_{0\neq\tilde{w}\in\mathbb{W}} \frac{b(\tilde{w},q)}{\|\tilde{w}\|_{\mathbb{W}}} \leq \|q\|_{\mathbb{Q}} \qquad \forall\, q \in \mathbb{Q}$.
4. *The bilinear form* $b(\cdot,\cdot)$ *fulfills the inf-sup-condition* $\sup_{0\neq\tilde{w}\in\mathbb{W}} \frac{b(\tilde{w},q)}{\|\tilde{w}\|_{\mathbb{W}}} \geq \sqrt{\frac{3}{4}}\|q\|_{\mathbb{Q}} \qquad \forall\, q \in \mathbb{Q}$.

Proof. The proof is separated in four parts and follows the proof [44, Lemma 4.1], see also [55]. There, the case $D = c = 1$ is proven. Furthermore the proof uses ideas of [26, Lemma 3.1].

1. For proving the first inequality, one starts with the bilinear form, applies the triangle inequality and Cauchy-Schwarz inequality and yields

$$|a(w,\tilde{w})| = \left|\int_\Omega y\tilde{y}\,\mathrm{d}x + \int_\Omega \alpha u\tilde{u}\,\mathrm{d}x\right| \leq \|y\|_{L_2(\Omega)}\|\tilde{y}\|_{L_2(\Omega)} + \alpha\|u\|_{L_2(\Omega)}\|\tilde{u}\|_{L_2(\Omega)}.$$

   The assertion follows by using the Cauchy-Schwarz inequality in $\mathbb{R}^2$.

2. The proof **coercivity on** $\ker B$ **of the bilinear form** $a(\cdot,\cdot)$ starts from $b(\tilde{w},\tilde{q}) = 0$ with $\tilde{w} = (\tilde{y},\tilde{u})$ i.e.

$$\mathbb{a}(\tilde{y},\tilde{q}) = \int_\Omega D\nabla\tilde{y}\cdot\nabla\tilde{q}\,\mathrm{d}x + \int_\Omega c\tilde{y}\tilde{q}\,\mathrm{d}x = \int_\Omega \tilde{u}\tilde{q}\,\mathrm{d}x \leq \|\tilde{u}\|_{L_2(\Omega)}\|\tilde{q}\|_{L_2(\Omega)}. \tag{7}$$

   This gives the estimate

$$\begin{aligned}\|\tilde{w}\|_{\mathbb{W}}^2 \quad = \quad \|\tilde{y}\|_{\mathbb{Y}}^2 + \|\tilde{u}\|_{\mathbb{U}}^2 \quad &= \quad \|\tilde{y}\|_{L_2(\Omega)}^2 + \sqrt{\alpha}\,\mathbb{a}(\tilde{y},\tilde{y}) + \alpha\|\tilde{u}\|_{L_2(\Omega)}^2 \\ &\overset{\text{(7) with }\tilde{q}=\tilde{y}}{\leq} \quad \|\tilde{y}\|_{L_2(\Omega)}^2 + \sqrt{\alpha}\,\|\tilde{u}\|_{L_2(\Omega)}\|\tilde{y}\|_{L_2(\Omega)} + \alpha\|\tilde{u}\|_{L_2(\Omega)}^2.\end{aligned}$$

   This is equivalent to $a(\tilde{w},\tilde{w}) \geq \frac{2}{3}\|\tilde{w}\|_{\mathbb{W}}^2$ for all $\tilde{w} \in \ker B$.

3. For the **boundedness of the bilinear form** $b(\cdot,\cdot)$ an application of Cauchy-Schwarz leads to

$$|b(w,q)| = \left|\mathbb{a}(y,q) - \int_\Omega uq\,\mathrm{d}x\right| \leq \quad \sqrt{\mathbb{a}(y,y)}\sqrt{\mathbb{a}(q,q)} + \|u\|_{L_2(\Omega)}\|q\|_{L_2(\Omega)} \leq \|w\|_{\mathbb{W}}\|q\|_{\mathbb{Q}}.$$

4. It remains to prove the **inf-sup condition** $\sup_{0\neq\tilde{w}\in\mathbb{W}} \frac{b(\tilde{w},q)}{\|\tilde{w}\|_{\mathbb{W}}} \geq b_1\|q\|_{\mathbb{Q}}$. We choose the function $w = (y,u) = (\sqrt{\alpha}q, -2q)$. This gives

$$\begin{aligned}\sup_{0\neq\tilde{w}\in\mathbb{W}} \frac{b^2(\tilde{w},q)}{\|\tilde{w}\|_{\mathbb{W}}^2} \geq \frac{b^2(w,q)}{\|w\|_{\mathbb{W}}^2} \quad &= \quad \frac{\left(\sqrt{\alpha}\,\mathbb{a}(q,q) + 2\langle q,q\rangle_\Omega\right)^2}{\alpha\sqrt{\alpha}\,\mathbb{a}(q,q) + 5\alpha\langle q,q\rangle_\Omega} \\ &= \quad \frac{1}{\alpha}\frac{\left(\sqrt{\alpha}\,\mathbb{a}(q,q) + 2\langle q,q\rangle_\Omega\right)^2}{\sqrt{\alpha}\,\mathbb{a}(q,q) + 5\langle q,q\rangle_\Omega} \\ &\overset{(a+2b)^2\geq\frac{3}{4}(a+b)(a+5b)}{\geq} \quad \frac{3}{4\alpha}\left(\sqrt{\alpha}\,\mathbb{a}(q,q) + \langle q,q\rangle_\Omega\right) = \frac{3}{4}\|q\|_{\mathbb{Q}}^2,\end{aligned}$$

   which proves the result.

□

Since the assumptions of Brezzis theorem are fulfilled, the existence and uniqueness of the solution follows.

Theorem 2.5. *The saddle point problem* (5) *with the non-standard norms* $\|w\|_{\mathbb{W}}, \|q\|_{\mathbb{Q}}$ *has a unique solution.*

Proof. As proven in Lemma 2.4, the assumptions for Theorem 2.2 hold with the α-independent constants $a_0 = 1, a_1 = \frac{2}{3}, b_0 = 1$ and $b_1 = \sqrt{\frac{3}{4}}$. □

Remark 2.6. *This paper investigates the case of distributed control only. For boundary control two cases have to be distinguished:*

1. *observation on the boundary: If one replaces the L_2 terms on Ω by the L_2 terms on $\Gamma_{\mathcal{N}}$ in the norms in $\mathbb{U}$, $\mathbb{Y}$ and $\mathbb{Q}$, respectively, this case is completely analogous.*
2. *observation on the domain: Here, the norms in $\mathbb{U}$ are the $L_2(\Gamma_{\mathcal{N}})$ norm with factor α. The norms on $\mathbb{Y}$ and $\mathbb{Q}$ contain the $L_2(\Omega)$ norm. Using the trace theorem and the Poincare Friedrichs inequality, the uniqueness of a weak solution can be proved. However, the robustness with respect to α goes lost. Therefore, the norms on $\mathbb{Y}$ and $\mathbb{Q}$ are defined without the regularization parameter α, see [46].*

3 Galerkin discretization

The next step is to discretize the saddle point problem (5). For the discretization, *hp* finite elements are used. Let $\mathbb{W}_N = \mathbb{Y}_N \times \mathbb{U}_N \subset \mathbb{W}$, $\mathbb{Q}_N \subset \mathbb{Q}$ be finite dimensional spaces. Then, the discrete formulation of the saddle point problem reads: Find $(w_N, q_N) \in \mathbb{W}_N \times \mathbb{Q}_N$ such that

$$\begin{aligned} a(w_N, \tilde{w}_N) + b(\tilde{w}_N, q_N) &= \langle F, \tilde{w}_N \rangle_\Omega \qquad \forall\, \tilde{w}_N \in \mathbb{W}_N, \\ b(w_N, \tilde{q}_N) &= \langle G, \tilde{q}_N \rangle_\Omega \qquad \forall\, \tilde{q}_N \in \mathbb{Q}_N. \end{aligned} \tag{8}$$

For existence and uniqueness of the solution Brezzis theorem shall be applied.

Theorem 3.1. *Assume that $\mathbb{W}_N \subset \mathbb{W}$, $\mathbb{Q}_N \subset \mathbb{Q}$ and $\mathbb{W}_N = \mathbb{Y}_N \times \mathbb{U}_N$ with $\mathbb{Y}_N = \mathbb{Q}_N \subset \mathbb{U}_N$. Then, the discrete saddle point problem* (8) *possesses a unique solution. Furthermore, the constants in Brezzis theorem are independent of mesh parameters and the regularization parameter α.*

Proof. The result follows from Theorem 2.2. The assumptions are proved as in Lemma 2.4. □

In this paper, the choice $\mathbb{Y}_N = \mathbb{Q}_N = \mathbb{U}_N$ is made. Let $[\Phi] = [\phi_1, \ldots, \phi_N]$ be a basis for $\mathbb{Y}_N = \mathbb{Q}_N = \mathbb{U}_N$. Then, the variational formulation (8) is equivalent to the linear system solve

$$\mathcal{A} \begin{bmatrix} \underline{w}_N \\ \underline{q}_N \end{bmatrix} = \begin{bmatrix} \underline{f}_N \\ \underline{s}_N \end{bmatrix} \quad \text{with} \quad \mathcal{A} = \begin{bmatrix} A_N & B_N^\top \\ B_N & \mathbf{0} \end{bmatrix}, \quad A_N = \begin{bmatrix} M_N & \mathbf{0} \\ \mathbf{0} & \alpha M_N \end{bmatrix} \quad \text{and} \quad B_N = \begin{bmatrix} K_N \\ -M_N \end{bmatrix}^\top, \tag{9}$$

where M_N denotes the mass matrix representing the $L_2(\Omega)$ inner product and K_N denotes the stiffness matrix with respect to the bilinear form $\mathfrak{a}(\cdot,\cdot)$, i.e.

$$K_N = [\mathfrak{a}(\phi_i, \phi_j)]_{i,j=1}^N, \quad M_N = [\langle \phi_i, \phi_j \rangle_\Omega]_{i,j=1}^N. \tag{10}$$

Remark 3.2. *In matrix vector notation, the assumptions of Theorem 2.2 mean that*

$$\frac{2}{3} \begin{pmatrix} M_N + \sqrt{\alpha} K_N & \mathbf{0} \\ \mathbf{0} & \alpha M_N \end{pmatrix} \leq \qquad A_N \qquad \leq \begin{pmatrix} M_N + \sqrt{\alpha} K_N & \mathbf{0} \\ \mathbf{0} & \alpha M_N \end{pmatrix} \quad \textit{on } \ker B_N, \tag{11}$$

$$\frac{\sqrt{3}}{2} \left(M_N + \sqrt{\alpha} K_N \right) \leq \quad \alpha B_N A_N^{-1} B_N^\top \quad \leq M_N + \sqrt{\alpha} K_N. \tag{12}$$

Next, we specify the space $\mathbb{Y}_N = \mathbb{Q}_N = \mathbb{U}_N$ and its basis functions $[\Phi]$. The polygonal Lipschitz domain Ω is decomposed into a triangulation $\mathcal{T}_s$ consisting of isotropic quadrilateral (2D) or hexahedral (3D) elements $\mathcal{R}_s$. With $\mathcal{R}$ we denote the reference element $(-1,1)^d$ for $d = 2, 3$. Φ_s is the bi-/tri-linear mapping from the reference element $\mathcal{R}$ to the element $\mathcal{R}_s$. $\mathcal{Q}_k$ denotes the space of polynomials on $(-1,1)^d$ with maximal degree k in each variable. The general hp finite element space

$$\mathbb{Y}_N = \mathbb{Q}_N = \mathbb{U}_N = \{y \in H^1_{\Gamma_D}(\Omega) : \quad y|_{\mathcal{R}_s} = \tilde{y} \circ \Phi_s^{-1}, \tilde{y} \in \mathcal{Q}_k^d\}$$

is introduced. In the next step, the spaces are equipped with basis functions. Note that the Legendre polynomials on $[-1, 1]$ are defined as

$$L_i(x) = \frac{1}{2^i i!} \frac{\mathrm{d}^i}{\mathrm{d}x^i} \left(x^2 - 1\right)^i.$$

Moreover, the integrated Legendre polynomials with some scaling factor γ_i are defined via the relation

$$\widehat{L}_i(x) = \gamma_i \int_{-1}^{x} L_{i-1}(s)\,\mathrm{d}s \quad i \geq 2 \qquad \text{and} \qquad \widehat{L}_{0/1}(x) = \frac{1 \pm x}{2}.$$

In the following, we use the scaling $\gamma_i = 1$. The shape functions for 2D and 3D can now be constructed by taking the products

$$\begin{aligned} \widehat{L}_{ij}(x_1, x_2) &= \widehat{L}_i(x_1)\widehat{L}_j(x_2), \quad 0 \leq i, j \leq p, \\ \widehat{L}_{ijm}(x_1, x_2, x_3) &= \widehat{L}_i(x_1)\widehat{L}_j(x_2)\widehat{L}_m(x_3), \quad 0 \leq i, j, m \leq p, \end{aligned}$$

on the reference element, respectively. The global functions ϕ_i can be divided into four groups,

- the vertex functions (V),
- the edge bubble functions,
- face bubble functions (only for $d = 3$),
- the interior bubble functions,

locally on each element $\mathcal{R}_s$, and globally on the domain Ω. We denote them with $[\Phi] = [\phi_1, \ldots, \phi_N]$, see e.g. [6], for more details.

4 The solution to the linear system

This section considers the efficient solution to the linear system of algebraic equations (9). In this paper, iterative solvers with preconditioners are the method of choice. Note that the system matrix is symmetric but indefinite. We start with a brief summary of two possible solution methods for systems of the form (9) and its main convergence results.

4.1 Solvers for saddle point systems

In this subsection, a system of the form

$$\mathcal{A} \begin{bmatrix} \underline{x} \\ \underline{q} \end{bmatrix} = \begin{bmatrix} \underline{f} \\ \underline{s} \end{bmatrix} \quad \text{with} \quad \mathcal{A} = \begin{bmatrix} A & B^\top \\ B & 0 \end{bmatrix} \tag{13}$$

is considered under the following assumptions:

Assumption 4.1. *Let $A \in \mathbb{R}^{N_A \times N_A}$ be a symmetric and positive semidefinite matrix, $B \in \mathbb{R}^{N_B \times N_A}$ has full rank $N_B \leq N_A$. Moreover, it is set $N = N_A + N_B$ and it holds*

$$\langle A\underline{x}, \underline{x} \rangle > 0 \quad \textit{for all } \underline{x} \in \ker B \textit{ with } \underline{x} \neq \underline{0}.$$

The first method is the modified PCG method by Schöberl and Zulehner [44]. The Schöberl-Zulehner PCG method is based on choosing a suitable preconditioner $\mathcal{P}_{\mathrm{cg},\hat{A},\hat{S}}$ in order to apply the CG method with respect to a non-standard scalar product. The preconditioner for the matrix $\mathcal{A}$ (13) is chosen as

$$\mathcal{P}_{\mathrm{cg},\hat{A},\hat{S}} = \begin{bmatrix} \hat{A} & B^\top \\ B & B\hat{A}^{-1}B^\top - \hat{S} \end{bmatrix}, \tag{14}$$

where $\hat{A}$ and $\hat{S}$ are symmetric and positive definite matrices with respect to the standard scalar product.

Theorem 4.2. *[44, Theorems 2.1,2.2] Let assumption 4.1 be fulfilled. Furthermore, it is assumed that the relations $\hat{A} > 0$ and $\hat{S} > 0$ with*

$$\langle A\underline{w}, \underline{w} \rangle \geq \nu_1 \langle \hat{A}\underline{w}, \underline{w} \rangle \qquad \textit{for all } \underline{w} \in \ker B \textit{ and } \hat{A} \geq A,$$

and

$$\hat{S} \leq B\hat{A}^{-1}B^\top \leq \nu_2 \hat{S},$$

with constants ν_1, ν_2 and $0 < \nu_1 \leq 1$ and $\nu_2 \geq 1$ hold. Let $\mathcal{P}_{\mathrm{cg},\hat{A},\hat{S}}$ be defined by (14)*. Then, all eigenvalues of $\mathcal{P}^{-1}_{\mathrm{cg},\hat{A},\hat{S}}\mathcal{A}$ are real and positive. Moreover, the maximal and minimal eigenvalue satisfy the estimates*

$$\lambda_{\max}(\mathcal{P}_{\mathrm{cg}}^{-1}\mathcal{A}) \leq \nu_2 \left(1 + \sqrt{1 - \nu_2^{-1}}\right) \textit{ and } \lambda_{\min}(\mathcal{P}_{\mathrm{cg}}^{-1}\mathcal{A}) \geq \nu_1 \left[\frac{2}{\sqrt{1 - \nu_2^{-1}} + \sqrt{5 - \nu_2^{-1}}}\right]^2 > 0.$$

Remark 4.3. *If in addition, $\hat{A} > A$ and $\hat{S} < B\hat{A}^{-1}B^\top$, then $\mathcal{P}^{-1}_{\mathrm{cg},\hat{A},\hat{S}}\mathcal{A}$ is symmetric and positive definite with respect to the scalar product*

$$\left\langle \begin{bmatrix} \underline{x} \\ \underline{p} \end{bmatrix}, \begin{bmatrix} \underline{w} \\ \underline{q} \end{bmatrix} \right\rangle_{\mathcal{D}} = \langle (\hat{A} - A)\underline{x}, \underline{w} \rangle + \langle (B\hat{A}^{-1}B^\top - \hat{S})\underline{p}, \underline{q} \rangle.$$

This method requires the explicit eigenvalue bounds $\hat{A} > A$ and $\hat{S} < B\hat{A}^{-1}B^\top$ for the preconditioner (14). This can be done by explicit estimates or an inverse vector iteration. In order to avoid this, the MINRES can be used. This method has been designed for symmetric indefinite system solves. The preconditioned version uses the preconditioner

$$\mathcal{P}_{\mathrm{minres},\tilde{A},\tilde{S}} = \begin{bmatrix} \tilde{A} & 0 \\ 0 & \tilde{S} \end{bmatrix}, \tag{15}$$

where $\tilde{A}$ and $\tilde{S}$ are spd preconditioners for A and the Schur complement $S = BA^{-1}B^\top$, respectively. Then, the following result can be proved:

Theorem 4.4. *([2, Corollary 2]) Let $\mathcal{A}$ and $\mathcal{P}_{\mathrm{minres},\tilde{A},\tilde{S}}$ be defined by* (13) *and* (15)*, respectively, where $\tilde{A}$ and $\tilde{S}$ are symmetric and positive definite preconditioners for A and S. Then, all eigenvalues $\mathcal{P}^{-1}_{\mathrm{minres},\tilde{A},\tilde{S}}\mathcal{A}$ are contained in the intervals*

$$\left[-\lambda_{\max}(\tilde{S}^{-1}S), \frac{-\lambda_{\min}(\tilde{S}^{-1}S)}{1 + \frac{1}{\lambda_{\min}(\tilde{A}^{-1}A)}}\right] \cup \left[\lambda_{\min}(\tilde{A}^{-1}A), \lambda_{\max}(\tilde{A}^{-1}A) + \lambda_{\max}(\tilde{S}^{-1}S)\right].$$

4.2 Additive Schwarz preconditioners

Summarizing, both proposed iterative methods require preconditioners for the matrix A_N and the Schur complement $S_N = B_N A_N^{-1} B_N^\top$. Due to the definition of A_N and B_N, see (9), and the spectral equivalence relations (11), (12) this means the construction of preconditioners for the mass matrix M_N and the linear combination $Y_N = \sqrt{\alpha} K_N + M_N$ of stiffness and mass matrix in (10). Therefore, efficient solvers for M_N and Y_N are required. Here, overlapping additive Schwarz methods are preferred.
For the definition of the preconditioner, some notation is introduced. Let

$$\mathbb{V}_h = \{y \in H^1_{\Gamma_1}(\Omega) : \quad y \mid_{\mathcal{R}_s} = \tilde{y}(F_s^{-1}(x_1, \ldots, x_d)), \tilde{y} \in \mathcal{Q}_1\} \tag{16}$$

be the space of all finite element functions of maximal polynomial degree 1. For a given node v, let $\Omega_\mathrm{v} = \{\cup_s \overline{\mathcal{R}}_s : \quad \mathrm{v} \subset \overline{\mathcal{R}_s}\}$ be the closed patch associated to a node v of the finite element mesh. Then, for each node v of the finite element mesh, we introduce

$$\mathbb{V}_\mathrm{v} = \{y \in \mathbb{Y}_N : \quad \text{supp } y \subset \Omega_\mathrm{v}\} \tag{17}$$

as the patch space, cf. a two-dimensional example in Figure 1.

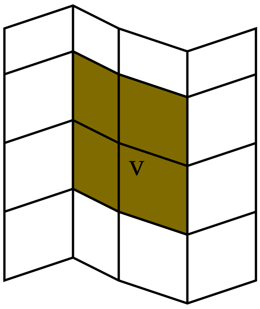

Figure 1: Patch Ω_v of a node v (2D) (marked colored).

Theorem 4.5. *Let $\mathbb{V}_\mathrm{v}$ and $\mathbb{V}_h$ be defined via* (17) *and* (16)*, respectively. Then, for all $y \in \mathbb{Y}_N$ there exists a decomposition $y = y_h + \sum_\mathrm{v} y_\mathrm{v}$ with $y_h \in \mathbb{V}_h$ and $y_\mathrm{v} \in \mathbb{V}_\mathrm{v}$ such that*

$$\begin{aligned} \mathfrak{a}(y,y) &\succeq \inf_{y=y_h+y_\mathrm{v}} \left(\mathfrak{a}(y_h,y_h) + \sum_\mathrm{v} \mathfrak{a}(y_\mathrm{v},y_\mathrm{v}) \right) \quad \textit{and} \\ \langle y,y\rangle_\Omega &\succeq \inf_{y=y_h+y_\mathrm{v}} \left(\langle y_h,y_h\rangle_\Omega + \sum_\mathrm{v} \langle y_\mathrm{v},y_\mathrm{v}\rangle_\Omega \right) \end{aligned}$$

hold. Moreover, for all decompositions $y = y_h + \sum_\mathrm{v} y_\mathrm{v}$, $y_h \in \mathbb{V}_h, y_\mathrm{v} \in \mathbb{V}_\mathrm{v}$, the estimates

$$\begin{aligned} \mathfrak{a}(y,y) &\preceq \mathfrak{a}(y_h,y_h) + \sum_\mathrm{v} \mathfrak{a}(y_\mathrm{v},y_\mathrm{v}) \quad \textit{and} \\ \langle y,y\rangle_\Omega &\preceq \langle y_h,y_h\rangle_\Omega + \sum_\mathrm{v} \langle y_\mathrm{v},y_\mathrm{v}\rangle_\Omega. \end{aligned}$$

The constants depend neither on h nor p.

Proof. This result in the $\mathfrak{a}$-norm has been proven by Pavarino [38]. Using the same decomposition as in [38], see also [39], the result for the L_2 norm is proved, see [9]. The upper estimates follow by a coloring argument. □

Since, a preconditioner for $Y_N = \sqrt{\alpha} K_N + M_N$ is required, the bilinear form

$$\mathfrak{p}(y, \tilde{y}) = \sqrt{\alpha}\mathfrak{a}(y, \tilde{y}) + \langle y, \tilde{y} \rangle_\Omega \tag{18}$$

on $\mathbb{Y} \times \mathbb{Y}$ is introduced.

Corollary 4.6. *Let $\mathfrak{p}$ be defined by* (18)*. Then, for all $y \in \mathbb{Y}_N$ there exists a decomposition $y = y_h + \sum_{\mathtt{v}} y_{\mathtt{v}}$ with $y_h \in \mathbb{V}_h$ and $y_{\mathtt{v}} \in \mathbb{V}_{\mathtt{v}}$ such that*

$$\mathfrak{p}(y, y) \quad \succeq \quad \inf_{y = y_h + y_{\mathtt{v}}} \left(\mathfrak{p}(y_h, y_h) + \sum_{\mathtt{v}} \mathfrak{p}(y_{\mathtt{v}}, y_{\mathtt{v}}) \right).$$

Moreover, for all decompositions $y = y_h + \sum_{\mathtt{v}} y_{\mathtt{v}}$ with $y_h \in \mathbb{V}_h$ and $y_{\mathtt{v}} \in \mathbb{V}_{\mathtt{v}}$

$$\mathfrak{p}(y, y) \quad \preceq \quad \mathfrak{p}(y_h, y_h) + \sum_{\mathtt{v}} \mathfrak{p}(y_{\mathtt{v}}, y_{\mathtt{v}}).$$

The constants depend neither on α, h nor p.

Proof. The independence of h and p follows from Theorem 4.5. The robustness in α is also a consequence of Theorem 4.5, since the decomposition $y = y_h + \sum_{\mathtt{v}} y_{\mathtt{v}}$ is the same in both involved norms. □

Remark 4.7. *The bilinear form $\mathfrak{b}(\cdot, \cdot) = \inf_{y = y_h + y_{\mathtt{v}}} \left(\mathfrak{a}(y_h, y_h) + \sum_{\mathtt{v}} \mathfrak{a}(y_{\mathtt{v}}, y_{\mathtt{v}}) \right)$ in Theorem 4.5 defines a preconditioner C_K for K_N* (10) *in the following way. Let $J(\mathtt{v}) = \left[j_1^{\mathtt{v}}, \ldots, j_{n_{\mathtt{v}}}^{\mathtt{v}} \right]$ be the index set of all basis functions ϕ_j with supp $(\phi_j) \subset \Omega_{\mathtt{v}}$ and $J(h)$ the index set of all vertex functions (V). Due to the partition of $[\Phi]$ into vertex, edge, face and interior functions, the set $[\phi_j]_{j \in J(\mathtt{v})}$ forms a $n_{\mathtt{v}} = \mathcal{O}(p^d)$ dimensional basis of the space $\mathbb{V}_{\mathtt{v}}$. Let $P_{\mathtt{v}} \in \mathbb{R}^{n_{\mathtt{v}} \times N}$ be the Boolean matrix with the entries*

$$[P_{\mathtt{v}}]_{ij} = \begin{cases} 1 & \text{if} \quad j = j_i^{\mathtt{v}}, 1 \le i \le n_{\mathtt{v}} \\ 0 & \text{else} \end{cases}.$$

Finally, let

$$K_{\mathtt{v}} = \left[\mathfrak{a}(\phi_{j_i^{\mathtt{v}}}, \phi_{j_k^{\mathtt{v}}}) \right]_{i,k=1}^{n_{\mathtt{v}}}. \tag{19}$$

be the stiffness matrix on $\mathbb{V}_{\mathtt{v}}$. In the same way, P_h and K_h corresponding to the set $J(h)$ are introduced. Then, the splitting in Theorem 4.5 introduces the preconditioner

$$C_K^{-1} = P_h^\top K_h^{-1} P_h + \sum_{\mathtt{v}} P_{\mathtt{v}}^\top K_{\mathtt{v}}^{-1} P_{\mathtt{v}} \tag{20}$$

with $K_N \sim C_K$, see e.g. [49] for more details. In the same way, we introduce the preconditioners

$$C_Y^{-1} = P_h^\top Y_h^{-1} P_h + \sum_{\mathtt{v}} P_{\mathtt{v}}^\top Y_{\mathtt{v}}^{-1} P_{\mathtt{v}}, \quad C_M^{-1} = P_h^\top M_h^{-1} P_h + \sum_{\mathtt{v}} P_{\mathtt{v}}^\top M_{\mathtt{v}}^{-1} P_{\mathtt{v}} \tag{21}$$

with

$$Y_{\mathtt{v}} = \left[\mathfrak{p}(\phi_{j_i^{\mathtt{v}}}, \phi_{j_k^{\mathtt{v}}}) \right]_{i,k=1}^{n_{\mathtt{v}}}, \quad M_{\mathtt{v}} = \left[\langle \phi_{j_i^{\mathtt{v}}}, \phi_{j_k^{\mathtt{v}}} \rangle_\Omega \right]_{i,k=1}^{n_{\mathtt{v}}}.$$

Due to Theorem 4.5 and Corollary 4.6 they satisfy the spectral equivalence relations

$$Y_N \sim C_Y \text{ and } M_N \sim C_M. \tag{22}$$

4.3 Final condition number estimates

Summarizing, instead of the solution to systems with the matrices Y_N and M_N we have to solve systems with the patch matrices $Y_{\mathtt{v}}$ and $M_{\mathtt{v}}$ as well as systems with the h-FEM matrices Y_h and M_h, see (21). In the following, these solvers are briefly considered:

- The mass matrix M_h for h-FEM is well conditioned. Therefore, it can be replaced by its diagonal part $D_{M,h} = \operatorname{diag}(M_h)$.
- The matrix Y_h corresponds to the stiffness matrix according to the bilinear form
$$\mathfrak{y}(y,\tilde{y}) = \sqrt{\alpha}\int_\Omega \left(D(x)\nabla y\cdot\nabla\tilde{y} + c(x)y\tilde{y}\right)\,\mathrm{d}x + \int_\Omega y\tilde{y}\,\mathrm{d}x$$
Here, one can use multigrid methods, [25], or multilevel preconditioners as the BPX-preconditioner, [54], [12]. In order to get robustness with respect to the parameter α, the multigrid method of [36] should be used.
- The mass matrix $M_{\mathtt{v}}$ has a tensor product structure and the solution to a system with $M_{\mathtt{v}}$ can be performed in optimal arithmetical complexity, i.e. $\mathcal{O}(p^d)$ operations, see [9].
- The matrix $Y_{\mathtt{v}}$ has a size of $\mathcal{O}(p^d)$, where p denotes the polynomial degree. In the case of boundary concentrated FEM (BC-FEM), see [29], $p \approx \log N$. Then, a system with $Y_{\mathtt{v}}$ is performed in optimal arithmetical complexity. In the case of very high polynomial degrees, the matrix $Y_{\mathtt{v}}$ can be replaced by a wavelet preconditioner $C_{wavelet,Y}$, see [6]. It is based on the tensor product structure of one dimensional mass and stiffness matrices. Hence, it can be designed such that it is robust with respect to the regularization parameter α. However, the condition number $\kappa(C_{wavelet,Y}^{-1}Y_{\mathtt{v}})$ grows as $(1 + \log p)^3 \log^\chi \log p)$ for any $\chi > 1$.

Based on (21), we introduce the preconditioners

$$C_{Y,\mathtt{type}}^{-1} = P_h^\top C_{h,Y,\mathtt{type}}^{-1} P_h + \sum_{\mathtt{v}} P_{\mathtt{v}}^\top Y_{\mathtt{v}}^{-1} P_{\mathtt{v}} \quad \text{and} \quad C_{M,D}^{-1} = P_h^\top D_{M,h}^{-1} P_h + \sum_{\mathtt{v}} P_{\mathtt{v}}^\top M_{\mathtt{v}}^{-1} P_{\mathtt{v}}, \tag{23}$$

for Y_N and M_N respectively, where `type=BPX` stands for the BPX preconditioner and `type=Mult` stands for the multigrid preconditioner of [36].
Summarizing, the properties of the multigrid preconditioner, see [36] allow us to conclude that

$$C_{h,Y,\mathtt{Mult}} \sim Y_h. \tag{24}$$

By (24), (23) and (22) one obtains the spectral equivalence relation $Y_N \sim C_{Y,\mathtt{Mult}}$. In the same way, the relation $C_{M,D} \sim M_h$ implies $M_N \sim C_{M,D}$.
This means that we are able to prove the eigenvalue bounds

$$c_{1,Y} C_{Y,\mathtt{type}} \leq Y_N \leq c_{2,Y} C_{Y,\mathtt{type}} \quad \text{and} \quad c_{1,M} C_{M,D} \leq M_N \leq c_{2,M} C_{M,D} \tag{25}$$

with generic constants $c_{2,M} \geq c_{1,M} > 0$, $c_{2,Y} \geq c_{1,Y} > 0$. With (23), (25) the preconditioners

$$\hat{A}_N = \sigma \begin{bmatrix} c_{2,Y} C_{Y,\mathtt{Mult}} & \mathbf{0} \\ \mathbf{0} & \alpha c_{2,M} C_{M,D} \end{bmatrix} \quad \text{with } \sigma < 1 \tag{26}$$

for A_N and

$$\hat{S}_N = \tau \frac{\sqrt{3}}{2\alpha} \frac{c_{1,Y}}{c_{2,Y}} \left(c_{2,Y} C_{Y,\mathtt{Mult}}\right) \quad \text{with } \tau < 1 \tag{27}$$

for $S_N = B_N A_N^{-1} B_N^\top$, respectively, are introduced.
Therefore, we can prove the following two theorems.

Theorem 4.8. *Let $\hat{A}_N$ and $\hat{S}_N$ be defined by* (26) *and* (27)*, respectively. Moreover, let us assume that the estimates of the eigenvalue bounds* (25) *can be obtained in $\mathcal{O}(N)$ operations.*
If the maximal polynomial degree p grows as $\mathcal{O}(\log N)$, the saddle point system (9) *can be solved by the Schöberl-Zulehner PCG with the preconditioners $\mathcal{P}_{\mathrm{cg},\hat{A}_N,\hat{S}_N}$* (14) *in $\mathcal{O}(N)$ operations.*
For general polynomial degree p, the saddle point system (9) *can be solved by the Schöberl-Zulehner PCG in $\mathcal{O}(N\sqrt{(1+\log N)^3 \log^{\chi} \log N}))$ operations for any $\chi > 1$. The results are robust with respect to the regularization parameter α.*

Proof. If the eigenvalue bounds in (25) are known, one obtains for the scaled preconditioners $c_{2,Y}C_{Y,\mathtt{Mult}}$ and $c_{2,M}C_{M,D}$ the eigenvalue bounds

$$c_{2,Y}^{-1}c_{1,Y}(c_{2,Y}C_{Y,\mathtt{Mult}}) \leq Y_N \leq (c_{2,Y}C_{Y,\mathtt{Mult}}) \quad \text{and} \quad c_{2,M}^{-1}c_{1,M}(c_{2,M}C_{M,D}) \leq M_N \leq (c_{2,M}C_{M,D})$$

with the upper eigenvalue estimates one. Using (11) and (12), the results

$$\begin{aligned}\frac{2}{3}\min\left\{\frac{c_{1,Y}}{c_{2,Y}},\frac{c_{1,M}}{c_{2,M}}\right\}\begin{bmatrix}c_{2,Y}C_{Y,\mathtt{Mult}} & \mathbf{0}\\ \mathbf{0} & \alpha c_{2,M}C_{M,D}\end{bmatrix} \leq \quad A_N \quad &\leq \begin{bmatrix}c_{2,Y}C_{Y,\mathtt{Mult}} & \mathbf{0}\\ \mathbf{0} & \alpha c_{2,M}C_{M,D}\end{bmatrix}\\ &= \frac{1}{\sigma}\hat{A}_N < \hat{A}_N. \end{aligned} \tag{28}$$

on $\ker B_N$ and

$$\frac{\sqrt{3}}{2}\frac{c_{1,Y}}{c_{2,Y}}(c_{2,Y}C_{Y,\mathtt{Mult}}) \leq \quad \alpha B_N A_N^{-1}B_N^{\top} \quad \leq c_{2,Y}C_{Y,\mathtt{Mult}} \tag{29}$$

follow. In order to apply the Schöberl-Zulehner PCG with $\hat{A}_N$ (26) and $\hat{S}_N$ (27), the assumptions $\hat{S}_N \leq B_N\hat{A}_N^{-1}B_N^{\top}$ and $A_N < \hat{A}_N$ have to be checked. The last one has been proved in (28).
Moreover, with $\hat{S}_N = \tau\frac{\sqrt{3}}{2\alpha}\frac{c_{1,Y}}{c_{2,Y}}(c_{2,Y}C_{Y,\mathtt{Mult}})$ and $\tau < 1$, one gets

$$B_N\hat{A}_N^{-1}B_N^{\top} \overset{(28)}{\geq} B_N A_N^{-1}B_N^{\top} \overset{(29)}{\geq} \frac{\sqrt{3}}{2\alpha}\frac{c_{1,Y}}{c_{2,Y}}(c_{2,Y}C_{Y,\mathtt{Mult}}) \overset{(27)}{=} \tau^{-1}\hat{S}_N > \hat{S}_N.$$

Therefore, the Schöberl-Zulehner PCG can be applied. Note that all eigenvalue estimates are independent of α, h and p. Hence, the algorithm stops (for a fixed relative accuracy ε) after a bounded number of iterations. Since systems with $\hat{S}_N^{-1}$ and $\hat{A}_N$, or equivalently, with $C_{Y,\mathtt{Mult}}$ and $C_{M,D}$ can be performed in optimal complexity, this proves the assertion for the case $p \sim \mathcal{O}(\log N)$.
In the general case, the eigenvalue bounds of [6] for $\kappa(C_{wavelet,Y}^{-1}Y_{\mathrm{v}})$ enter the estimates. □

The preconditioners $\hat{A}_N$ (26) and $\hat{S}_N$ (27) require the constants in the eigenvalue bounds in (25). The application of the MINRES does not need these values. Using (23), the preconditioners

$$\tilde{A}_N = \begin{bmatrix}C_{Y,\mathtt{Mult}} & 0\\ 0 & \alpha C_{M,D}\end{bmatrix} \quad \text{and} \quad \tilde{S}_N = C_{Y,\mathtt{Mult}} \tag{30}$$

for A_N and $S_N = B_N A_N^{-1}B_N^{\top}$, respectively, are introduced.

Theorem 4.9. *Let $\tilde{A}_N$ and $\tilde{S}_N$ be defined by* (30)*, respectively. If the maximal polynomial degree p grows as $\mathcal{O}(\log N)$, the saddle point system* (9) *can be solved by the MINRES method with the preconditioner $\mathcal{P}_{\mathrm{minres},\tilde{A}_N,\tilde{S}_N}$* (15) *in $\mathcal{O}(N)$ operations.*
For general polynomial degree p, the saddle point system (9) *can be solved by the MINRES method in $\mathcal{O}(N\sqrt{(1+\log N)^3 \log^{\chi} \log N}))$ operations for any $\chi > 1$.*

Proof. The proof is similar to the proof of the previous theorem. As in (28) and (29), the estimates

$$A_N \sim \tilde{A}_N \quad \text{and} \quad \tilde{S}_N \sim B_N A_N^{-1} B_N^\top$$

follow. Therefore, all eigenvalues of the MINRES preconditioner $\mathcal{P}_{\text{minres}}$ (15) lie in the two real intervals $(-\mu_4, -\mu_3)$ and (μ_2, μ_1), where the constants $\mu_i > 0$ do not depend on the discretization parameter. Therefore, the convergence rate of MINRES is independent of h and p, see [23, Theorem 6.13]. If $p \sim \log N$, the systems with $\tilde{A}_N$ and $\tilde{S}_N$ can be solved in $\mathcal{O}(N)$ operations. □

Remark 4.10. *The disadvantage of the MINRES method is that the robustness with respect to α gets lost.*

5 Numerical experiments

Finally, some numerical experiments are presented.

5.1 Setting of the problems

In all examples, we choose $D(x) = c(x) = 1$ and a Neumann boundary, i.e. $\Gamma_{\mathcal{D}} = \emptyset$. For the domain Ω, two cases are distinguished:

- **Square**: Here, the domain is the unit square $\Omega = (0,1)^2$. The data y_d is chosen such that the solution to this optimal control problem is given by

$$y(x) = e^{\frac{1}{3}x_1^3 - x_1} e^{\frac{1}{3}x_2^3 - x_2}, \qquad\qquad q(x) = -y(x).$$

- **Hole:** In this case $\Omega = (0,3)^2 \backslash [1,2]^2$ and the desired state is given by

$$y_d(x) = 10 \sin(\pi x_1) + 5 \cos(\pi x_2^2),$$

whereas the exact solution is unknown.

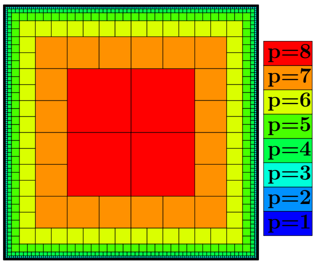

Figure 2: BC-refinement: unit square (left), hole (right)

Problem (2) is discretized by p-FEM or by BC-FEM, see Figure 2. The system (9) is solved by means of the Schöberl-Zulehner PCG or the preconditioned MINRES method with a relative accuracy of $\varepsilon = 10^{-10}$ for BC-FEM and $\varepsilon = 10^{-12}$ for p-FEM, respectively. The preconditioners of (21), i.e.

$$C^{-1}_{Y,\texttt{type}} = P_h^\top C^{-1}_{h,Y,\texttt{type}} P_h + \sum_{\mathtt{v}} P_{\mathtt{v}}^\top Y_{\mathtt{v}}^{-1} P_{\mathtt{v}} \quad \text{and} \quad C^{-1}_{M,D} = P_h^\top D^{-1}_{M,h} P_h + \sum_{\mathtt{v}} P_{\mathtt{v}}^\top M_{\mathtt{v}}^{-1} P_{\mathtt{v}}$$

are chosen for Y_N and M_N, respectively, where

- `type=BPX` uses the multilevel diagonally scaled BPX preconditioner for $C^{-1}_{h,Y,\mathtt{type}}$ and
- `type=E` uses the matrix $C_{h,Y,\mathtt{type}} = Y_h$, e.g. the h part is solved exactly.

In some examples, also the exact matrices Y_N instead of $C_{Y,\mathtt{type}}$ and M_N instead of $C_{M,D}$ are used as preconditioner, in order to check the influence of the different preconditioners.

5.2 Schöberl-Zulehner PCG

This solver requires the upper eigenvalue bounds of the preconditioner in the estimates (25). This is performed by an inverse vector or vector iteration. The results for the test example **Square** are displayed in Table 1.

N	$\alpha = 10^6$		$\alpha = 10^2$		$\alpha = 10^{-6}$		$\alpha = 10^{-2}$	
	$\kappa_{1,M}$	$\kappa_{1,Y}$	$\kappa_{1,M}$	$\kappa_{1,Y}$	$\kappa_{1,M}$	$\kappa_{1,Y}$	$\kappa_{1,M}$	$\kappa_{1,Y}$
75	4.5	1.5	4.5	1.5	4.5	1.3	4.5	2.2
243	5.8	4.0	5.8	4.0	5.8	4.1	5.8	5.7
507	5.9	4.0	5.9	4.0	5.9	4.1	5.9	5.8
867	6.1	4.0	6.1	4.0	5.9	4.1	5.9	5.9
1323	6.1	4.0	6.1	4.0	6.1	4.1	6.1	5.9
1875	6.2	4.0	6.2	4.0	6.2	4.1	6.2	5.9
2523	6.2	4.0	6.2	4.0	6.2	4.1	6.2	5.9
3267	6.2	4.0	6.2	4.0	6.2	4.1	6.2	5.9

Table 1: estimated parameters for different α and uniform p-refinement for $\kappa_{1,Y} = \lambda_{max}(C^{-1}_{Y,\mathrm{E}} Y_N)$ and $\kappa_{1,M} = \lambda_{max}(C^{-1}_{M,D} M_N)$.

From, the results one can see that the eigenvalue bounds are robust with respect to the regularization parameter. In all other examples, we choose now $\alpha = 1$. Table 2 displays the iteration numbers for the **square** example with p-FEM. In the experiments, the iteration numbers are bounded. Although the condition numbers of $C^{-1}_{Y,\mathrm{E}} Y_N$ and $C^{-1}_{M,D} M_N$ are quite low, the iteration numbers are about 200. This due to the scaling in order to guarantee the positive definiteness of the preconditioner.

N	75	243	507	867	1323	1875	2523	3267
M_N, Y_N	16	18	19	19	19	19	19	19
$C_{M,D}, C_{Y,\mathtt{PE}}$	74	196	199	224	199	224	202	216

Table 2: PCG iteration numbers for different preconditioners using p-FEM on the unit **square**.

5.3 MINRES

The first test example for the MINRES method uses the test example **square**. The iteration numbers are displayed in Figure 3. For the uniform p-FEM the iteration numbers are bounded and lower in comparison to the Schöberl-Zulehner-PCG. For the BC-FEM, there is a strong adaptive refinement to the boundary including hanging nodes. This influences the quality of the BPX preconditioner as h-FEM preconditioner, cf. if one compares the iteration numbers of the preconditioners $C_{Y,\mathtt{BPX}}$ and $C_{Y,\mathtt{E}}$ (red vs. yellow curve or green vs. light curve in the right picture of Figure 3).
An improvement is expected by using multigrid methods for the h-FEM part and or triangular/tetrahedral finite elements with red/green refinement in order to avoid hanging nodes.

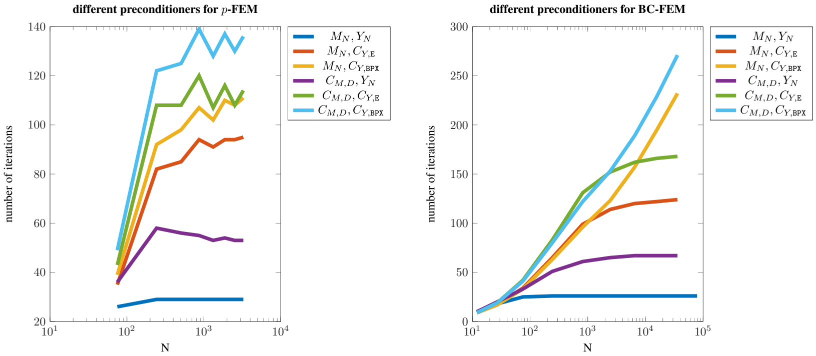

Figure 3: Iteration numbers of MINRES for uniform p refinement (left) and BC refinement (right) for the test example **Square**.

Figure 4 displays the iteration numbers for the preconditioned MINRES method for the test example **hole**. The behavior of the iteration numbers is the same. For uniform p-FEM the absolute numbers are higher than for the **square**.

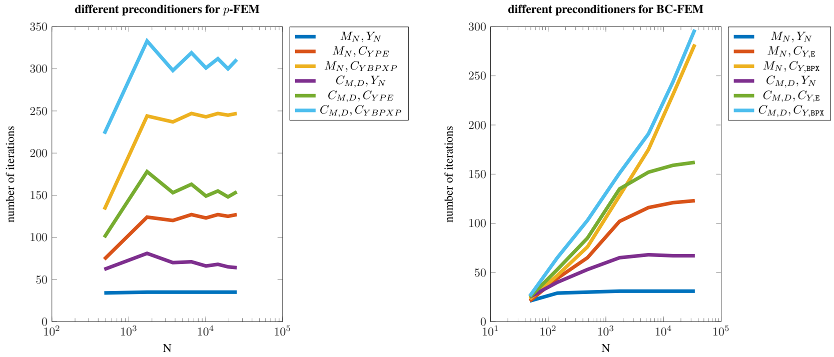

Figure 4: Iteration numbers of MINRES for uniform p refinement (left) and BC refinement (right) for the test example **Hole**.

Acknowledgement: The authors want to thank Prof. Walter Zulehner from the Institute for Computational Mathematics of the University of Linz, Austria for helpful hints.
This work has been supported by the FWF project P23484 and by the BIGS of the Haussdorff Center for Mathematics (HCM) .

References

[1] T. Apel, A. Rösch, and D. Sirch, *L^∞-error estimates on graded meshes with application to optimal control*, SIAM J. Control Optim., 48 (2009), pp. 1771–1796.

[2] O. Axelsson and M. Neytcheva, *Eigenvalue estimates for preconditioned saddle point matrices*, Numer. Linear Algebra Appl., 13 (2006), pp. 339–360.

[3] I. Babuška, A. Craig, J. Mandel, and J. Pitkäranta, *Efficient preconditioning for the p-version finite element method in two dimensions*, SIAM J. Numer. Anal., 28 (1991), pp. 624–661.

[4] R. Becker, H. Kapp, and R. Rannacher, *Adaptive finite element methods for optimal control of partial differential equations: basic concept*, SIAM J. Control Optim., 39 (2000), pp. 113–132 (electronic).

[5] S. Beuchler, *Wavelet solvers for hp-FEM discretizations in 3D using hexahedral elements*, Comput. Methods Appl. Mech. Engrg., 198 (2009), pp. 1138–1148.

[6] S. Beuchler, *Wavelet solvers for hp-FEM discretizations in 3D using hexahedral elements*, Comput. Methods Appl. Mech. Engrg., 198 (2009), pp. 1138–1148.

[7] S. Beuchler, K. Hofer, D. Wachsmuth, and J.-E. Wurst, *Boundary concentrated finite elements for optimal control problems with distributed observation*, Comput. Optim. Appl., 62 (2015), pp. 31–65.

[8] S. Beuchler, C. Pechstein, and D. Wachsmuth, *Boundary concentrated finite elements for optimal boundary control problems of elliptic PDEs*, Comput. Optim. Appl., 51 (2012), pp. 883–908.

[9] S. Beuchler and M. Purrucker, *Schwarz type solvers for hp-FEM discretizations of mixed problems*, Comput. Methods Appl. Math., 12 (2012), pp. 369–390.

[10] S. Beuchler, R. Schneider, and C. Schwab, *Multiresolution weighted norm equivalences and applications*, Numer. Math., 98 (2004), pp. 67–97.

[11] A. Borzì, *Multigrid methods for parabolic distributed optimal control problems*, J. Comput. Appl. Math., 157 (2003), pp. 365–382.

[12] J. H. Bramble, J. E. Pasciak, and J. Xu, *Parallel multilevel preconditioners*, Math. Comp., 55 (1990), pp. 1–22.

[13] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer Verlag, New York, 1994.

[14] F. Brezzi and M. Fortin, *Mixed and hybrid finite element methods*, vol. 15 of Springer Series in Computational Mathematics, Springer-Verlag, New York, 1991.

[15] E. Casas and M. Mateos, *Error estimates for the numerical approximation of Neumann control problems*, Comput. Optim. Appl., 39 (2008), pp. 265–295.

[16] E. Casas, M. Mateos, and F. Tröltzsch, *Necessary and sufficient optimality conditions for optimization problems in function spaces and applications to control theory*, in Proceedings of 2003 MODE-SMAI Conference, vol. 13 of ESAIM Proc., EDP Sci., Les Ulis, 2003, pp. 18–30 (electronic).

[17] Y. Chen, N. Yi, and W. Liu, *A Legendre-Galerkin spectral method for optimal control problems governed by elliptic equations*, SIAM J. Numer. Anal., 46 (2008), pp. 2254–2275.

[18] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[19] K. Deckelnick, A. Günther, and M. Hinze, *Finite element approximation of elliptic control problems with constraints on the gradient*, Numer. Math., 111 (2009), pp. 335–350.

[20] L. Demkowicz, *Computing with hp Finite Elements*, CRC Press, Taylor and Francis, 2006.

[21] L. Demkowicz, J. Kurtz, D. Pardo, M. Paszyński, W. Rachowicz, and A. Zdunek, *Computing with hp-adaptive finite elements. Vol. 2*, Chapman & Hall/CRC Applied Mathematics and Nonlinear Science Series, Chapman & Hall/CRC, Boca Raton, FL, 2008. Frontiers: three dimensional elliptic and Maxwell problems with applications.

[22] T. EIBNER AND J. M. MELENK, *Multilevel preconditioning for the boundary concentrated hp-FEM*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 3713–3725.

[23] H. C. ELMAN, D. J. SILVESTER, AND A. J. WATHEN, *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*, Numerical Mathematics and Scientific Computation, Oxford University Press, Oxford, second ed., 2014.

[24] M. ENGEL AND M. GRIEBEL, *A multigrid method for constrained optimal control problems*, J. Comput. Appl. Math., 235 (2011), pp. 4368–4388.

[25] W. HACKBUSCH, *Multigrid methods and applications*, vol. 4 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 1985.

[26] R. HERZOG AND E. SACHS, *Preconditioned conjugate gradient method for optimal control problems with control and state constraints*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 2291–2317.

[27] G. KARNIADAKIS AND S.J.SHERWIN, *Spectral/HP Element Methods for CFD*, Oxford University Press. Oxford, 1999.

[28] B. N. KHOROMSKIJ AND J. M. MELENK, *An efficient direct solver for the boundary concentrated FEM in 2D*, Computing, 69 (2002), pp. 91–117.

[29] B. N. KHOROMSKIJ AND J. M. MELENK, *Boundary concentrated finite element methods*, SIAM J. Numer. Anal., 41 (2003), pp. 1–36.

[30] V. KORNEEV, U. LANGER, AND L. S. XANTHIS, *On fast domain decomposition solving procedures for hp-discretizations of 3-D elliptic problems*, Comput. Methods Appl. Math., 3 (2003), pp. 536–559 (electronic).

[31] V. G. KORNEEV AND S. JENSEN, *Domain decomposition preconditioning in the hierarchical p-version of the finite element method*, Appl. Numer. Math., 29 (1999), pp. 479–518.

[32] A. KUNOTH, *Fast iterative solution of saddle point problems in optimal control based on wavelets*, Comput. Optim. Appl., 22 (2002), pp. 225–259.

[33] J.-F. MAITRE AND O. POURQUIER, *Condition number and diagonal preconditioning: comparison of the p-version and the spectral element methods*, Numer. Math., 74 (1996), pp. 69–84.

[34] H. MAURER AND H. D. MITTELMANN, *Optimization techniques for solving elliptic control problems with control and state constraints. II. Distributed control*, Comput. Optim. Appl., 18 (2001), pp. 141–160.

[35] C. MEYER AND A. RÖSCH, *Superconvergence properties of optimal control problems*, SIAM J. Control Optim., 43 (2004), pp. 970–985 (electronic).

[36] M. A. OLSHANSKII AND A. REUSKEN, *On the convergence of a multigrid method for linear reaction-diffusion problems*, Computing, 65 (2000), pp. 193–202.

[37] L. F. PAVARINO, *Additive Schwarz methods for the p-version finite element method*, Numer. Math., 66 (1994), pp. 493–515.

[38] L. F. PAVARINO, *Additive schwarz methods for the p-version finite element method*, Numer. Math., 66 (1994), pp. 493–515.

[39] L. F. PAVARINO, *Schwarz methods with local refinement for the p-version finite element method*, Numer. Math., 69 (1994), pp. 185–211.

[40] V. Pillwein and S. Takacs, *A local Fourier convergence analysis of a multigrid method using symbolic computation*, J. Symbolic Comput., 63 (2014), pp. 1–20.

[41] J. Schöberl, J. M. Melenk, C. Pechstein, and S. Zaglmayr, *Additive Schwarz preconditioning for p-version triangular and tetrahedral finite elements*, IMA J. Numer. Anal., 28 (2008), pp. 1–24.

[42] J. Schöberl, R. Simon, and W. Zulehner, *A robust multigrid method for elliptic optimal control problems*, SIAM J. Numer. Anal., 49 (2011), pp. 1482–1503.

[43] ——, *A robust multigrid method for elliptic optimal control problems*, SIAM J. Numer. Anal., 49 (2011), pp. 1482–1503.

[44] J. Schöberl and W. Zulehner, *Symmetric indefinite preconditioners for saddle point problems with applications to PDE-constrained optimization problems*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 752–773 (electronic).

[45] C. Schwab, *$p-$ and $hp-$finite element methods. Theory and applications in solid and fluid mechanics.*, Clarendon Press. Oxford, 1998.

[46] R. Simon and W. Zulehner, *On Schwarz-type smoothers for saddle point problems with applications to PDE-constrained optimization problems*, Numer. Math., 111 (2009), pp. 445–468.

[47] P. Solin, K. Segeth, and I. Dolezel, *Higher-Order Finite Element Methods*, Chapman and Hall, CRC Press, 2003.

[48] A. Toselli and O. Widlund, *Domain decomposition methods—algorithms and theory*, vol. 34 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 2005.

[49] A. Toselli and O. Widlund, *Domain Decomposition Methods- Algorithms and Theory*, Springer, 2005.

[50] F. Tröltzsch, *Optimale Steuerung partieller Differentialgleichungen: Theorie, Verfahren und Anwendungen*, Vieweg+Teubner Verlag, 2005.

[51] D. Wachsmuth and J.-E. Wurst, *Exponential convergence of hp-finite element discretization of optimal boundary problems with elliptic partial differential equations*, Preprint 328, Institut für Mathematik, Universität Würzburg, (2015).

[52] ——, *Optimal control of interface problems with hp-finite elements*, Numer. Funct. Anal. Optim., 37 (2016), pp. 363–390.

[53] H. Yserentant, *Über die Konvergenz von Mehrgitterverfahren für nichtuniform verfeinerte Familien von Gittern*, Z. Angew. Math. Mech., 64 (1984), pp. 324–326.

[54] X. Zhang, *Multilevel Schwarz methods*, Numer. Math., 63 (1992), pp. 521–539.

[55] W. Zulehner, *Nonstandard norms and robust estimates for saddle point problems*, SIAM J. Matrix Anal. Appl., 32 (2011), pp. 536–560.