MULTIGRID PRECONDITIONING OF LINEAR SYSTEMS FOR SEMISMOOTH NEWTON METHODS APPLIED TO OPTIMIZATION PROBLEMS CONSTRAINED BY SMOOTHING OPERATORS

ANDREI DRĂGĂNESCU*

Abstract. This article is concerned with the question of constructing efficient multigrid preconditioners for the linear systems arising when applying semismooth Newton methods to large-scale linear-quadratic optimization problems constrained by smoothing operators with box-constraints on the controls. It is shown that, for certain discretizations of the optimization problem, the linear systems to be solved at each semismooth Newton iteration reduce to inverting principal minors of the Hessian of the associated unconstrained problem. As in the case when box-constraints on the controls are absent, the multigrid preconditioner introduced here is shown to increase in quality as the mesh-size decreases, resulting in a number of iterations that decreases with mesh-size. However, unlike the unconstrained case, the spectral distance between the preconditioners and the Hessian is shown to be of suboptimal order in general.

Key words. multigrid, semismooth Newton methods, optimization with PDE constraints, large-scale optimization

AMS subject classifications. 65K10, 65M55, 65M32, 90C06

1. Introduction.

The objective of this article is to develop efficient multigrid preconditioners for the linear systems arising in the solution process of large-scale optimal control problems constrained by partial differential equations (PDEs) using semismooth Newton methods (SSNMs). The model problems under scrutiny have the form

$$\begin{cases} \text{minimize} & \mathcal{J}_\beta(u) \overset{\text{def}}{=} \frac{1}{2}\|\mathcal{K}u - y_d\|^2 + \frac{\beta}{2}\|u\|^2, \qquad \beta > 0 \text{ fixed} \\ \text{subject to}: & a \le u \le b \ \text{ a.e.}, \end{cases} \tag{1.1}$$

where $\Omega \subset \mathbb{R}^d$ is a bounded, open set, $\mathcal{K} : \mathcal{U} \to \mathcal{U}$ is a linear compact operator in the Hilbert space $\mathcal{U} = L^2(\Omega)$, and $a, b, y_d \in \mathcal{U}$ are given functions so that $a(x) < b(x)$ for almost all $x \in \Omega$. More detailed conditions will be given in Section 2.

In the PDE-constrained optimization literature $\mathcal{K}$ appears oftentimes as the composition of two operators $\mathcal{K} = I_{\mathcal{Y}\to\mathcal{U}}\mathcal{S}$, where $I_{\mathcal{Y}\to\mathcal{U}} : \mathcal{Y} \to \mathcal{U}$ is the compact embedding of a space $\mathcal{Y}$ into $\mathcal{U}$, and $\mathcal{S} : \mathcal{U} \to \mathcal{Y}$ is the solution operator of a linear PDE: if $e : \mathcal{Y} \times \mathcal{U} \to \mathcal{Y}^*$ defines the linear PDE

$$e(y, u) = 0 \ ,$$

then $e(y, u) = 0$ if and only if $y = \mathcal{S}u$. This way we obtain an equivalent formulation of (1.1) that has become standard in the PDE-constrained optimization literature [17]:

$$\begin{cases} \text{minimize} \ \ \frac{1}{2}\|y - y_d\|^2 + \frac{\beta}{2}\|u\|^2 \\ \text{subject to}: e(y, u) = 0, \ u \in \mathcal{U}_{\text{ad}} = \{u \in \mathcal{U} \ : \ a \le u \le b \ \text{ a.e.}\} \ . \end{cases} \tag{1.2}$$

[*] Department of Mathematics and Statistics, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, Maryland 21250 (draga@umbc.edu). This work was supported by the Department of Energy under contract no. de-sc0005455, and by the National Science Foundation under award DMS-1016177. Computations were performed in part on the UMBC's High Performance Computing Facility which was supported in part by the National Science Foundation under awards CNS-0821258 and DMS-0821311.

In this formulation u is the control and we shall call y the state. We shall also refer to (1.1) as the reduced form of (1.2). There are many applications, however, for which the formulation (1.1) may not correspond to a problem of the form (1.2), as is the case when $\mathcal{K}$ is an explicitly integral operator as in image deblurring. In other instances it may simply be that the reduced form (1.1) is more natural than (1.2). For example, in the solution process of the backwards parabolic equation discussed in [7] (see also the related work [1]) $\mathcal{K}u$ represents the time-T solution of a linear parabolic equation with initial value u; hence, the purpose of the optimization problem is to find an initial value u for which the time-T solution $\mathcal{K}u$ is close to a given state y_d. While this problem can be formulated as a PDE-constrained optimization (1.2) with space-time constraints, in a truly large-scale, four-dimensional setting the three-dimensional reduced form (1.1) of the optimization problem is more trackable. We should remark that both for the image-deblurring and the backwards heat-equation examples the operator $\mathcal{K}$ can be discretized at various resolutions, and the adjoint $\mathcal{K}^*$ can be computed at a cost comparable to that of $\mathcal{K}$. Hence, in this work we focus on the reduced problem (1.1).

Due to the availability of increasingly powerful parallel computers, the scientific community has shown a growing interest over the last decade in developing scalable solvers for large-scale optimization problems with PDE constraints. Multigrid methods have long been associated with large-scale linear systems, the paradigm being that the solution process can be significantly accelerated by using multiple resolutions of the same problem. However, the exact embodiment of the multigrid paradigm depends strongly on the class of problems considered, with multigrid methods for differential equations (elliptic, parabolic, flow problems) being significantly different from methods for integral equations. To place our problem in the context of multigrid methods we consider the simplified version of (1.1) obtained by removing the inequality constraints on the control, e.g. $\mathcal{U}_{\mathrm{ad}} = \mathcal{U}$, case in which (1.1) reduces to the linear system

$$(\mathcal{K}^*\mathcal{K} + \beta I)u = \mathcal{K}^* y_d \ , \tag{1.3}$$

which represents the normal equations associated with the Tikhonov regularization of the ill-posed problem

$$\mathcal{K}u = y_d \ . \tag{1.4}$$

Beginning with the works of Hackbusch [11] (see also [13]) much effort has been devoted to developing efficient multigrid methods for solving equations like (1.3) and (1.4), e.g. see [20, 21, 14, 19, 2, 7] and the references therein. For example, Drăgănescu and Dupont [7] have constructed a multigrid preconditioner $\mathcal{M}_h$ which satisfies

$$1 - C\frac{h^p}{\beta} \le \frac{\langle \mathcal{G}_h u, u\rangle}{\langle \mathcal{M}_h u, u\rangle} \le 1 + C\frac{h^p}{\beta}, \quad \forall u \neq 0, \tag{1.5}$$

where h is the mesh-size, $\mathcal{G}_h$ is the discretized version of $(\mathcal{K}^*\mathcal{K} + \beta I)$, and p is the order of the discretization ($p = 2$ for piecewise linear finite elements). We regard (1.5) as optimal-order scalability since it implies that the condition number of the unpreconditioned system $\mathrm{cond}(\mathcal{G}_h) = O(1/\beta)$ is being reduced by a factor of h^p in the $\mathcal{M}_h$-preconditioned system, namely $\mathrm{cond}(\mathcal{M}_h^{-1}\mathcal{G}_h) = O(h^p/\beta)$; this reduction is of optimal order given that the discretization order is p. A similar situation is encountered in classical multigrid for elliptic problems where multigrid is used to reduce the condition number from $O(h^{-2})$ to $O(1)$, the latter implying the desired mesh-independence property. We should point out that (1.5) implies that the number of iterations actually *decreases* with $h \downarrow 0$ to the point where, asymptotically, only one iteration is needed on very fine meshes.

The presence of explicit box constraints on the controls and/or states in PDE-constrained optimization problems is sometimes critical both for practical (design constraints) and theoretical reasons (e.g. states representing densities or concentrations of substances that have to be nonnegative). Methods for solving optimization problems with inequality constraints are fundamentally different and more involved than those for unconstrained problems; they generally fall into two competing categories: *interior point methods* (IPMs) and active-set methods such as SSNMs. Both types of methods exhibit superlinear local convergence and can be formulated and analyzed both in finite dimensional spaces as well as in function spaces [25, 27, 28], the latter being a critical step towards proving mesh-independence for the number of optimization steps. Both IPMs and SSNMs are iterative procedures that require the equivalent of a few PDE solves, i.e., applications of $\mathcal{K}$, for each iteration (called here *outer iteration*) and the solution one or two *inner* linear systems; for SSNMs only one inner linear solve is required, while for Mehrotra's predictor-corrector IPM implementation two inner linear solves are needed for each outer iteration. Efficiency of the solution process is measured by the number of outer iterations (ideally mesh-independent) needed to solve the problem to a desired tolerance and by the ability to solve the inner linear systems efficiently. In this work we concentrate on the latter.

Even though SSNMs and IPMs essentially solve the same problem, the linear algebra requirements for IPMs are different from those of SSNMs. If formulated in the reduced form (1.1), that is, with the PDE constraints eliminated (e.g. when the application of $\mathcal{K}$ is treated as a black-box) the structure of the systems arising in the IPM solution process is shown to be similar to the system (1.3) for the unconstrained problem [9]. More precisely, for IPMs we need to solve systems of the form

$$\mathcal{G}_{h,\lambda} u \overset{\text{def}}{=} (\mathcal{K}^*\mathcal{K} + \mathcal{D}_\lambda) u = b \ , \tag{1.6}$$

where $\mathcal{D}_\lambda$ is the multiplication operator with a relatively smooth function λ. Moreover, under specific conditions and for a natural discrete formulation of (1.1), Drăgănescu and Petra [9] have constructed multigrid preconditioners for the linear systems (1.6) that exhibit a certain degree of optimality similar to the one multigrid preconditioners in [7]: the resulting multigrid preconditioner $\mathcal{M}_h$ for the $\mathcal{G}_{h,\lambda}$ satisfies

$$1 - C\frac{h^2}{\beta}\|\lambda^{-\frac{1}{2}}\|_{W^2_\infty} \le \frac{\langle \mathcal{G}_{h,\lambda} u, u\rangle}{\langle \mathcal{M}_h u, u\rangle} \le 1 + C\frac{h^2}{\beta}\|\lambda^{-\frac{1}{2}}\|_{W^2_\infty}, \quad \forall u \neq 0 \ . \tag{1.7}$$

We recognize in (1.7) the optimal-order h^2-term (linear splines were used for discretization), but also remark that the quality of the preconditioner normally is affected by the lack of smoothness of λ: in general, as the solution approaches the boundary, the smoothness of λ is expected to degrade.

In this article we use similar ideas to design preconditioners for the linear systems arising in the SSNM solution process. Our main contribution is two-fold: we construct suitable coarse spaces and transition operators, and we give a detailed analysis of the resulting two- and multigrid preconditioners. We note that the basic elements of our analysis are related to those in [9], so in this sense the present work could be regarded as a companion of [9]. However, we should point out that the algorithms for

SSNMs are essentially different from those for IPMs, and so are the structures of their analyses. For SSNMs we will show that the linear systems to be solved are essentially principal subsystems of (1.3) where the selected rows (and columns) correspond to the constraints that are deemed inactive at some point in the solution process. The constructed multigrid preconditioner is shown to essentially satisfy (1.5) with $p = \frac{1}{2}$. While this order of approximation is clearly suboptimal, it still brings a significant reduction of the condition number if $\sqrt{h} \ll \beta$, and still results in a solution process that requires fewer and fewer inner linear iterations as $h \downarrow 0$.

The method developed and analyzed in this article is related to the multigrid method of the second kind developed by Hackbusch (see [12], Ch. 16). In fact, when using the multigrid iteration described in [12] in connection with the coarse spaces and transition operators described here we observe that the number of inner linear iterations decreases as $h \downarrow 0$. However, our numerical experiments in Section 5 indicate that our method is more efficient in absolute terms. We should also note that the coarse spaces defined in this work are related to the symmetric multigrid preconditioner developed by Hoppe and Kornhuber in [18] for obstacle problems, where the matrices to be preconditioned are subsystems of elliptic operators.

While our strategy is mainly designed for the reduced form (1.1), a significant literature is devoted to multigrid methods applied to the complementarity problem representing the Karush-Kuhn-Tucker (KKT) system of (1.2). Of these techniques we mention the collective smoothing multigrid method of Borzì and Kunisch [3]. For further references we refer the reader to the review article of Borzì and Schulz [4]. Also, in a recent article Stoll and Wathen [22] develop preconditioners for the linear systems arising in the SSNMs solution process of the unreduced optimal control problem (1.2); in their approach the linear systems are indefinite (since they correspond to derivatives of Lagrangians) and sparse, since it is not $\mathcal{K}$, but $\mathcal{K}^{-1}$ that is explicitly present in the system. Yet another alternative strategy for solving the linear systems for SSNMs applied to the unreduced problem (1.2), presented by Ulbrich in [25], p. 219ff (see also [24]) involves reducing the linear systems to solving the discrete PDEs for which efficient solvers are assumed to be readily available (such as classical multigrid for elliptic problems). The question of which method is the most efficient is difficult to answer for a general setting. However, we emphasize that the technique proposed in this article will work when the operator $\mathcal{K}$ is given only as a black-box, or when solvers are available for computing $\mathcal{K}u$ efficiently.

This article is organized as follows: in Section 2 we give the formal introduction of the problem, briefly discuss SSNMs, and present the main results. Section 3 is essentially devoted to proving the main result, Theorem 2.2, concerning the two-grid preconditioner, while in Section 4 we extend the analysis to the multigrid preconditioner. In Section 5 we show some numerical results to support our theoretical work, and we formulate some conclusions in Section 6. Appendix A contains some technical results used in Section 4.

2. Problem formulation and main results.

Our solution strategy follows the discretize-then-optimize paradigm, where we first formulate a discrete optimization problem associated with (1.1), which we then solve using SSNMs. After introducing the discrete framework in Section 2.1, we discuss the optimality conditions and their semismooth formulation in Section 2.2. In Section 2.3 we derive the linear systems needed to be solved at each SSNM iteration. The two-grid preconditioner and main two-grid results are given in Section 2.4. Furthermore, we discuss the multigrid preconditioner in Section 2.5.

2.1. Notation and discrete problem formulation. Let $\Omega \subset \mathbb{R}^d$ ($d = 1, 2$, or 3) be a bounded domain which, for simplicity, we assume to be polygonal (if $d = 2$) or polyhedral (for $d = 3$). We denote by $W_p^m(\Omega), H^m(\Omega), H_0^m(\Omega)$ (with $p \in [1, \infty], m \in \mathbb{N}$) the standard Sobolev spaces, and by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ the L^2-norm and inner product, respectively. Let $\widetilde{H}^{-m}(\Omega)$ be the dual (with respect to the L^2-inner product) of $H^m(\Omega) \cap H_0^1(\Omega)$ for $m > 0$, with the norm given by

$$\|u\|_{\widetilde{H}^{-m}(\Omega)} = \sup_{v \in H^m(\Omega) \cap H_0^1(\Omega)} \langle u, v\rangle / \|v\|_{H^m(\Omega)} \ .$$

The space of bounded linear operators on a Banach space X is denoted by $\mathfrak{L}(X)$. We regard square $n \times n$ matrices as operators in $\mathfrak{L}(\mathbb{R}^n)$ and we write matrices and vectors using bold font. If $\mathbf{A}$ is a symmetric positive definite matrix, we denote by $(\mathbf{u}, \mathbf{v})_{\mathbf{A}} = \mathbf{v}^T \mathbf{A} \mathbf{u}$ the $\mathbf{A}$-dot product of two vectors $\mathbf{u}, \mathbf{v}$, and by $|\mathbf{u}|_{\mathbf{A}} = \sqrt{(\mathbf{u}, \mathbf{u})_{\mathbf{A}}}$ the $\mathbf{A}$-norm; if $\mathbf{A} = \mathbf{I}$ we drop the subscript from the inner product and norm. The space of $m \times n$ matrices is denoted by $M_{m\times n}$; if $m = n$ we write M_n instead of $M_{n\times n}$. Given some norm $\|\cdot\|_s$ on a vector space $\mathcal{X}$, and $T \in \mathfrak{L}(\mathcal{X})$, we denote by $\|T\|_s$ the induced operator-norm

$$\|T\|_s = \sup_{u \in \mathcal{X},\ \|u\|_s = 1} \|Tu\|_s \ .$$

Consequently, if $T \in \mathfrak{L}(L^2(\Omega))$ then $\|T\|$ (no subscripts) is the L^2 operator-norm of T. If $\mathcal{X}$ is a Hilbert space and $T \in \mathfrak{L}(\mathcal{X})$ then $T^* \in \mathfrak{L}(\mathcal{X})$ denotes the adjoint of T. The defining elements of the discrete optimization problem are: the discrete analogues of $\mathcal{K}$, discrete norms, and discrete inequality constraints, all of which we introduce below.

To discretize the optimal control problem (1.1) we consider a sequence of quasi-uniform (in the sense of [6]) meshes $\mathcal{T}_j,\ j = 0, 1, 2, \ldots$, which we assume to be either simplicial (triangular if $d = 2$, tetrahedral if $d = 3$) or rectangular, and let

$$h_j = \max\{\mathrm{diam}(T)\ :\ T \in \mathcal{T}_j\}\ ,\quad j = 0, 1, 2, \ldots\ .$$

It is assumed that there are mesh-independent constants $0 < \underline{f} \le \overline{f} < 1$ (usually $\underline{f} = \overline{f} = 1/2$) so that

$$\underline{f} \le h_j / h_{j-1} \le \overline{f}\ .$$

We define the standard finite element spaces: for simplicial elements let

$$\mathcal{V}_j^s = \{u \in \mathcal{C}(\overline{\Omega})\ :\ \forall T \in \mathcal{T}_j,\quad u|_T \text{ is linear},\quad u|_{\partial\Omega} \equiv 0\}\ ,$$

and for rectangular we use piecewise tensor-products of linear polynomials

$$\mathcal{V}_j^r = \{u \in \mathcal{C}(\overline{\Omega})\ :\ \forall T \in \mathcal{T}_j,\quad u|_T\ \in \mathcal{Q}_1,\quad u|_{\partial\Omega} \equiv 0\}\ ,$$

where

$$\mathcal{Q}_1 = \left\{ \sum_j c_j \prod_{k=1}^d l_{j,k}(x_k)\ :\ l_{j,k} \text{ linear polynomial of one variable} \right\}\ .$$

For simplicity we assume that $\mathcal{T}_{j+1}$ is a uniform refinement of $\mathcal{T}_j$ so the associated spaces are nested

$$\mathcal{V}_j \subset \mathcal{V}_{j+1} \subset H_0^1(\Omega)\ .$$

Since the algorithms and results are the same for both types of finite element spaces we will denote by $\mathcal{V}_j$ either $\mathcal{V}_j^s$ or $\mathcal{V}_j^r$. Let $N_j = \dim(\mathcal{V}_j)$ and $P_1^{(j)}, \dots, P_{N_j}^{(j)}$ the nodes of $\mathcal{T}_j$ that lie in the interior of Ω, and define $\mathcal{I}_j : \mathcal{C}(\Omega) \to \mathcal{V}_j$ to be the standard interpolation operator

$$\mathcal{I}_j(u) = \sum_{i=1}^{N_j} u(P_i^{(j)})\varphi_i^{(j)} \ ,$$

where $\varphi_i^{(j)}, i = 1, \dots, N_j$ are the standard nodal basis functions. If we replace exact integration on an element T with vertices $P_1, \dots, P_\nu$ by the cubature

$$\int_T f(x)dx \approx \frac{\mathrm{vol}(T)}{\nu} \sum_{P \text{ vertex of } T} f(P) \ ,$$

then the L^2-inner product is approximated by the mesh-dependent inner product

$$\langle u, v\rangle_j = \sum_{i=1}^{N_j} w_i^{(j)} \ u(P_i^{(j)})v(P_i^{(j)}), \quad \text{for } u, v \in \mathcal{V}_j \ ,$$

where

$$w_i^{(j)} {=} \nu^{-1} \sum_{P_i^{(j)} \text{ vertex of } T} \mathrm{vol}(T) \ . \tag{2.1}$$

The discrete norms are then given by

$$\|\!|u|\!\|_j \stackrel{\text{def}}{=} \sqrt{\langle u, u\rangle_j} \ .$$

Since the quadrature/cubature is exact for linear functions, or tensor-products of linear functions, respectively, we have

$$\langle u, v\rangle_j = \int_\Omega \mathcal{I}_j(uv), \text{ for all } u, v \in \mathcal{V}_j \ .$$

Moreover, due to quasi-uniformity, there exist positive constants C_1, C_2 independent of $j \geq 0$ such that

$$C_1\|u\| \leq \|\!|u|\!\|_j \leq C_2\|u\|, \ \forall u \in \mathcal{V}_j \ . \tag{2.2}$$

We should point out that the norm-equivalence (2.2) extends to show mesh-independent equivalence of the associated operator-norms. We say that the weights $w_i^{(j)}$ are uniform with respect to the mesh $\mathcal{T}_j$ if there exists $w_j > 0$ independent of i so that

$$w_i^{(j)} = \omega_j h^d \ \text{ for } \ i = 1, \dots, N_j \ .$$

We call a triangulation *locally symmetric* if for every vertex P the associated nodal basis function φ is symmetric with respect to the reflection in P, that is,

$$\varphi(2P - x) = \varphi(x), \quad \forall x \in \Omega \ .$$

If a mesh is uniform, then it is locally symmetric and the weights $w_i^{(j)}$ are uniform.

On each space $\mathcal{V}_j$ consider an operator $\mathcal{K}_j \in \mathfrak{L}(\mathcal{V}_j)$ representing a discretization of $\mathcal{K}$. For the discrete operators we denote $\mathcal{K}_j^*$ to be the adjoint of $\mathcal{K}_j$ with respect to $\langle\cdot,\cdot\rangle_j$, that is,

$$\left\langle \mathcal{K}_j^* u, v\right\rangle_j = \left\langle u, \mathcal{K}_j v\right\rangle_j , \ \forall u, v \in \mathcal{V}_j \ .$$

We assume that the operators satisfy the following condition.

Condition 2.1. *There exists a constant $C = C(\mathcal{K})$ depending on $\mathcal{K}, \Omega, \mathcal{T}_0$ and independent of j so that the following hold:*

[a] *smoothing:*

$$\max(\|\mathcal{K}u\|_{H^m(\Omega)}, \|\mathcal{K}^*u\|_{H^m(\Omega)}) \le C\,\|u\|, \quad \forall u \in L^2(\Omega),\ m = 0,1,2\ ; \tag{2.3}$$

[b] *smoothed approximation:*

$$\|\mathcal{K}u - \mathcal{K}_j u\|_{H^m(\Omega)} \le C h_j^{2-m}\,\|u\|, \quad \forall u \in \mathcal{V}_j,\ m = 0,1,\ j \ge 0\ ; \tag{2.4}$$

[c] *uniform boundedness of discrete operators and their adjoints:*

$$\max(\|\mathcal{K}_j^* u\|_{L^\infty(\Omega)}, \|\mathcal{K}_j u\|_{L^\infty(\Omega)}) \le C\,\|u\|, \quad \forall u \in \mathcal{V}_j, \quad j \ge 0\ . \tag{2.5}$$

We now formulate the discrete optimization problem using the discrete norms and we enforce the inequality-constraints at the vertices:

$$\begin{cases} \text{minimize} & \mathcal{J}_{j,\beta}(u) \overset{\text{def}}{=} \frac{1}{2}\|\mathcal{K}_j u - y_d\|_j^2 + \frac{\beta}{2}\|u\|_j^2, \quad \beta > 0 \text{ fixed} \\ \text{subject to :} & u \in \mathcal{V}_j, \ \ a_i^{(j)} \le u(P_i^{(j)}) \le b_i^{(j)} \text{ for } i = 1,\dots,N_j\ , \end{cases} \tag{2.6}$$

where the vectors $\mathbf{a}^{(j)} = [a_1^{(j)},\dots,a_{N_j}^{(j)}]^T$ (resp.,$\mathbf{b}^{(j)} = [b_1^{(j)},\dots,b_{N_j}^{(j)}]^T$) represent L^2-projections of the functions a (resp., b). onto $\mathcal{V}_j$.

If $\mathbf{K}_j$ is the matrix representation of $\mathcal{K}_j$ in the nodal basis, $\mathbf{W}_j$ is the diagonal matrix with diagonal entries $w_1^{(j)}, w_2^{(j)},\dots,w_{N_j}^{(j)}$, and with $\mathbf{a}^{(j)}$, $\mathbf{b}^{(j)}$ defined earlier, the problem (2.6) reads in matrix form

$$\begin{cases} \text{minimize} & J_{j,\beta}(\mathbf{u}) \overset{\text{def}}{=} \frac{1}{2}\left|\mathbf{K}_j\mathbf{u} - \mathbf{y}_d\right|_{\mathbf{W}_j}^2 + \frac{\beta}{2}\left|\mathbf{u}\right|_{\mathbf{W}_j}^2, \qquad \beta > 0 \text{ fixed} \\ \text{subject to :} & \mathbf{u} \in \mathbb{R}^{N_j},\ \mathbf{a}^{(j)} \le \mathbf{u} \le \mathbf{b}^{(j)}\ . \end{cases} \tag{2.7}$$

Furthermore, we write $J_{j,\beta}(\mathbf{u}) = \frac{1}{2}\mathbf{u}^T\mathbf{C}_j\mathbf{u} - \mathbf{f}_j^T\mathbf{u} + \boldsymbol{\gamma}_j$, where

$$\mathbf{C}_j = \mathbf{K}_j^T\mathbf{W}_j\mathbf{K}_j + \beta\mathbf{W}_j, \ \ \mathbf{f}_j = \mathbf{K}_j^T\mathbf{W}_j\mathbf{y}_d, \ \ \boldsymbol{\gamma}_j = \frac{1}{2}\mathbf{y}_d^T\mathbf{W}_j\mathbf{y}_d\ . \tag{2.8}$$

We also point out that the adjoint operator $\mathcal{K}_j^*$ is represented by the matrix $\mathbf{W}_j^{-1}\mathbf{K}_j^T\mathbf{W}_j\mathbf{K}_j$, so we denote

$$\mathbf{K}_j^* \overset{\text{def}}{=} \mathbf{W}_j^{-1}\mathbf{K}_j^T\mathbf{W}_j\mathbf{K}_j\ .$$

2.2. Optimality conditions and SSNMs. Since $\mathcal{J}_\beta$ is strictly convex and quadratic, (1.1) has a unique solution $u \in L^2(\Omega)$ satisfying the KKT conditions (e.g. see [17, 23]): there exist $\lambda_a, \lambda_b \in L^2(\Omega)$ so that

$$\left\{ \begin{array}{l} \nabla \mathcal{J}_\beta + \lambda_b - \lambda_a = 0 \ , \\ a - u \leq 0, \ \ \lambda_a \geq 0, \ \ (a-u)\lambda_a = 0 \ \ \text{a.e.}, \\ u - b \leq 0, \ \ \lambda_b \geq 0, \ \ (u-b)\lambda_b = 0 \ \ \text{a.e.} \end{array} \right. \tag{2.9}$$

After denoting $\lambda = \lambda_a - \lambda_b$, the complementarity system (2.9) can be written as a nonlinear, non-smooth system

$$\left\{ \begin{array}{l} \nabla \mathcal{J}_\beta - \lambda = 0 \ , \\ \lambda - \max(0, \lambda + \sigma(a-u)) - \min(0, \lambda + \sigma(b-u)) = 0 \ , \end{array} \right. \tag{2.10}$$

where $\sigma > 0$ is an arbitrary constant, and $\lambda_a = \max(0, \lambda), \ \lambda_b = -\min(0, \lambda)$. Since $\nabla \mathcal{J}_\beta(u) = \mathcal{K}^*(\mathcal{K}u - y_d) + \beta u$, for $\sigma = \beta$ the system (2.10) is equivalent to

$$\nabla \mathcal{J}_\beta - \max(0, \mathcal{K}^*(\mathcal{K}u - y_d) + \beta a) - \min(0, \mathcal{K}^*(\mathcal{K}u - y_d) + \beta b) = 0 \ . \tag{2.11}$$

Cf. [26], if $a, b \in L^s(\Omega)$ with $s > 2$, then Condition 2.1 implies that the nonlinear function in (2.11) is semismooth (slantly differentiable); therefore Newton's method converges superlinearly when applied to (2.11).

To simplify notation, for the remainder of this section we omit the sub- or superscripts 'j' if there is no danger of confusion; hence $J_\beta = J_{j,\beta}$, $\mathbf{C} = \mathbf{C}_j$, $\mathbf{a} = \mathbf{a}^{(j)}$, $N = N_j$, etc. The KKT and associated semismooth equation for (2.7) is similar to the continuous case, namely the solution $\mathbf{u}$ satisfies the following complementarity problem: there exist vectors $\boldsymbol{\lambda}_a, \boldsymbol{\lambda}_b \in \mathbb{R}^N$ so that

$$\left\{ \begin{array}{l} \nabla J_\beta + \boldsymbol{\lambda}_b - \boldsymbol{\lambda}_a = \mathbf{0} \ , \\ \mathbf{a} - \mathbf{u} \leq \mathbf{0}, \ \ \boldsymbol{\lambda}_a \geq 0, \ \ (\mathbf{a} - \mathbf{u}) \cdot \boldsymbol{\lambda}_a = \mathbf{0} \ , \\ \mathbf{u} - \mathbf{b} \leq \mathbf{0}, \ \ \boldsymbol{\lambda}_b \geq 0, \ \ (\mathbf{u} - \mathbf{b}) \cdot \boldsymbol{\lambda}_b = \mathbf{0} \ , \end{array} \right. \tag{2.12}$$

where $\mathbf{u} \cdot \mathbf{v}$ is the componentwise vector multiplication. Since $\nabla J_\beta(u) = \mathbf{C}\mathbf{u} - \mathbf{f}$ and $\mathbf{W}$ is diagonal, after denoting $\hat{\boldsymbol{\lambda}} = \mathbf{W}^{-1}(\boldsymbol{\lambda}_a - \boldsymbol{\lambda}_b)$ and left-multiplying the first equation in (2.12) by $\mathbf{W}^{-1}$, the complementarity system (2.12) is written as a nonlinear, non-smooth system

$$\left\{ \begin{array}{l} (\mathbf{K}^*\mathbf{K} + \beta \mathbf{I})\mathbf{u} - \mathbf{K}^*\mathbf{y}_d - \hat{\boldsymbol{\lambda}} = \mathbf{0} \ , \\ \hat{\boldsymbol{\lambda}} - \max(\mathbf{0}, \hat{\boldsymbol{\lambda}} + \sigma(\mathbf{a} - \mathbf{u})) - \min(\mathbf{0}, \hat{\boldsymbol{\lambda}} + \sigma(\mathbf{b} - \mathbf{u})) = \mathbf{0} \ , \end{array} \right. \tag{2.13}$$

since $\mathbf{W}^{-1}\mathbf{f} = \mathbf{K}^*\mathbf{y}_d$. For $\sigma = \beta$, (2.13) is equivalent to

$$\left\{ \begin{array}{l} \hat{\boldsymbol{\lambda}} = (\mathbf{K}^*\mathbf{K} + \beta \mathbf{I})\mathbf{u} - \mathbf{K}^*\mathbf{y}_d \ , \\ \hat{\boldsymbol{\lambda}} - \max(\mathbf{0}, \mathbf{K}^*(\mathbf{K}\mathbf{u} - \mathbf{y}_d) + \beta \mathbf{a}) - \min(\mathbf{0}, \mathbf{K}^*(\mathbf{K}\mathbf{u} - \mathbf{y}_d) + \beta \mathbf{b}) = \mathbf{0} \ . \end{array} \right. \tag{2.14}$$

While the resemblance between the discrete system (2.14) and its continuous counterpart (2.11) extends beyond notation, it is neither evident nor is it the purpose of this paper to prove mesh-independence of the convergence rate of Newton's method for (2.14); however, we *do* observe mesh-independence in numerical computations even for $\sigma \neq \beta$. We should also note that the mesh-independence results of Hintermüller and Ulbrich in [16] do not apply directly to (2.14) mainly because here we discretize the control using continous piecewise linear functions rather than piecewise constants as in [16], so the operator $\varphi \mapsto \max(0, \varphi)$ is not well defined in $\mathcal{V}_j$: if $\varphi \in \mathcal{V}_j$ has both negative and positive values on an element $T \in \mathcal{T}_j$, then $\max(0, \varphi)$ is no longer in $\mathcal{V}_j$. This is the main reason for which we resort to a matrix formulation of the discrete optimization problem (2.6) prior to formulating a discrete semismooth Newton system, instead of formulating a semismooth Newton system directly for finite element spaces. However, our primary interest is to construct multigrid preconditioners for the linear systems arising in the solution process of (2.14) which we discuss in the next section.

2.3. Primal-dual formulation and linear systems. As shown in [15], the semismooth Newton method for (2.7) can be regarded as a primal-dual active set method which we now describe briefly. If $\mathbf{u}$, $\hat{\boldsymbol{\lambda}}$ solve (2.13), then we define the following sets:

$$\begin{aligned}
\mathcal{I} &= \{i \in \{1,\dots,N\} \ : \ \hat{\boldsymbol{\lambda}}_i + \sigma(\mathbf{a}_i - \mathbf{u}_i) < 0 \ \text{ and } \ \hat{\boldsymbol{\lambda}}_i + \sigma(\mathbf{b}_i - \mathbf{u}_i) > 0\} \ ,\\
\mathcal{A}^a &= \{i \in \{1,\dots,N\} \ : \ \hat{\boldsymbol{\lambda}}_i + \sigma(\mathbf{a}_i - \mathbf{u}_i) \geq 0 \ \text{ and } \ \hat{\boldsymbol{\lambda}}_i + \sigma(\mathbf{b}_i - \mathbf{u}_i) > 0\} \ ,\\
\mathcal{A}^b &= \{i \in \{1,\dots,N\} \ : \ \hat{\boldsymbol{\lambda}}_i + \sigma(\mathbf{a}_i - \mathbf{u}_i) < 0 \ \text{ and } \ \hat{\boldsymbol{\lambda}}_i + \sigma(\mathbf{b}_i - \mathbf{u}_i) \leq 0\} \ .
\end{aligned}$$

A simple argument shows that if $i \in \mathcal{I}$, then $\mathbf{a}_i < \mathbf{u}_i < \mathbf{b}_i$ and $\hat{\boldsymbol{\lambda}}_i = 0$; on the other hand, if $i \in \mathcal{A}^a$, then $\mathbf{a}_i = \mathbf{u}_i < \mathbf{b}_i$ and $\hat{\boldsymbol{\lambda}}_i \geq 0$; finally, if $i \in \mathcal{A}^b$, then $\mathbf{a}_i < \mathbf{u}_i = \mathbf{b}_i$ and $\hat{\boldsymbol{\lambda}}_i \leq 0$. Also, since $\mathbf{a}_i < \mathbf{b}_i$ for all i, we have $\mathcal{I} \cup \mathcal{A}^a \cup \mathcal{A}^b = \{1,\dots,N\}$. The primal-dual active set method produces a sequence of sets $(\mathcal{I}_k, \mathcal{A}^a_k, \mathcal{A}^b_k)_{k=1,2,\dots}$ that approximate $(\mathcal{I}, \mathcal{A}^a, \mathcal{A}^b)$. Given $(\mathcal{I}_k, \mathcal{A}^a_k, \mathcal{A}^b_k)$, we set the system

$$\begin{cases}
(\mathbf{K}^*\mathbf{K} + \beta\mathbf{I})\mathbf{u}^{(k+1)} - \hat{\boldsymbol{\lambda}}^{(k+1)} = \mathbf{K}^*\mathbf{y}_d \ ,\\
\mathbf{u}_i^{(k+1)} = \mathbf{a}_i, \ \text{ for } \ i \in \mathcal{A}^a_k, \quad \mathbf{u}_i^{(k+1)} = \mathbf{b}_i, \ \text{ for } \ i \in \mathcal{A}^b_k \ ,\\
\hat{\boldsymbol{\lambda}}_i^{(k+1)} = 0, \ \text{ for } \ i \in \mathcal{I}_k \ .
\end{cases} \tag{2.15}$$

The new set-iterates are then given by

$$\begin{aligned}
\mathcal{I}_{k+1} &= \{i \ : \ \hat{\boldsymbol{\lambda}}_i^{(k+1)} + \sigma(\mathbf{a}_i - \mathbf{u}_i^{(k+1)}) < 0 \ \text{ and } \ \hat{\boldsymbol{\lambda}}_i^{(k+1)} + \sigma(\mathbf{b}_i - \mathbf{u}_i^{(k+1)}) > 0\} \ ,\\
\mathcal{A}^a_{k+1} &= \{i \ : \ \hat{\boldsymbol{\lambda}}_i^{(k+1)} + \sigma(\mathbf{a}_i - \mathbf{u}_i^{(k+1)}) \geq 0 \ \text{ and } \ \hat{\boldsymbol{\lambda}}_i^{(k+1)} + \sigma(\mathbf{b}_i - \mathbf{u}_i^{(k+1)}) > 0\} \ ,\\
\mathcal{A}^b_{k+1} &= \{i \ : \ \hat{\boldsymbol{\lambda}}_i^{(k+1)} + \sigma(\mathbf{a}_i - \mathbf{u}_i^{(k+1)}) < 0 \ \text{ and } \ \hat{\boldsymbol{\lambda}}_i^{(k+1)} + \sigma(\mathbf{b}_i - \mathbf{u}_i^{(k+1)}) \leq 0\} \ .
\end{aligned}$$

Now consider the splitting of $\mathbf{G} = \mathbf{K}^*\mathbf{K} + \beta\mathbf{I}$ according to the sets of indices $\mathcal{I}_k$ and $\mathcal{A}_k = \mathcal{A}^a_k \cup \mathcal{A}^b$. More precisely, using MATLAB syntax, we define

$$\mathbf{G}^{\text{in},k} = \mathbf{G}(\mathcal{I}_k, \mathcal{I}_k) \ , \quad \mathbf{G}^{\text{ia},k} = \mathbf{G}(\mathcal{I}_k, \mathcal{A}_k), \quad \mathbf{G}^{\text{ai},k} = \mathbf{G}(\mathcal{A}_k, \mathcal{I}_k), \quad \mathbf{G}^{\text{aa},k} = \mathbf{G}(\mathcal{A}_k, \mathcal{A}_k) \ .$$

Since $\mathbf{u}_i^{(k+1)}$ is determined for $i \in \mathcal{A}_k$, and $\hat{\boldsymbol{\lambda}}_i^{(k+1)} = 0$ for $i \in \mathcal{I}_k$, the equations corresponding to indices from $\mathcal{I}_k$ in the first system of (2.15) reduce to

$$\mathbf{G}^{\text{in},k}\mathbf{u}_{\text{in}}^{(k+1)} = (\mathbf{K}^*\mathbf{y}_d)_{\text{in}} - \mathbf{G}^{\text{ia},k}\mathbf{u}_{\text{a}}^{(k+1)} \ , \tag{2.16}$$

where $\mathbf{u}_{\text{in}}^{(k+1)} = \mathbf{u}^{(k+1)}(\mathcal{I}_k)$, $\mathbf{u}_{\text{a}}^{(k+1)} = \mathbf{u}^{(k+1)}(\mathcal{A}_k)$, $(\mathbf{K}^*\mathbf{y}_d)_{\text{in}} = (\mathbf{K}^*\mathbf{y}_d)(\mathcal{I}_k)$. After solving (2.16), the remaining components of $\hat{\boldsymbol{\lambda}}^{(k+1)}$ are identified by

$$\hat{\boldsymbol{\lambda}}^{(k+1)}(\mathcal{A}_{k+1}) = \mathbf{G}^{\text{ai},k}\mathbf{u}_{\text{in}}^{(k+1)} + \mathbf{G}^{\text{aa},k}\mathbf{u}_{\text{a}}^{(k+1)} - (\mathbf{K}^*\mathbf{y}_d)(\mathcal{A}_{k+1}) \ .$$

Thus the critical step in solving (2.15) is the reduced system (2.16).

We should point out that for large-scale problems the matrices $\mathbf{K}$, and thus also $\mathbf{G}$ and $\mathbf{G}^{\mathrm{in},k}$, are expected to be dense and are not formed. Therefore we can solve (2.16) only by means of iterative solvers. Hence, efficient matrix-free preconditioners are necessary for accelerating the solution process. Also note that if $\mathcal{A}_k = \emptyset$, then the multigrid preconditioners developed in [7] can be used. Our main contribution in this work is the design of two- and multigrid preconditioners for (2.16) for the case when $\mathcal{A}_k \neq \emptyset$.

2.4. The two-grid preconditioner.

Let $j \geq 1$ be a fixed level, to which we shall refer as the fine level. We will first define a two-grid preconditioner for the matrix $\mathbf{G}_j^{\mathrm{in},k}$ of (2.16) that will involve inverting a certain principal minor of $\mathbf{G}_{j-1}$. To design the two-grid preconditioner we will regard the matrix $\mathbf{G}_j^{\mathrm{in},k}$ as an operator between finite element spaces.

2.4.1. Construction of the two-grid preconditioner.

To simplify notation, consider a fixed index set $\mathcal{I}^{(j)} \subseteq \{1, \ldots, N_j\}$; this should be regarded as one of the iterates $\mathcal{I}_k^{(j)}$ encountered in the solution process described in the previous section. The linear system we investigate has the form

$$\mathbf{G}_j^{\mathrm{in}} \tilde{\mathbf{u}} = \tilde{\mathbf{r}} \ , \tag{2.17}$$

where $\mathbf{G}_j^{\mathrm{in}} = \mathbf{G}_j(\mathcal{I}^{(j)}, \mathcal{I}^{(j)}), \tilde{\mathbf{u}}, \tilde{\mathbf{r}} \in \mathbb{R}^{|\mathcal{I}^{(j)}|}$. We will call vertex P_i or an index i *inactive* if $i \in \mathcal{I}^{(j)}$. Define the *fine inactive space* by

$$\mathcal{V}_j^{\mathrm{in}} = \mathrm{span}\{\varphi_i^{(j)} \ : \ i \in \mathcal{I}^{(j)}\} \ ,$$

and the *fine inactive domain* by

$$\Omega_j^{\mathrm{in}} = \bigcup_{i \in \mathcal{I}^{(j)}} \mathrm{supp}(\varphi_i^{(j)}) \ , \tag{2.18}$$

where $\mathrm{supp}(u)$ is the support of the function u. The critical component of the preconditioner is the definition of the *coarse inactive index set*:

$$\mathcal{I}^{(j-1)} \overset{\mathrm{def}}{=} \{i \in \{1, \ldots, N_{j-1}\} \ : \ \mathrm{supp}(\varphi_i^{(j-1)}) \subseteq \Omega_j^{\mathrm{in}}\} \ . \tag{2.19}$$

To be more precise, if i_f is the index in the fine numbering associated to the coarse index i_c, that is, $P_{i_f}^{(j)} = P_{i_c}^{(j-1)}$, the definition above is equivalent to saying that $i_c \in \mathcal{I}^{(j-1)}$ if i_f together with all its neighbouring *fine* indices are inactive. In Figure 2.1 we depict a set of inactive fine nodes by filled circles, and the associated coarse inactive nodes by hollow circles. The coarse nodes that are not inactive are shown with a square hollow marker. The *coarse inactive space* is now defined to be

$$\mathcal{V}_{j-1}^{\mathrm{in}} = \mathrm{span}\{\varphi_i^{(j-1)} \ : \ i \in \mathcal{I}^{(j-1)}\} \ , \tag{2.20}$$

and the *coarse inactive domain* is given by

$$\Omega_{j-1}^{\mathrm{in}} = \bigcup_{i \in \mathcal{I}^{(j-1)}} \mathrm{supp}(\varphi_i^{(j-1)}) \ .$$

Note that $\mathcal{V}_{j-1}^{\mathrm{in}} \subseteq \mathcal{V}_j^{\mathrm{in}}$ and $\Omega_{j-1}^{\mathrm{in}} \subseteq \Omega_j^{\mathrm{in}}$. In connection with Ω_j^{in} we also define *numerical interior* $\mathrm{Int_n}\Omega_j^{\mathrm{in}}$ of Ω_j^{in} (relative to $\Omega_{j-1}^{\mathrm{in}}$) to be the union of all *coarse* elements T

included in the fine inactive domain, that is, $T \subseteq \Omega_j^{\mathrm{in}}$, and whose vertices are either in $\mathcal{I}^{(j-1)}$ or lie on the boundary of Ω (see Figure 2.1). Furthermore, let the *numerical boundary* of Ω_j^{in} (relative to $\Omega_{j-1}^{\mathrm{in}}$) be given by

$$\partial_{\mathrm{n}}\Omega_j^{\mathrm{in}} = \Omega_j^{\mathrm{in}} \setminus \mathrm{Int}_{\mathrm{n}}\Omega_h^{\mathrm{in}} \ .$$

Note that $\mathrm{Int}_{\mathrm{n}}\Omega_j^{\mathrm{in}} \subseteq \Omega_{j-1}^{\mathrm{in}}$.

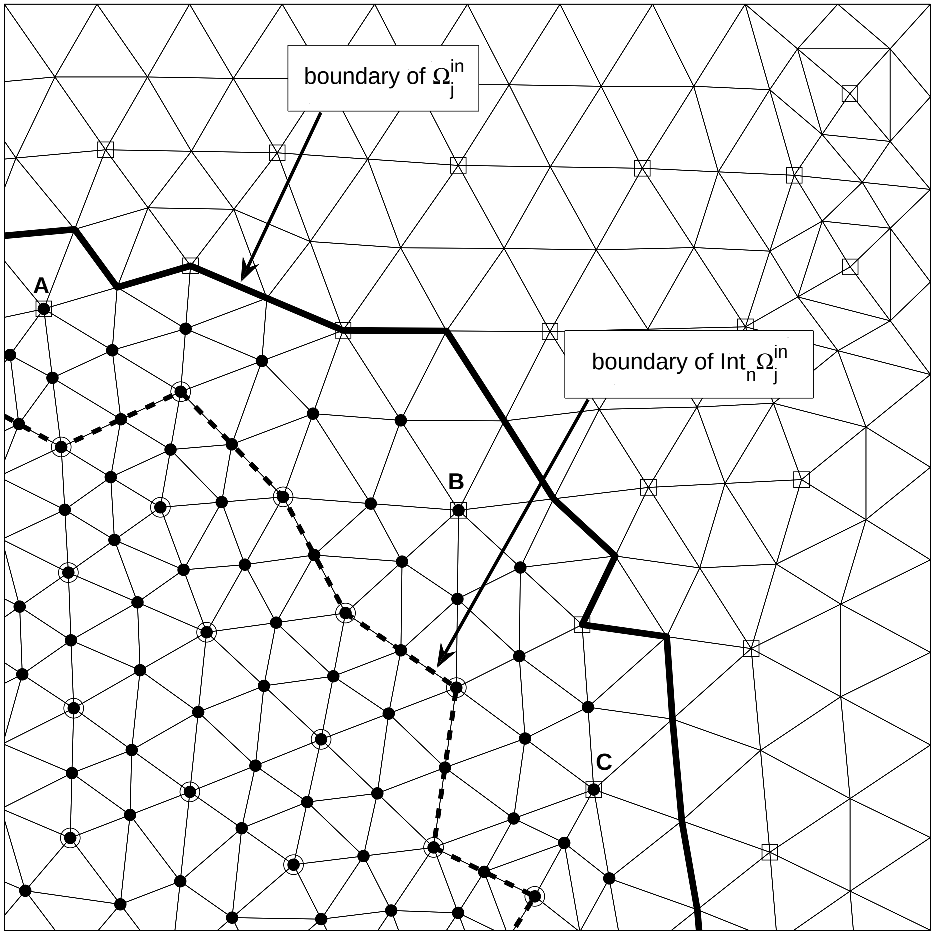
boundary of Ω_j^{in}
boundary of $\mathrm{Int}_n\Omega_j^{in}$
A
B
C

FIG. 2.1. *Inactive fine nodes are marked with a dot, inactive coarse nodes are marked with a circle, and active coarse nodes are marked with a square. Note that the coarse nodes A, B, C are active because the supports of the corresponding nodal basis functions are not included in Ω_j^{in}, even though the nodes themselves lie in the interior of Ω_j^{in}. The area between the solid and the dotted lines lies in the numerical boundary of Ω_j^{in}.*

Let now $\mathcal{W}_{j-1}^{\mathrm{in}}$ be the L^2-orthogonal complement of $\mathcal{V}_{j-1}^{\mathrm{in}}$ in $\mathcal{V}_j^{\mathrm{in}}$ and define the L^2-projectors

$$\pi_{j-1}^{\mathrm{in}} : \mathcal{V}_j^{\mathrm{in}} \to \mathcal{V}_{j-1}^{\mathrm{in}} \ , \qquad \rho_{j-1}^{\mathrm{in}} : \mathcal{V}_j^{\mathrm{in}} \to \mathcal{W}_{j-1}^{\mathrm{in}} \ ,$$

so $\pi_{j-1}^{\mathrm{in}} + \rho_{j-1}^{\mathrm{in}}$ is the identity on $\mathcal{V}_j^{\mathrm{in}}$. Furthermore, let $\mathcal{E}_j^{\mathrm{in}} : \mathcal{V}_j^{\mathrm{in}} \to \mathcal{V}_j$ be the natural embedding obtained by extending a function with zero outside of Ω_j^{in}. We may occasionally omit the explicit use of the embedding operators to ease notation. We also define the restriction $R_{j-1} : \mathcal{V}_j \to \mathcal{V}_{j-1}$ as the adjoint with respect to $\langle \cdot, \cdot \rangle_j$ to the embedding of $\mathcal{V}_{j-1}$ in $\mathcal{V}_j$, that is, for $u \in \mathcal{V}_j$

$$\langle u, v \rangle_j = \langle R_{j-1} u, v \rangle_{j-1} \ , \quad \forall v \in \mathcal{V}_{j-1} \ , \tag{2.21}$$

and let $P_j^{\mathrm{in}} : \mathcal{V}_j \to \mathcal{V}_j^{\mathrm{in}}$ the projection with respect to $\langle \cdot, \cdot \rangle_j$ given by

$$P_j^{\mathrm{in}} \left(\sum_{i=1}^{N_j} u_i \varphi_i^{(j)} \right) = \sum_{i \in \mathcal{I}^{(j)}} u_i \varphi_i^{(j)} \ .$$

Furthermore, denote by π_j the orthogonal L^2-projector onto the space $\mathcal{V}_j$. From the equivalence (2.2) of the discrete norms with the L^2-norm it follows that

$$\|R_j u\| \le C \|u\| \ , \ j = 0, 1, \dots, \ . \tag{2.22}$$

for some mesh-independent constant C.

The matrix $\mathbf{G}_j^{\mathrm{in}}$ in (2.17) represents the operator

$$\mathcal{G}_j^{\mathrm{in}} = P_j^{\mathrm{in}} \left(\mathcal{K}_j^* \mathcal{K}_j + \beta I \right) \mathcal{E}_j^{\mathrm{in}} \ , \tag{2.23}$$

and thus matrix-vector products for $\mathbf{G}_j^{\mathrm{in}}$ are computed accordingly. We define a *two-grid preconditioner* M_j^{in} for $\mathcal{G}_j^{\mathrm{in}}$ as in Drăgănescu and Dupont [7] for the unconstrained case:

$$\mathcal{M}_j^{\mathrm{in}} = \overbrace{P_{j-1}^{\mathrm{in}} \left(\mathcal{K}_{j-1}^* \mathcal{K}_{j-1} + \beta I \right) \mathcal{E}_{j-1}^{\mathrm{in}}}^{\mathcal{G}_{j-1}^{\mathrm{in}}} \pi_{j-1}^{\mathrm{in}} + \beta \rho_{j-1}^{\mathrm{in}} \ . \tag{2.24}$$

Note that the inverse $\mathcal{S}_j^{\mathrm{in}}$ of the preconditioner M_j^{in} has the explicit form

$$\mathcal{S}_j^{\mathrm{in}} \overset{\mathrm{def}}{=} (\mathcal{M}_j^{\mathrm{in}})^{-1} = \left(\mathcal{G}_{j-1}^{\mathrm{in}} \right)^{-1} \pi_{j-1}^{\mathrm{in}} + \beta^{-1} \rho_{j-1}^{\mathrm{in}} \ . \tag{2.25}$$

Since none of the matrices representing $\mathcal{G}_j^{\mathrm{in}}$ or $\mathcal{M}_j^{\mathrm{in}}$ are formed, it is the operator $\mathcal{S}_j^{\mathrm{in}}$ that we need to apply in practice, so the explicit formula (2.25) for $(\mathcal{M}_j^{\mathrm{in}})^{-1}$ is essential. In light of this fact we should remark that, if the projection π_{j-1}^{in} were to be replaced by a restriction operator, as is the case in classical multigrid, then (2.25) would no longer hold.

2.4.2. Matrix-form of the preconditioner. In order to describe the matrix-form of the preconditioner in MATLAB form we introduce the mass matrix $\mathbf{L}_j$ on $\mathcal{V}_j$, and we denote by $\mathbf{J}_j \in M_{N_j \times N_{j-1}}$ the matrix representing the interpolation operator $\mathcal{I}_j|_{\mathcal{V}_{j-1}} \in \mathfrak{L}(\mathcal{V}_{j-1}, \mathcal{V}_j)$. We assume that the inactive indices on level j are stored in the vector $\mathbf{i}_j$, and define the matrices

$$\mathbf{L}_j^I = \mathbf{L}_j(\mathbf{i}_j, \mathbf{i}_j), \quad \mathbf{J}_j^I = \mathbf{J}_j(\mathbf{i}_j, \mathbf{i}_{j-1}), \quad \mathbf{E}_j^I = \mathbf{I}(:, \mathbf{i}_j), \quad \mathbf{P}_j^I = (\mathbf{E}_j^I)^T,$$

where we used MATLAB syntax for the selection of submatrices. Note that $\mathbf{E}_j^I$ represents the extension operator $\mathcal{E}_j^{\text{in}}$ and $\mathbf{P}_j^I$ the operator P_j^{in}. We can now write the projector-operator π_{j-1}^{in} in matrix-form as

$$\mathbf{\Pi}_{j-1}^I = \left(\mathbf{L}_{j-1}^I\right)^{-1} \cdot (\mathbf{J}_j^I)^T \cdot \mathbf{L}_j^I \ ,$$

and ρ_{j-1}^{in} is represented by $(\mathbf{I} - \mathbf{J}_j^I \mathbf{\Pi}_{j-1}^I)$. So $\mathcal{M}_j^{\text{in}}$ is represented by the matrix

$$\mathbf{M}_j^I = \mathbf{J}_j^I \overbrace{\mathbf{P}_{j-1}^I(\mathbf{K}_{j-1}^T\mathbf{K}_{j-1} + \beta\mathbf{I})\mathbf{E}_{j-1}^I}^{\text{represents } \mathcal{G}_{j-1}^{\text{in}}} \mathbf{\Pi}_{j-1}^I + \beta(\mathbf{I} - \mathbf{J}_j^I\mathbf{\Pi}_{j-1}^I) \ ,$$

and $\mathcal{S}_j^{\text{in}}$ is represented by

$$\mathbf{S}_j^I = \mathbf{J}_j^I \left(\mathbf{P}_{j-1}^I(\mathbf{K}_{j-1}^T\mathbf{K}_{j-1} + \beta\mathbf{I})\mathbf{E}_{j-1}^I\right)^{-1} \mathbf{\Pi}_{j-1}^I + \beta^{-1}(\mathbf{I} - \mathbf{J}_j^I\mathbf{\Pi}_{j-1}^I) \ . \tag{2.26}$$

We should point out that, due to the presence of $\mathbf{\Pi}_{j-1}^I$, the matrices $\mathbf{M}_j^I$ and $\mathbf{S}_j^I$ are slightly nonsymmetric, hence one has to employ solvers for nonsymmetric systems in connection with the $\mathbf{M}_j^I$ preconditioner. We found that conjugate gradient squared (CGS) is quite efficient (see Section 5).

A symmetric alterative to $\mathbf{S}_j^I$ is obtained by replacing the orthogonal projection $\mathbf{\Pi}_{j-1}^I$ in (2.26) with the matrix $\boldsymbol{R}_{j-1}^I = \boldsymbol{R}_{j-1}(\mathbf{i}_{j-1}, \mathbf{i}_j)$, where the matrix $\boldsymbol{R}_{j-1}$ represents the restriction operator R_{j-1} defined in (2.21). This gives rise to the preconditioner

$$\widehat{\mathbf{S}}_j^I = \mathbf{J}_j^I \left(\mathbf{P}_{j-1}^I(\mathbf{K}_{j-1}^T\mathbf{K}_{j-1} + \beta\mathbf{I})\mathbf{E}_{j-1}^I\right)^{-1} \boldsymbol{R}_{j-1}^I + \beta^{-1}(\mathbf{I} - \mathbf{J}_j^I\boldsymbol{R}_{j-1}^I) \ . \tag{2.27}$$

It can be easily verified that $\widehat{\mathbf{S}}_j^I$ is symmetric, hence it can be used in practice as a preconditioner to conjugate gradient (CG). We conduct some of our numerical experiments in Section 5 using multigrid preconditioners defived from $\widehat{\mathbf{S}}_j^I$ rather than from $\mathbf{S}_j^I$. Note that for a uniform grid in d dimensions we have $\boldsymbol{R}_{j-1} = 2^{-d}\boldsymbol{J}_j^T$.

2.4.3. Spectral distance estimation. To quantify the quality of the two-grid preconditioner $\mathcal{M}_j^{\text{in}}$ we will estimate the *spectral distance* d_σ between $(\mathcal{G}_j^{\text{in}})^{-1}$ and its inverse $\mathcal{S}_j^{\text{in}}$ given by (2.25). The use of the spectral distance ensures that such estimates extend automatically to multigrid preconditioners, as discussed in Section 2.5 and Section 4 (see also [7, 9]). We briefly recall the definition of the spectral distance, as introduced in [7]. Given a Hilbert space $(\mathcal{X}, \langle\cdot,\cdot\rangle)$ we denote by $\mathfrak{L}_+(\mathcal{X})$ the set of operators with positive definite symmetric part:

$$\mathfrak{L}_+(\mathcal{X}) = \{T \in \mathfrak{L}(\mathcal{X}) : \langle Tu, u\rangle > 0, \ \ \forall u \in \mathcal{X} \setminus \{0\}\} \ .$$

Let the joined numerical range of $S, T \in \mathfrak{L}_+(\mathcal{X})$ be given by

$$W(S,T) = \left\{ \frac{\langle S_{\mathbb{C}} w, w\rangle}{\langle T_{\mathbb{C}} w, w\rangle} \ : \ w \in \mathcal{X}^{\mathbb{C}} \setminus \{0\} \right\} \ ,$$

where $T_{\mathbb{C}}(u + \mathbf{i}v) = T(u) + \mathbf{i}T(v)$ is the complexification of T. The spectral distance between $S, T \in \mathfrak{L}_+(\mathcal{X})$, is a measure of spectral equivalence between S and T, and it is defined by

$$d_\sigma(S,T) = \sup\{|\ln z| \ : \ z \in W(S,T)\} \ ,$$

where ln is the branch of the logarithm corresponding to $\mathbb{C} \setminus (-\infty, 0]$. Following Lemma 3.2 in [7], if $W(S,T) \subseteq \mathcal{B}_\alpha(1) = \{z \in \mathbb{C} \ : \ |z-1| < \alpha\}$ with $\alpha \in (0,1)$, then

$$d_\sigma(S,T) \leq \frac{|\ln(1-\alpha)|}{\alpha} \sup\{|z-1| \ : \ z \in W(S,T)\} \ , \tag{2.28}$$

which offers a practical way to estimate the spectral distance when it is small. The spectral distance serves both as a means to quantify the quality of a preconditioner and also as a convenient analysis tool for multigrid algorithms. Essentially, if two operators S, T satisfy

$$1 - \delta \leq \left| \frac{\langle S_\mathbb{C} w, w\rangle}{\langle T_\mathbb{C} w, w\rangle} \right| \leq 1 + \delta \ , \quad \forall w \in \mathcal{X}^\mathbb{C} \setminus \{0\},$$

with $\delta \ll 1$, then $d_\sigma(S,T) \approx \delta$. If $N \approx G^{-1}$ is a preconditioner for G, then both $d_\sigma(N, G^{-1})$ and $d_\sigma(N^{-1}, G)$ (quantities which are equal if G, N are symmetric) are shown to control the spectral radius $\rho(I - NG)$ (see [8]) which is an accepted quality-measure for a preconditioner. The advantage of using d_σ over $\rho(I-NG)$ is that the former is a true distance function.

The main result of this article is the following theorem.

Theorem 2.2. *If the operators $\mathcal{K}$ and $\mathcal{K}_j$ satisfy Condition* 2.1 *and the weights $w_i^{(j)}$ are uniform, then there exists $\delta > 0$ and a constant $C(\mathcal{K})$ (see Condition* 2.1 *) independent of j and of the inactive set so that*

$$d_\sigma \left(\left(\mathcal{G}_j^{\mathrm{in}}\right)^{-1}, \mathcal{S}_j^{\mathrm{in}} \right) \leq C\beta^{-1} \left(h_j^2 + \sqrt{\mu_j^{\mathrm{in}}} \right) \ , \tag{2.29}$$

where μ_j^{in} is the Lebesgue measure of $\partial_\mathrm{n}\Omega_j^{\mathrm{in}}$, provided that

$$\beta^{-1} \left(h_j^2 + \sqrt{\mu_j^{\mathrm{in}}} \right) < \delta \ .$$

We postpone the proof of Theorem 2.2 until Section 3.

Remark 2.3. *The natural question arises as to how to estimate μ_j^{in}. In the worst case scenario there are no coarse inactive nodes, so $\Omega_{j-1}^{\mathrm{in}} = \emptyset$; therefore $\partial_\mathrm{n}\Omega_j^{\mathrm{in}} = \Omega$, case in which the two-grid preconditioner is βI, so essentially there is no preconditioner. However, if u is the solution of* (1.1) *and the continuous inactive set defined by $\Omega^{\mathrm{in}} = \{x \in \Omega \ : \ u(x) > 0\}$ is a domain with Lipschitz boundary, then the discrete inactive set Ω_j^{in} is expected to be close to Ω^{in} provided that a good initial guess at the inactive set is available. In this case we expect that $\partial_\mathrm{n}\Omega_j^{\mathrm{in}}$ will lie within Ch_j of the topological boundary of Ω^{in}, therefore*

$$\mu_j^{\mathrm{in}} \approx Ch_j \ ,$$

where C is proportional to the $(d-1)$-dimensional measure of $\partial\Omega^{\mathrm{in}}$. Hence, the estimate (2.29) *truly implies*

$$d_\sigma \left(\left(\mathcal{G}_j^{\mathrm{in}}\right)^{-1}, \mathcal{S}_j^{\mathrm{in}} \right) \leq C \frac{\sqrt{h_j}}{\beta} \ , \tag{2.30}$$

which is consistent with the numerical experiments in Section 5. Under certain, reasonable assumptions we will show that the estimate (2.30) *extends to multilevel preconditioners.*

In case the grid is quasi-uniform but not uniform we apply Theorem 2.2 to the matrix $\widetilde{\mathbf{K}}_j \stackrel{\text{def}}{=} \mathbf{W}_j\mathbf{K}_j$. The important aspect in the estimate is the verification of Condition 2.1 by $\widetilde{\mathbf{K}}_j$. Following [9] we introduce the following indirect measure of grid-smoothness: for each j we consider a C^2-function $w_j : \overline{\Omega} \to \mathbb{R}$ so that

$$w_j(P_i^{(j)}) = w_i^{(j)}, \quad \forall i = 1, 2, \ldots, N_j, \tag{2.31}$$

with $w_i^{(j)}$ given by (2.1). With this notation the matrix $\widetilde{\mathbf{K}}_j$ represents the operator $\widetilde{\mathcal{K}}_j \in \mathfrak{L}(\mathcal{V}_j)$ defined by

$$\widetilde{\mathcal{K}}_j u \stackrel{\text{def}}{=} \mathcal{I}_j(w_j \cdot (\mathcal{K}_j u)) \ . \tag{2.32}$$

Note that, because the grids are hierarchical ($\mathcal{T}_j$ is obtained from $\mathcal{T}_{j-1}$ by adding nodes), the function w_j can serve for defining all operators $\widetilde{\mathcal{K}}_l$, for $l = 0, 1, \ldots, j$. Consider j fixed, and define $\widetilde{\mathcal{K}} \in \mathfrak{L}(L^2(\Omega))$ by

$$\widetilde{\mathcal{K}}u \stackrel{\text{def}}{=} w_j \cdot (\mathcal{K}_j u) \ .$$

By Proposition 4.8 in [9], the operators $(\widetilde{\mathcal{K}}_l)_{l=0,1,\ldots,j}$ together with the continuous operator $\widetilde{\mathcal{K}}$ satisfy Condition 2.1 with

$$C(\widetilde{\mathcal{K}}) = \|w_j\|_{W_2^\infty(\Omega)} C(\mathcal{K}) \ .$$

Thus we establish the following corollary.

COROLLARY 2.4. *If the operators $\mathcal{K}$ and $\mathcal{K}_j$ satisfy Condition* 2.1 *and $w_j \in C^2(\overline{\Omega})$ satisfies* (2.31)*, then there exists $\delta > 0$ and a constant $C(\mathcal{K})$ independent of j and the inactive set so that*

$$d_\sigma\left(\left(\mathcal{G}_j^{\text{in}}\right)^{-1}, \mathcal{S}_j^{\text{in}}\right) \le C\beta^{-1}\|w_j\|_{W_2^\infty(\Omega)}\left(h_j + \sqrt{\mu_j^{\text{in}}}\right) \ , \tag{2.33}$$

where μ_j^{in} is as in Theorem 2.2, provided that

$$\beta^{-1}\|w_j\|_{W_2^\infty(\Omega)}\left(h_j + \sqrt{\mu_j^{\text{in}}}\right) < \delta \ .$$

Corollary 2.4 marks a decrease of the power of h_j from two to one compared to the estimate in Theorem 2.2. This fact is due to the nonuniformity of the grid as can be seen in the analysis of Section 3. However, in general, the larger of the two terms in (2.33) is expected to be $\sqrt{\mu_j^{\text{in}}}$, so the main effect of the nonuniformity on the estimate is the presence of the factor $\|w_j\|_{W_2^\infty(\Omega)}$.

2.5. The multigrid preconditioner. We now assume the level $j_{\max} \ge 0$ to be fixed; we will refer to $j_{\max}$ as the finest level. The goal is to construct a multigrid operator $\mathcal{Z}_{j_{\max}}^{\text{in}} \approx (\mathcal{G}_{j_{\max}}^{\text{in}})^{-1}$ which satisfies the following conditions: (i) if $j_{\max} = 1$ then $\mathcal{Z}_1^{\text{in}} = \mathcal{S}_1^{\text{in}}$; (ii) an estimate like (2.29) holds if we replace $\mathcal{S}_{j_{\max}}^{\text{in}}$ with $\mathcal{Z}_{j_{\max}}^{\text{in}}$. In order to construct $\mathcal{Z}_{j_{\max}}^{\text{in}}$ we must first specify the coarser inactive domains Ω_j^{in}, inactive index-sets $\mathcal{I}^{(j)}$, and inactive spaces $\mathcal{V}_j^{\text{in}}$ for $j = j_{\max} - 1, \ldots, 0$. All these entities are defined recursively using (2.18), (2.19), (2.20), and are essentially specified by the sets $\mathcal{I}^{(j)}$. Hence, we give below the algorithm for computing the inactive-index sets $\mathcal{I}^{(j)}$ for $j = j_{\max} - 1, \ldots, 0$ (recall that $\mathcal{I}^{(j)}$ is one of the set-iterates in the SSNM solution process). Given a vertex $P_{i_c}^{(j-1)}$ of the triangulation $\mathcal{T}_{j-1}$, let $P_{i_f}^{(j)}$ be its fine label. We define the 'fine neighborhood' of $P_{i_c}^{(j-1)}$ by

$$\mathcal{N}_j(P_{i_c}^{(j-1)}) = \{R^{(j)} \ : \ R^{(j)} \text{ neighbor in } \mathcal{T}_j \text{ of } P_{i_f}^{(j)}\} \cup \{P_{i_f}^{(j)}\} \ .$$

In principle, the algorithm for finding the coarse inactive sets (see Algorithm 2.5 below) can be easily implemented using MATLAB's set operations. However, we should point out that for medium- and large-scale problems Algorithm 2.5 should be implemented using a divide-and-conquer strategy in order to avoid repeated memory allocation at step 5.

ALGORITHM 2.5 (Inactive set definition).

```
1. for $j = j_{\max} \ : \ -1 \ : \ 1$
2.     $\mathcal{I}^{(j-1)} = \emptyset$
3.     for $i = 1 \ : \ N_{j-1}$
4.         if $\mathcal{N}_j(P_i^{(j-1)}) \subseteq \mathcal{I}^{(j)}$
5.             $\mathcal{I}^{(j-1)} = \mathcal{I}^{(j-1)} \cup \{P_i^{(j-1)}\}$
```

If we define the operator

$$\mathfrak{I}_{j-1}^j : \mathfrak{L}(\mathcal{V}_{j-1}^{\text{in}}) \to \mathfrak{L}(\mathcal{V}_j^{\text{in}}), \ \ \mathfrak{I}_{j-1}^j(\mathcal{X}) = \mathcal{X} \cdot \pi_j^{\text{in}} + \beta^{-1}(I - \pi_j^{\text{in}}) \ ,$$

then $\mathcal{S}_j^{\text{in}}$ can be written as

$$\mathcal{S}_j^{\text{in}} = \mathfrak{I}_{j-1}^j \left((\mathcal{G}_{j-1}^{\text{in}})^{-1}\right) \ . \tag{2.34}$$

In light of (2.34) and the continuity of the affine operator $\mathfrak{I}_{j-1}^j$, it is tempting to define recursively the following multigrid preconditioner:

$$\widetilde{\mathcal{Z}}_j^{\text{in}} = \begin{cases} (\mathcal{G}_j^{\text{in}})^{-1} & , \ \text{if } j = 0 \ , \\ \mathfrak{I}_{j-1}^j(\widetilde{\mathcal{Z}}_{j-1}^{\text{in}}) & , \ \text{if } j \geq 1 \ . \end{cases} \tag{2.35}$$

As shown in [7], the V-cycle type preconditioner $\widetilde{\mathcal{Z}}_j^{\text{in}}$ does not satisfy condition (ii) above. In fact one can show that, under the conditions set in Remark 2.3, $\widetilde{\mathcal{Z}}_j^{\text{in}}$ satisfies (2.30) with h_j replaced by h_0. As a result, the number of preconditioned iterations would no longer be decreasing with $h_j \downarrow 0$, as is the case for the two-grid preconditioner, but would be bounded. Instead of the definition (2.35), we employ the same strategy adopted in [7, 9], which guarantees that the estimate for the multigrid preconditioner will be similar to that for the two-grid preconditioner (except for a constant factor). In order to do so, we define the operator $\mathfrak{N}_j$ by

$$\mathfrak{N}_j : \mathfrak{L}(\mathcal{V}_j^{\text{in}}) \to \mathfrak{L}(\mathcal{V}_j^{\text{in}}), \ \ \mathfrak{N}_j(\mathcal{X}) \overset{\text{def}}{=} 2\mathcal{X} - \mathcal{X} \cdot \mathcal{G}_j^{\text{in}} \cdot \mathcal{X} \ .$$

Note that $\mathfrak{N}_j$ is the Newton iterator for the operator-equation

$$\mathcal{X}^{-1} - \mathcal{G}_j^{\text{in}} = 0 \ ,$$

as shown by Drăgănescu and Dupont in [7]. This implies that, if $\mathcal{X}_0$ is a good approximation of $(\mathcal{G}_j^{\text{in}})^{-1}$, then $\mathcal{X}_1 = \mathfrak{N}_j(\mathcal{X}_0)$ is the first Newton iterate of the above operator-equation starting with $\mathcal{X}_0$, and so $\mathcal{X}_1$ is significantly closer to $(\mathcal{G}_j^{\text{in}})^{-1}$ than $\mathcal{X}_0$. This idea was also used in [9] to construct multigrid preconditioners of the same quality as the two-grid preconditioners. The Algorithm 2.6 essentially has a W-cycle structure; we construct recursively a sequence of coarser-level operators $\mathcal{Z}_j$ for $0 \leq j \leq j_{\max} - 1$ so that each application $\mathcal{Z}_j$ requires one application of $\mathcal{G}_j^{\text{in}}$. At the finest level no application of $\mathcal{G}_{j_{\max}}^{\text{in}}$ is performed inside the preconditioner $\mathcal{Z}_{j_{\max}}^{\text{in}}$.

Algorithm 2.6 (Operator-form definition of $\mathcal{Z}_{j_{\max}}^{\text{in}}$).

```
if $j = 0$
        $\mathcal{Z}_0^{\text{in}} := (\mathcal{G}_0^{\text{in}})^{-1}$                          % coarsest level
else if $j < j_{\max}$
            $\mathcal{Z}_j^{\text{in}} := \mathfrak{N}_j(\mathfrak{I}_{j-1}^{j}(\mathcal{Z}_{j-1}^{\text{in}}))$           % intermediate level
        else                                      % here $j = j_{\max}$
            $\mathcal{Z}_j^{\text{in}} := \mathfrak{I}_{j-1}^{j}(\mathcal{Z}_{j-1}^{\text{in}})$                 % finest level
```

We give a detailed analysis of the multigrid preconditioner in Section 4. To anticipate, essentially we show that, under conditions set forth in Theorem 4.5, the operator $\mathcal{Z}_{j_{\max}}^{\text{in}}$ satisfies an estimate like (2.29).

We conclude our description of the multigrid preconditioner by commenting on its computational cost. First, it goes without saying that for large-scale problems the operator $\mathcal{Z}_{j_{\max}}^{\text{in}}$ is not formed; rather, its action on a vector is implemented as a multilevel, matrix-free iteration. This matrix-free implementation, described in [9], naturally calls for a matrix-free implementation of $\mathcal{G}_j^{\text{in}}$ which essentially relies on applying $\mathcal{G}_j$ matrix-free, as follows from (2.23). The computational cost of the application of the multigrid preconditioner is estimated in [9]; for completeness we briefly describe the result. Let $C(j)$ denote the cost of applying the operator $\mathcal{G}_j$ and assume that there exists $0 < \tau < 1/2$ so that $C(j-1) \leq \tau C(j)$. For example, if the cost of applying $\mathcal{K}_j$ is linear in the number of variables, as is the case if $\mathcal{K}_j$ is computed using classical multigrid, then for a problem with d spatial dimensions we expect $\tau = 2^{-d}$. We also assume that unpreconditioned CG solves the problem (2.17) in at most K_{cg} iterations to acceptable precision, with K_{cg} independent of j. This assumption is certainly justified by the uniform-boundedness of the condition number of $\mathcal{G}_j$ (see also Theorem 3.1 in [22]); however, due to the compactness of the operators, K_{cg} is in fact relatively small (see Section 5.2 for numerical results). If $T(j)$ denotes the cost of applying $\mathcal{Z}_{j_{\max}}^{\text{in}}$, then cf. [9] we have

$$T(j_{\max}) \leq \tau C(j_{\max}) \left(K_{cg}(2\tau)^{j_{\max}-1} + \frac{1 - (2\tau)^{j_{\max}-1}}{1 - 2\tau} \right) . \tag{2.36}$$

For example, if four levels are used ($j_{\max} = 3$), $\tau = 2^{-3}$, and with $K_{cg} = 20$, then $T(j_{\max}) \leq 0.3125 \cdot C(j_{\max})$; if 5 levels are used, then $T(j_{\max}) \leq 0.2031 \cdot C(j_{\max})$. The estimate (2.36) shows that for large-scale, multidimensional problems where τ is expected to be relatively small, the cost of applying the preconditioner is a fraction of the cost of computing a matrix-vector multiplication at the finest level if multiple levels are used. For $j_{\max} \gg 1$, $T(j_{\max}) \leq (\tau/(1-\tau))\, C(j_{\max})$.

3. Analysis of the two-grid preconditioner.

The main step in the analysis is to evaluate the norm-distance between the operators $\mathcal{G}_j^{\text{in}}$ and $\mathcal{M}_j^{\text{in}}$ which is done in Proposition 3.4. The plan of the analysis generally resembles that of the analysis of multigrid preconditioners for IPMs from [9], however, certain critical estimates related to the projection π_j^{in} are different for the case of SSNMs.

First we restate Lemma 4.3 in [9] as

Lemma 3.1. *With $(w_i^{(j)})_{1\le i\le N_j}$ chosen as in (2.1) there exists a constant C dependent on $\mathcal{T}_0$ and independent of j so that*

$$|\langle u,v\rangle_j - \langle u,v\rangle| \le Ch_j^2\|u\|_{H^1(\Omega)}\cdot\|v\|_{H^1(\Omega)}\ ,\ \forall u,v\in\mathcal{V}_j. \tag{3.1}$$

We also recall Lemma 4.4 in [9]:

Lemma 3.2. *If $\mathcal{K},\mathcal{K}_j$ satisfy Condition 2.1, there exist constants $C(\mathcal{K})$ and $C'=C'(\Omega)$ independent of j such that the following hold:*
(a) *H^1, L^2 - uniform stability of $\mathcal{K}_j$:*

$$\|\mathcal{K}_j u\|_{H^m(\Omega)} \le C(\mathcal{K})\|u\|\ ,\ \forall u\in\mathcal{V}_j,\ m=0,1,\quad j=0,1,\ldots; \tag{3.2}$$

(b) *smoothing of negative-index norm:*

$$\|\mathcal{K}u\| \le C(\mathcal{K})\,\|u\|_{\widetilde{H}^{-m}}\ \ ,\ \forall u\in\mathcal{V}_j, m=1,2\ ; \tag{3.3}$$

(c) *negative-index norm approximation of the identity by $\pi_{j-1},\mathcal{R}_{j-1}$:*

$$\|(I-\pi_{j-1})u\|_{\widetilde{H}^{-2}(\Omega)} \le C'h_j^2\,\|u\|\,,\quad \forall u\in\mathcal{V}_j; \tag{3.4}$$

$$\|(I-\mathcal{R}_{j-1})u\|_{\widetilde{H}^{-p}(\Omega)} \le C'h_j^p\,\|u\|\,,\quad \forall u\in\mathcal{V}_j\ , \tag{3.5}$$

where $p=1$ on a quasi-uniform grid, and $p=2$ on a locally symmetric grid;
(d) *$\mathcal{K}$ diminishes high-frequencies:*

$$\|\mathcal{K}(I-\pi_{j-1})u\| \le C(\mathcal{K})h_j^2\,\|u\|\,,\quad \forall u\in\mathcal{V}_j\ ; \tag{3.6}$$

$$\|\mathcal{K}(I-\mathcal{R}_{j-1})u\| \le C(\mathcal{K})h_j^p\,\|u\|\,,\quad \forall u\in\mathcal{V}_j\ , \tag{3.7}$$

where $p=1$ on an unstructured grid, and $p=2$ on a locally symmetric grid;
(e)

$$|\langle\mathcal{K}u,\mathcal{K}v\rangle - \langle\mathcal{K}_j u,\mathcal{K}_j v\rangle| \le C(\mathcal{K})h_j^2\|u\|\cdot\|v\|\ ,\ \forall u,v\in\mathcal{V}_j\ . \tag{3.8}$$

The main difference between the two-grid preconditioners for the unconstrained case vs. the constrained case is in the properties of the projectors on the coarse spaces. While the projector π_{j-1} on the entire coarse space $\mathcal{V}_{j-1}$ satisfies (3.4), the projector on the inactive space $\mathcal{V}_{j-1}^{\text{in}}$ satisfies the weaker estimate below.

Lemma 3.3. *If $d\le 3$ there exists a constant C depending on the domain Ω and the base triangulation $\mathcal{T}_0$ so that*

$$\|(I-\pi_{j-1}^{\text{in}})u\|_{\widetilde{H}^{-2}(\Omega)} \le C\left(h_j^2+\sqrt{\mu_j^{\text{in}}}\right)\|u\|,\quad \text{for all } u\in\mathcal{V}_j^{\text{in}}\ , \tag{3.9}$$

where μ_j^{in} is the Lebesgue measure of $\partial_{\text{n}}\Omega_j^{\text{in}}$. The constant C is independent of j and of the inactive set Ω_j^{in}.

Proof. Let $u\in\mathcal{V}_j^{\text{in}}$, $v\in H^2(\Omega)\cap H_0^1(\Omega)$ be arbitrary, and let $v_{j-1}\in\mathcal{V}_{j-1}^{\text{in}}$ be the natural interpolant of v in $\mathcal{V}_{j-1}^{\text{in}}$, that is

$$v_{j-1} = \sum_{i\in\mathcal{I}^{(j-1)}} v(P_i^{(j-1)})\varphi_i^{(j-1)}\ .$$

Let $T \in \mathcal{T}_{j-1}$ be a coarse element lying in Ω_j^{in}. If $T \subseteq \text{Int}_{\text{n}}\Omega_j^{\text{in}}$ then v_{j-1} agrees on T with the interpolant of v in $\mathcal{V}_{j-1}$. Therefore a standard interpolation estimate (see [6]) applied on $\text{Int}_n\Omega_j^{\text{in}}$ gives

$$\|v - v_{j-1}\|_{L^2(\text{Int}_n\Omega_j^{\text{in}})} \leq C h_{j-1}^2 |v|_{H^2(\text{Int}_n\Omega_j^{\text{in}})} \leq C\underline{f}^{-1} h_j^2 |v|_{H^2(\Omega)} \ .$$

On a coarse element T that satisfies $T \subseteq \partial_{\text{n}}\Omega_j^{\text{in}}$, if v_{j-1} is not identically zero on T then v_{j-1} and v agree at least for one vertex of T and potentially disagree at a vertex corresponding to an active constraint. In either case the bound $\|v_{j-1}\|_{L^\infty(T)} \leq \|v\|_{L^\infty(T)}$ holds. Hence,

$$\|v - v_{j-1}\|_{L^2(\partial_{\text{n}}\Omega_j^{\text{in}})} \leq 2\sqrt{\mu_j^{\text{in}}}\ \|v\|_{L^\infty(\partial_{\text{n}}\Omega_j^{\text{in}})} \leq C\sqrt{\mu_j^{\text{in}}}\ \|v\|_{H^2(\Omega)} \ ,$$

by Sobolev's inequality. We have

$$\begin{aligned}
&\left|\left\langle (I - \pi_{j-1}^{\text{in}})u, v\right\rangle\right| \\
&= \left|\left\langle (I - \pi_{j-1}^{\text{in}})u, v - v_{j-1}\right\rangle\right| = \left|\int_{\Omega_j^{\text{in}}} (u - \pi_{j-1}^{\text{in}}u)\ (v - v_{j-1})\right| && (3.10)\\
&\leq \left|\int_{\text{Int}_{\text{n}}\Omega_j^{\text{in}}} (u - \pi_{j-1}^{\text{in}}u)\ (v - v_{j-1})\right| + \left|\int_{\partial_{\text{n}}\Omega_j^{\text{in}}} (u - \pi_{j-1}^{\text{in}}u)\ (v - v_{j-1})\right| && (3.11)\\
&\leq C\|u - \pi_{j-1}^{\text{in}}u\|_{L^2(\Omega_j^{\text{in}})} \left(h_j^2 + \sqrt{\mu_j^{\text{in}}}\right) \|v\|_{H^2(\Omega)} \ . && (3.12)
\end{aligned}$$

Since $\|\pi_{j-1}^{\text{in}}u\| \leq \|u\|$, the result now follows after dividing by $\|v\|_{H^2(\Omega)}$ and taking the supremum over all $v \in H^2(\Omega)$. □

Proposition 3.4. *If the operators $\mathcal{K}$ and $\mathcal{K}_j$ satisfy Condition* 2.1 *and the weights $w_i^{(j)}$ are uniform, then there exists a constant C independent on j and the inactive set so that*

$$\|\mathcal{G}_j^{\text{in}} - \mathcal{M}_j^{\text{in}}\| \leq C\left(h_j^2 + \sqrt{\mu_j^{\text{in}}}\right) \ , \qquad (3.13)$$

where μ_j^{in} is the Lebesgue measure of $\partial_{\text{n}}\Omega_j^{\text{in}}$.

Proof. Since $P_j^{\text{in}}\mathcal{E}_j^{\text{in}} = \pi_{j-1}^{\text{in}} + \rho_{j-1}^{\text{in}} = I_{\mathcal{V}_j^{\text{in}}}$, we have

$$\mathcal{G}_j^{\text{in}} - \mathcal{M}_j^{\text{in}} = P_j^{\text{in}}\mathcal{K}_j^*\mathcal{K}_j\mathcal{E}_j^{\text{in}} - P_{j-1}^{\text{in}}\mathcal{K}_{j-1}^*\mathcal{K}_{j-1}\mathcal{E}_{j-1}^{\text{in}}\pi_{j-1}^{\text{in}} \ . \qquad (3.14)$$

The argument has a structure that is similar to the proof of Proposition 4.5 in [9] with changes due to the specific approximation properties of P_{j-1}^{in} and π_{j-1}^{in}. Let $u, v \in \mathcal{V}_j^{\text{in}}$ be arbitrary, and define $\widetilde{u} = \mathcal{E}_{j-1}^{\text{in}}\pi_{j-1}^{\text{in}}u$. Note that $\|\widetilde{u}\| \leq \|u\|$. We first examine the difference

$$\begin{aligned}
&|\left\langle \mathcal{K}_{j-1}^*\mathcal{K}_{j-1}\widetilde{u}, v\right\rangle_j - \left\langle P_{j-1}^{\text{in}}\mathcal{K}_{j-1}^*\mathcal{K}_{j-1}\widetilde{u}, v\right\rangle_j| \\
&= \left|\int_{\Omega_j^{\text{in}}} (I - P_{j-1}^{\text{in}})(\mathcal{K}_{j-1}^*\mathcal{K}_{j-1}\widetilde{u}) \cdot v\right| = \left|\int_{\partial_{\text{n}}\Omega_j^{\text{in}}} (I - P_{j-1}^{\text{in}})(\mathcal{K}_{j-1}^*\mathcal{K}_{j-1}\widetilde{u}) \cdot v\right| \\
&\leq \|(I - P_{j-1}^{\text{in}})(\mathcal{K}_{j-1}^*\mathcal{K}_{j-1}\widetilde{u})\|_{L^2(\partial_{\text{n}}\Omega_j^{\text{in}})} \cdot \|v\|_{L^2(\partial_{\text{n}}\Omega_j^{\text{in}})} \\
&\leq \sqrt{\mu_j^{\text{in}}}\|(I - P_{j-1}^{\text{in}})(\mathcal{K}_{j-1}^*\mathcal{K}_{j-1}\widetilde{u})\|_{L^\infty(\partial_{\text{n}}\Omega_j^{\text{in}})} \cdot \|v\|_{L^2(\partial_{\text{n}}\Omega_j^{\text{in}})} \ ,
\end{aligned}$$

where we used that $(I-P^{\rm in}_{j-1})(\mathcal{K}^*_{j-1}\mathcal{K}_{j-1}\widetilde{u})$ is zero in $\mathrm{Int}_n\Omega^{\rm in}_j$ and that v is zero outside of $\Omega^{\rm in}_j$. Since for $w\in\mathcal{V}_{j-1}$ and $x\in\Omega$ the function value $(I-P^{\rm in}_{j-1})w(x)$ lies between 0 and $w(x)$ we have

$$\|(I-P^{\rm in}_{j-1})(\mathcal{K}^*_{j-1}\mathcal{K}_{j-1}\widetilde{u})\|_{L^\infty(\partial_{\rm n}\Omega^{\rm in}_j)}\leq\|\mathcal{K}^*_{j-1}\mathcal{K}_{j-1}\widetilde{u}\|_{L^\infty(\Omega_j)}\overset{(2.5)}{\leq}C\|\widetilde{u}\|\leq C\|u\|\ .$$

Therefore

$$|\left\langle\mathcal{K}^*_{j-1}\mathcal{K}_{j-1}\widetilde{u},v\right\rangle_j-\left\langle P^{\rm in}_{j-1}\mathcal{K}^*_{j-1}\mathcal{K}_{j-1}\widetilde{u},v\right\rangle_j|\leq C\sqrt{\mu^{\rm in}_j}\,\|u\|\cdot\|v\|\ . \tag{3.15}$$

We return to estimating $(\mathcal{G}^{\rm in}_j-\mathcal{M}^{\rm in}_j)$. We have

$$\begin{aligned}
&\left\langle(\mathcal{G}^{\rm in}_j-\mathcal{M}^{\rm in}_j)u,v\right\rangle_j\\
&\quad=\ \left\langle P^{\rm in}_j\mathcal{K}^*_j\mathcal{K}_j\mathcal{E}^{\rm in}_j u,v\right\rangle_j-\left\langle\mathcal{K}^*_{j-1}\mathcal{K}_{j-1}\mathcal{E}^{\rm in}_{j-1}\pi^{\rm in}_{j-1}u,v\right\rangle_j\\
&\qquad+\overbrace{\left\langle\mathcal{K}^*_{j-1}\mathcal{K}_{j-1}\mathcal{E}^{\rm in}_{j-1}\pi^{\rm in}_{j-1}u,v\right\rangle_j-\left\langle P^{\rm in}_{j-1}\mathcal{K}^*_{j-1}\mathcal{K}_{j-1}\mathcal{E}^{\rm in}_{j-1}\pi^{\rm in}_{j-1}u,v\right\rangle_j}^{A_0}\\
&\quad\overset{(2.21)}{=}\ \left\langle\mathcal{K}^*_j\mathcal{K}_j\mathcal{E}^{\rm in}_j u,v\right\rangle_j-\left\langle\mathcal{K}^*_{j-1}\mathcal{K}_{j-1}\mathcal{E}^{\rm in}_{j-1}\pi^{\rm in}_{j-1}u,R_{j-1}v\right\rangle_{j-1}+A_0\\
&\quad=\ \left\langle\mathcal{K}_j\mathcal{E}^{\rm in}_j u,\mathcal{K}_j\mathcal{E}^{\rm in}_j v\right\rangle_j-\left\langle\mathcal{K}_{j-1}\mathcal{E}^{\rm in}_{j-1}\pi^{\rm in}_{j-1}u,\mathcal{K}_{j-1}\mathcal{E}^{\rm in}_{j-1}R_{j-1}v\right\rangle_{j-1}+A_0\\
&\quad=\ A_0+\overbrace{\left\langle\mathcal{K}_j\mathcal{E}^{\rm in}_j u,\mathcal{K}_j\mathcal{E}^{\rm in}_j v\right\rangle_j-\left\langle\mathcal{K}_j\mathcal{E}^{\rm in}_j u,\mathcal{K}_j\mathcal{E}^{\rm in}_j v\right\rangle}^{A_1}+\\
&\qquad\overbrace{\left\langle\mathcal{K}_j\mathcal{E}^{\rm in}_j u,\mathcal{K}_j\mathcal{E}^{\rm in}_j v\right\rangle-\left\langle\mathcal{K}\mathcal{E}^{\rm in}_j u,\mathcal{K}\mathcal{E}^{\rm in}_j v\right\rangle}^{A_2}+\\
&\qquad\overbrace{\left\langle\mathcal{K}\mathcal{E}^{\rm in}_j u,\mathcal{K}\mathcal{E}^{\rm in}_j v\right\rangle-\left\langle\mathcal{K}\mathcal{E}^{\rm in}_{j-1}\pi^{\rm in}_{j-1}u,\mathcal{K}\mathcal{E}^{\rm in}_{j-1}R_{j-1}v\right\rangle}^{A_3}+\\
&\qquad\overbrace{\left\langle\mathcal{K}\mathcal{E}^{\rm in}_{j-1}\pi^{\rm in}_{j-1}u,\mathcal{K}\mathcal{E}^{\rm in}_{j-1}R_{j-1}v\right\rangle-\left\langle\mathcal{K}_{j-1}\mathcal{E}^{\rm in}_{j-1}\pi^{\rm in}_{j-1}u,\mathcal{K}_{j-1}\mathcal{E}^{\rm in}_{j-1}R_{j-1}v\right\rangle}^{A_4}+\\
&\qquad\overbrace{\left\langle\mathcal{K}_{j-1}\mathcal{E}^{\rm in}_{j-1}\pi^{\rm in}_{j-1}u,\mathcal{K}_{j-1}\mathcal{E}^{\rm in}_{j-1}R_{j-1}v\right\rangle-\left\langle\mathcal{K}_{j-1}\mathcal{E}^{\rm in}_{j-1}\pi^{\rm in}_{j-1}u,\mathcal{K}_{j-1}\mathcal{E}^{\rm in}_{j-1}R_{j-1}v\right\rangle_{j-1}}^{A_5}\ .
\end{aligned}$$

By (3.15)

$$|A_0|\leq C\sqrt{\mu^{\rm in}_j}\,\|u\|\cdot\|v\|\ .$$

For A_1 and A_5 we use (3.1), (3.2), and (2.22) to conclude that

$$\begin{aligned}
|A_1|&\leq Ch_j^2\|\mathcal{K}_j\mathcal{E}^{\rm in}_j u\|_{H^1(\Omega)}\cdot\|\mathcal{K}_j\mathcal{E}^{\rm in}_j v\|_{H^1(\Omega)}\leq Ch_j^2\|u\|\cdot\|v\|\ ,\\
|A_5|&\leq Ch_{j-1}^2\|\mathcal{K}_{j-1}\mathcal{E}^{\rm in}_{j-1}\pi^{\rm in}_{j-1}u\|_{H^1(\Omega)}\cdot\|\mathcal{K}_j\mathcal{E}^{\rm in}_j R_{j-1}v\|_{H^1(\Omega)}\\
&\leq C\underline{f}^2h_j^2\|\pi^{\rm in}_{j-1}u\|\cdot\|R_{j-1}v\|\leq Ch_j^2\|u\|\cdot\|v\|\ .
\end{aligned}$$

Estimation of A_2, A_4, essentially involving (3.8), is handled exactly as in [9] to give

$$\max(|A_2|,|A_4|)\leq Ch_j^2\|u\|\cdot\|v\|\ .$$

Finally the term A_3 is estimated by

$$\begin{aligned}|A_3| &\le |\langle \mathcal{K}\mathcal{E}_j^{\text{in}}(I-\pi_{j-1}^{\text{in}})u, \mathcal{K}\mathcal{E}_j^{\text{in}}v\rangle| + |\langle \mathcal{K}\mathcal{E}_{j-1}^{\text{in}}\pi_{j-1}^{\text{in}}u, (\mathcal{K}\mathcal{E}_j^{\text{in}}(I-R_{j-1})v\rangle| \\ &\le C\left(h_j^2+\sqrt{\mu_j^{\text{in}}}\right)\|u\|\cdot\|v\| \ .\end{aligned}$$

We conclude that

$$\left|\left\langle(\mathcal{G}_j^{\text{in}}-\mathcal{M}_j^{\text{in}})u,v\right\rangle_j\right| \le C\left(h_j^2+\sqrt{\mu_j^{\text{in}}}\right)\|u\|\cdot\|v\|,\ \forall u,v\in\mathcal{V}_j^{\text{in}} \ ,$$

hence $\|\!|\mathcal{G}_j^{\text{in}}-\mathcal{M}_j^{\text{in}}|\!\|_j \le \left(h_j^2+\sqrt{\mu_j^{\text{in}}}\right)$, and the result follows from the equivalence of the operator-norms $\|\!|\cdot|\!\|_j,\ \|\cdot\|$. □

Theorem 2.2 now follows easily from Proposition 3.4:

Proof. First note that for $u\in\mathcal{V}_j^{\text{in}}$

$$\left\langle\mathcal{G}_j^{\text{in}}u,u\right\rangle_j = \left\langle\mathcal{K}_j^{\text{in}}\mathcal{E}^{\text{in}}u,\mathcal{K}_j^{\text{in}}\mathcal{E}^{\text{in}}u\right\rangle_j+\beta\left\langle u,u\right\rangle_j \ge \beta\left\langle u,u\right\rangle_j \ , \tag{3.16}$$

so $\sigma(\mathcal{G}_j^{\text{in}})\subseteq[\beta,\infty)$, which implies, due to the symmetry of $\mathcal{G}_j^{\text{in}}$ with respect to $\langle\cdot,\cdot\rangle_j$ and the equivalence of the operator-norms $\|\!|\cdot|\!\|_j,\ \|\cdot\|$, that

$$\|\left(\mathcal{G}_j^{\text{in}}\right)^{-\frac{1}{2}}\|\le C\beta^{-\frac{1}{2}} \ .$$

By Proposition 3.4

$$\|I-\left(\mathcal{G}_j^{\text{in}}\right)^{-\frac{1}{2}}\mathcal{M}_j^{\text{in}}\left(\mathcal{G}_j^{\text{in}}\right)^{-\frac{1}{2}}\|\le C\|\left(\mathcal{G}_j^{\text{in}}\right)^{-\frac{1}{2}}\|^2\cdot\|\mathcal{G}_j^{\text{in}}-\mathcal{M}_j^{\text{in}}\|\le C\beta^{-1}\left(h_j^2+\sqrt{\mu_j^{\text{in}}}\right) \ .$$

The remainder of the argument proceeds as in the proof of Theorem 4.9 in [9]. □

4. Analysis of the multigrid preconditioner.

We begin with a few technical results concerning the spectral distance.

LEMMA 4.1. *If* $w_1,w_2\in\mathbb{H}=\{z\in\mathbb{C}\ :\ \text{Re}(z)>0\}$*, then*

$$\left|\ln\frac{w_1+x}{w_2+x}\right|\le\left|\ln\frac{w_1}{w_2}\right|,\quad\forall x\ge0\ , \tag{4.1}$$

where $\text{Re}(z)$ *denotes the real part of a complex number* z*.*

We postpone the proof of Lemma 4.1 until Appendix A. The next result shows that $\mathfrak{I}_{j-1}^j$ is non-expansive when measured in the spectral distance.

LEMMA 4.2. *For all* $j=1,2,...,$ *and* $\mathcal{G},\mathcal{H}\in\mathfrak{L}_+(\mathcal{V}_{j-1})$

$$d_\sigma(\mathfrak{I}_{j-1}^j(\mathcal{G}),\mathfrak{I}_{j-1}^j(\mathcal{H}))\le d_\sigma(\mathcal{G},\mathcal{H})\ , \tag{4.2}$$

where the spectral distance is computed using the L^2*-inner product.*

Proof. Let $u\in\mathcal{V}_j^{\mathbb{C}}$ be arbitrary so that $\pi_{j-1}^{\text{in}}u\ne0$. Then

$$\begin{aligned}&\left|\ln\frac{\left\langle\mathcal{G}\pi_{j-1}^{\text{in}}u+\beta^{-1}\rho_{j-1}^{\text{in}}u,u\right\rangle}{\left\langle\mathcal{H}\pi_{j-1}^{\text{in}}u+\beta^{-1}\rho_{j-1}^{\text{in}}u,u\right\rangle}\right| \\ &=\left|\ln\frac{\left\langle\mathcal{G}\pi_{j-1}^{\text{in}}u,\pi_{j-1}^{\text{in}}u\right\rangle+\beta^{-1}\left\langle\rho_{j-1}^{\text{in}}u,\rho_{j-1}^{\text{in}}u\right\rangle}{\left\langle\mathcal{H}\pi_{j-1}^{\text{in}}u,\pi_{j-1}^{\text{in}}u\right\rangle+\beta^{-1}\left\langle\rho_{j-1}^{\text{in}}u,\rho_{j-1}^{\text{in}}u\right\rangle}\right|\overset{(4.1)}{\le}\left|\ln\frac{\left\langle\mathcal{G}\pi_{j-1}^{\text{in}}u,\pi_{j-1}^{\text{in}}u\right\rangle}{\left\langle\mathcal{H}\pi_{j-1}^{\text{in}}u,\pi_{j-1}^{\text{in}}u\right\rangle}\right| \\ &\le\sup_{v\in\mathcal{V}_j^{\mathbb{C}}\setminus\{0\}}\left|\ln\frac{\langle\mathcal{G}v,v\rangle}{\langle\mathcal{H}v,v\rangle}\right|=d_\sigma(\mathcal{G},\mathcal{H})\ .\end{aligned}$$

The above holds trivially if $\pi_{j-1}^{\mathrm{in}} u = 0$ and $u \neq 0$, so (4.2) follows by taking the supremum over all $u \in \mathcal{V}_j^{\mathbb{C}} \setminus \{0\}$. □

Note that for symmetric operators Lemma 4.2 reduces to Lemma 5.1 in [7]. However, as mentioned earlier, neither $\mathcal{G}_j^{\mathrm{in}}$ nor $\mathcal{M}_j^{\mathrm{in}}$ are expected to be symmetric with respect to the L^2-inner product, hence the need for Lemma 4.2.

Lemma 4.3. *If the operators $\mathcal{K}, \mathcal{K}_j$ satisfy Condition* 2.1 *with constant C, then there exist $\delta > 0$ and $C_1 = C_1(\beta, C)$ so that, if $\mathcal{X} \in \mathfrak{L}(\mathcal{V}_j^{\mathrm{in}})$ satisfies $d_\sigma(\mathcal{X}, (\mathcal{G}_j^{\mathrm{in}})^{-1}) < \delta$, then*

$$d_\sigma(\mathfrak{N}_j(\mathcal{X}), (\mathcal{G}_j^{\mathrm{in}})^{-1}) \leq C_1 \left(d_\sigma(\mathcal{X}, (\mathcal{G}_j^{\mathrm{in}})^{-1}) \right)^2 \ . \tag{4.3}$$

The proof of Lemma 4.3 is similar to that of Theorem 3.12 in [7], except that here none of the operators involved is symmetric with respect to the L^2-inner product. Unfortunately, the lack of appropriate analysis tools for the non-symmetric case (see also the comment following Lemma A.1) allow for a weaker than desired result, with C_1 depending on β instead of being universal. For completeness we give the proof of Lemma 4.3 in Appendix A.

We recall Lemma 5.3 from [7] as

Lemma 4.4. *Let $(e_j)_{j \geq 0}$ and $(a_j)_{j \geq 0}$ be positive numbers satisfying*

$$e_j \leq C(e_{j-1} + a_j)^2 \ , \quad a_j \leq a_{j-1} \leq f^{-1} a_j, \quad j = 1, 2, \ldots, \tag{4.4}$$

for some $0 < f < 1$. If $a_0 \leq \frac{f}{4C}$ and if $e_0 \leq 4Ca_0^2$, then

$$e_j \leq 4Ca_j^2, \quad \forall j > 0 \ . \tag{4.5}$$

Theorem 4.5. *Assume that the operators $\mathcal{K}$ and $\mathcal{K}_j$ satisfy Condition* 2.1 *and the weights $w_i^{(j)}$ are uniform. Given an index $j_{\max} \geq 1$ let the index-sets $\mathcal{I}^{(j)}$, $0 \leq j \leq j_{\max}$, be given as in Algorithm* 2.5, *and consider the corresponding inactive domains Ω_j^{in}. If μ_j^{in} is the Lebesgue measure of $\partial_{\mathrm{n}} \Omega_j^{\mathrm{in}}$, it is assumed that there exists $f \in (1/4, 1)$ so that $\mu_j^{\mathrm{in}} \leq \mu_{j-1}^{\mathrm{in}} \leq f^{-2} \mu_j^{\mathrm{in}}$ for $j = 1, 2, \ldots, j_{\max}$ and that $\beta^{-1}(h_0^2 + \sqrt{\mu_0^{\mathrm{in}}}) \leq f/(4C_1)$ with C_1 as in Lemma* 4.3. *Then there exists a constant $C_2 = C_2(C_1, \beta)$ so that*

$$d_\sigma\left(\mathcal{Z}_{j_{\max}}^{\mathrm{in}}, (\mathcal{G}_{j_{\max}}^{\mathrm{in}})^{-1}\right) \leq C_2 \beta^{-1} \left(h_{j_{\max}}^2 + \sqrt{\mu_{j_{\max}}^{\mathrm{in}}} \right) \ . \tag{4.6}$$

Proof. With $\mathcal{Z}_j^{\mathrm{in}}$ given by Algorithm 2.6, denote by $e_j = d_\sigma(\mathcal{Z}_j^{\mathrm{in}}, (\mathcal{G}_j^{\mathrm{in}})^{-1})$ and by $a_j = C\beta^{-1}(h_j^2 + \sqrt{\mu_j^{\mathrm{in}}})$ with C as in Theorem 2.2. The assumption on μ_j^{in} implies that $a_j \leq a_{j-1} \leq f^{-1} a_j$. From Theorem 2.2 we know that $d_\sigma(\mathcal{S}_j^{\mathrm{in}}, (\mathcal{G}_j^{\mathrm{in}})^{-1}) \leq a_j$. Then, for $1 \leq j \leq j_{\max} - 1$

$$\begin{aligned}
e_j &= d_\sigma(\mathcal{Z}_j^{\mathrm{in}}, (\mathcal{G}_j^{\mathrm{in}})^{-1}) = d_\sigma(\mathfrak{N}_j(\mathfrak{I}_{j-1}^{j}(\mathcal{Z}_{j-1}^{\mathrm{in}})), (\mathcal{G}_j^{\mathrm{in}})^{-1}) \\
&\overset{(4.3)}{\leq} C_1 \left(d_\sigma(\mathfrak{I}_{j-1}^{j}(\mathcal{Z}_{j-1}^{\mathrm{in}}), (\mathcal{G}_j^{\mathrm{in}})^{-1}) \right)^2 \\
&\leq C_1 \left(d_\sigma(\mathfrak{I}_{j-1}^{j}(\mathcal{Z}_{j-1}^{\mathrm{in}}), \mathcal{S}_j^{\mathrm{in}}) + d_\sigma(\mathcal{S}_j^{\mathrm{in}}, (\mathcal{G}_j^{\mathrm{in}})^{-1})) \right)^2 \\
&\overset{(2.34)}{=} C_1 \left(d_\sigma(\mathfrak{I}_{j-1}^{j}(\mathcal{Z}_{j-1}^{\mathrm{in}}), \mathfrak{I}_{j-1}^{j}((\mathcal{G}_{j-1}^{\mathrm{in}})^{-1}) + d_\sigma(\mathcal{S}_j^{\mathrm{in}}, (\mathcal{G}_j^{\mathrm{in}})^{-1})) \right)^2 \\
&\overset{(4.2)}{\leq} C_1 \left(d_\sigma(\mathcal{Z}_{j-1}^{\mathrm{in}}, (\mathcal{G}_{j-1}^{\mathrm{in}})^{-1}) + a_j \right)^2 = C_1 \left(e_{j-1} + a_j \right)^2 \ .
\end{aligned}$$

By Lemma 4.4 we have

$$e_{j_{\max}-1} \leq 4C_1 a^2_{j_{\max}-1} \ . \tag{4.7}$$

Finally

$$\begin{aligned} e_{j_{\max}} &= d_\sigma(\mathcal{Z}^{\text{in}}_{j_{\max}}, (\mathcal{G}^{\text{in}}_{j_{\max}})^{-1}) \leq d_\sigma(\mathcal{Z}^{\text{in}}_{j_{\max}}, \mathcal{S}^{\text{in}}_{j_{\max}}) + d_\sigma(\mathcal{S}^{\text{in}}_{j_{\max}}, (\mathcal{G}^{\text{in}}_{j_{\max}})^{-1}) \\ &\overset{(4.2)}{\leq} e_{j_{\max}-1} + a_{j_{\max}} \overset{(4.7)}{\leq} 4C_1 a^2_{j_{\max}-1} + a_{j_{\max}} \leq (4C_1 f^{-2} + 1) a_{j_{\max}} \ , \end{aligned}$$

which proves (4.6) with $C_2 = (4C_1 f^{-2} + 1)$. ☐

REMARK 4.6. *As stated in Remark* 2.3*, under certain, natural conditions it is expected that* $\mu^{\text{in}}_j \approx Ch_j$*, case in which the factor* f *in Theorem* 4.5 *is* $1/\sqrt{2}$*. However, while the estimate* (4.6) *is (asymptotically) as good as the two-grid estimate in terms of* $h_{j_{\max}}$ *and* $\mu^{\text{in}}_{j_{\max}}$*, the dependence of* C_2 *on* β *is suboptimal. However, we believe that this is only an artifact of the analysis of which the weak link is Lemma A*.1.

5. Numerical experiments.

We test our algorithm on the standard linear elliptic-constrained distributed optimal control problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\left\|y - y_d\right\|^2 + \frac{\beta}{2}\left\|u\right\|^2 \\ \text{subject to} & -\Delta y = u \ , \ y \in H^1_0(\Omega), \ a \leq u \leq b \ \ a.e. \text{ in } \Omega, \end{array} \tag{5.1}$$

where Δ is the Laplace operator acting on $H^1_0(\Omega)$, so $\mathcal{K} = (-\Delta)^{-1}$. The problem (5.1) is often used as a test example for multigrid algorithms in PDE constrained optimization (e.g., see [4, 9]). Parts **[a, b]** of Condition 2.1 follow from standard estimates for finite element solutions of elliptic equations, while condition **[c]** follows from standard L^∞-estimates [5]. Note that the L^∞-stability in **[c]** does not necessitate an optimal L^∞-convergence rate of the discrete solutions to their continuous counterparts, but merely convergence:

$$\lim_{h_j \to 0} \|\mathcal{K}_j u - \mathcal{K} u\|_{L^\infty} = 0 \ .$$

We conduct three kinds of experiments to test different aspects of the method: first we consider an 'in vitro' experiment where we artificially fix a nontrivial subset of the domain Ω based on which we construct (artificial) inactive sets on all grids. The goal of this experiment is to assess the change in quality of the two-grid preconditioner while varying only the mesh-size and the regularizing parameter β. For each grid we then construct the two-grid preconditioner and we estimate numerically the decay rate of the spectral distance $d_\sigma((\mathcal{G}^{\text{in}}_j)^{-1}, \mathcal{S}^{\text{in}}_j)$. The advantage of this approach is that we can isolate the effect of the inactive set on the quality of the preconditioner, since in the context of actually applying the SSNM, the inactive sets are expected to change from one resolution to another. The second test is an actual, 'in vivo' application of the two-preconditioner while solving (5.1) using the SSNM and we compare its performance with Hackbusch's multigrid method of the second kind. In the third set of tests we compare actual runtimes of the linear solves based on symmetric multigrid-preconditioned CG with those performed with unpreconditioned CG in an actual SSNM solution process.

5.1. One-dimensional 'in vitro' experiments.

Let $\Omega = [0, 1]$ and designate the interval $\Omega^{\text{in}} = [1/8, 3/4]$ as the set where 'inactive vertices' reside; this set is used

TABLE 5.1
Spectral distance decay for a fixed inactive domain $[1/8, 3/4] \subset \Omega = [0, 1]$.

grid-size (n_j)	16	32	64	128	256	512
d_j ($\beta = 1$)	$2.28 \cdot 10^{-3}$	$1.56 \cdot 10^{-3}$	$1.08 \cdot 10^{-3}$	$7.46 \cdot 10^{-4}$	$5.20 \cdot 10^{-4}$	$3.64 \cdot 10^{-4}$
$\frac{d_{j-1}}{d_j}$ ($\beta = 1$)	–	1.4610	1.4506	1.4421	1.4351	1.4295
d_j ($\beta = 0.1$)	$2.18 \cdot 10^{-2}$	$1.49 \cdot 10^{-2}$	$1.03 \cdot 10^{-2}$	$7.16 \cdot 10^{-3}$	$4.99 \cdot 10^{-3}$	$3.49 \cdot 10^{-3}$
$\frac{d_{j-1}}{d_j}$ ($\beta = 0.1$)	–	1.4589	1.4491	1.4410	1.4343	1.4290
d_j ($\beta = 0.01$)	$1.59 \cdot 10^{-1}$	$1.10 \cdot 10^{-1}$	$7.66 \cdot 10^{-2}$	$5.34 \cdot 10^{-2}$	$3.74 \cdot 10^{-2}$	$2.62 \cdot 10^{-2}$
$\frac{d_{j-1}}{d_j}$ ($\beta = 0.01$)	–	1.4459	1.4398	1.4343	1.4296	1.4257

for all grids. Note that in this experiment we do not solve a control problem, and the designated inactive nodes do not correspond to a SSNM iterate. Let $n_j = 16 \cdot 2^j, j = 0, 1, 2, \dots$, and define $\mathcal{T}_j$ to be the uniform grid with n_j intervals on Ω with mesh-size $h_j = 1/n_j$. The 'inactive indices' are given by

$$\mathcal{I}_j^{\text{in}} = \{i \in \{1, \dots, n_j - 1 \ : \ i\, h_j \in \Omega^{\text{in}}\} \ .$$

Correspondingly, we construct the matrices $\mathbf{G}_j^{\text{in}}$ and $\mathbf{M}_j^{\text{in}}$ as in Section 2.4.2 and we compute the quantities

$$d_j = \max\{|\ln \lambda| \ : \ \lambda \in \sigma(\mathbf{G}_j^{\text{in}}, \mathbf{M}_j^{\text{in}})\} \ , \ j = 1, 2, \dots \ ,$$

where $\sigma(\mathbf{A}, \mathbf{B})$ is the set of generalized eigenvalues of the matrices $\mathbf{A}, \mathbf{B}$. In general we have

$$d_j \leq d_\sigma(\mathbf{G}_j^{\text{in}}, \mathbf{M}_j^{\text{in}}) \ ,$$

but if both matrices are symmetric then the above inequality becomes an equality. In this case $\mathbf{G}_j^{\text{in}}$ is symmetric and $\mathbf{M}_j^{\text{in}}$ is close to being symmetric, so we expect that d_j is a good approximation of $d_\sigma(\mathbf{G}_j^{\text{in}}, \mathbf{M}_j^{\text{in}})$. For each of $\beta = 1, 0.1, 0.01$ we report in Table 5.1 both the numbers d_j and their ratios d_{j-1}/d_j for $j = 2, \dots, 6$, which clearly indicate that the numerical results are consistent with the estimate $d_j \leq C\, \sqrt{h_j}$ in Remark 2.3, namely we notice that $d_{j-1}/d_j \to \sqrt{2}$ for each β considered.

Another advantage of the one-dimensional example is that we can easily compute numerically all the generalized eigenvalues of $\mathbf{G}_j^{\text{in}}$ and $\mathbf{M}_j^{\text{in}}$. What may be surprising, is that, while the largest of them decay at the predicted rate of $\sqrt{h_j}$ as $h_j \downarrow 0$, for each grid most of the generalized eigenvalues are very close to 1. In fact, for a fixed grid, if we order the generalized eigenvalues according to their distance from 1, we notice an exponential decay of $|\lambda_i - 1|$ (see Figure 5.1). We also plot in Figure 5.2 the function $u \in \mathcal{V}_2^{\text{in}}$ associated with the generalized eigenvector $\mathbf{u}$ that corresponds to the generalized eigenvalue which is furthest from 1. What we notice is that u is large around the boundary of Ω^{in}. These facts suggest that, with the exception of a limited number of generalized eigenvalues associated with eigenvectors strongly related to the boundary of Ω^{in}, the generalized eigenvalues $\mathbf{G}_j^{\text{in}}$ and $\mathbf{M}_j^{\text{in}}$ are in fact quite close 1. We revisit this idea in Section 5.4.

5.2. Two-dimensional 'in-vivo' experiments.. We now consider the two-dimensional version of (5.1) with $\Omega = [0, 1]^2$, and, for simplicity, we impose only the lower bound-constraint $a \equiv 0$. To construct uniform triangular grids $\mathcal{T}_j$, $j = 0, 1, \dots,$

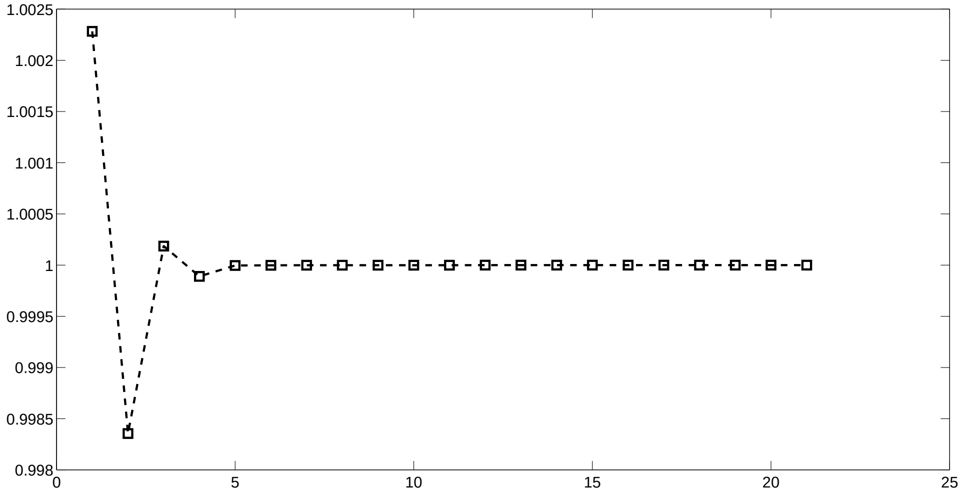

FIG. 5.1. *Absolute values of the generalized eigenvalues of* $\mathbf{G}_2^{\mathrm{in}}$ *and* $\mathbf{M}_2^{\mathrm{in}}$ $(n_2 = 32)$*. There are* 21 *eigenvalues, and their distance to* 1 *decays rapidly.*

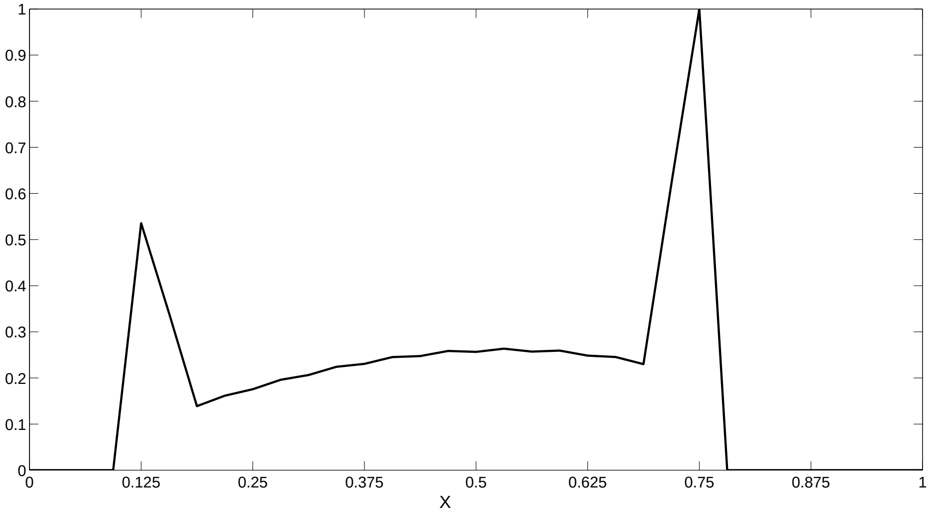

FIG. 5.2. *The generalized eigenvector corresponding to the generalized eigenvalue* $\mathbf{G}_2^{\mathrm{in}}$ *and* $\mathbf{M}_2^{\mathrm{in}}$ *which is furthest from* 1.

we first divide each side of Ω uniformly in $n_j = 64 \cdot 2^j$ intervals to obtain a uniform rectangular grid, and we further divide each resulting grid-square in two triangles along the diagonal of slope -1; the mesh-size is thus $h_j = 1/n_j$, and the number of variables is $N_j = (n_j - 1)^2$. The Poisson problem is then discretized using standard continuous piecewise linear elements on a triangular grid. We solve the problem using grid-sequencing, that is, the solution on level $(j-1)$ is used as an initial guess for the semismooth Newton iteration at level j. More precisely, if Ω_{j-1}^{in} denotes the coarse inactive domain as defined in Section 2.4, we define the initial guess at the inactive set on level j by

$$\mathcal{I}_0^{(j)} = \{i \in \{1, \dots, N_j\} \;:\; \text{supp}(\varphi_i^{(j)}) \subseteq \Omega_{j-1}^{\text{in}}\} \ .$$

Then we apply the semismooth Newton iteration described in Section 2.2, and we solve the reduced linear systems (2.16) at each outer iteration using CGS preconditioned by the two-grid explicit preconditioner $\mathcal{M}_j^{\text{in}}$. For comparison, we also solve (2.16) (using the same coarse spaces and transfer operators defined in Section 2.4) using Hackbusch's multigrid method of the second kind (e.g., see [12], Ch. 16, or [11] for the original source). In order to do so we remark that the matrix in (2.16) has the form $\mathbf{G}^{\text{in},k} = \beta\mathbf{I} + \mathbf{H}$, therefore we can rewrite the system $(\beta\mathbf{I} + \mathbf{H})\mathbf{u} = \mathbf{r}$ as a fixed point problem:

$$\mathbf{u} = \beta^{-1}\mathbf{r} - \beta^{-1}\mathbf{H}\mathbf{u} \ . \tag{5.2}$$

We also solve the systems (2.16) using unpreconditioned CG. Since the solution process is matrix-free (we solve the Poisson problem using classical multigrid) we use the explicit form $\mathcal{S}_j^{\text{in}}$ of the inverse of $\mathcal{M}_j^{\text{in}}$ as given in (2.25). For each outer iteration we record the number of two-grid preconditioned CGS iterations required for solving each linear system (2.16), as well as the number of iterations in which the multigrid (here, two-grid) method of the second kind converged. Note that each CGS iteration requires two fine-grid matrix-vector multiplications (mat-vecs), and the same holds for the multigrid method of the second kind, so this should be a fair comparison.

For the numerical example we choose as 'target' control the function defined by

$$u_d(x) = \begin{cases} r^{-4}(r^2 - \|x - x_0\|^2) + \alpha & , \quad \text{if } \|x - x_0\| < r \ , \\ \alpha & , \quad \text{otherwise} \end{cases}$$

for some $x_0 \in \Omega$ (see Figure 5.3), and let $\widetilde{y}$ be the solution of the Poisson equation

$$-\Delta\widetilde{y} = u_d, \quad \widetilde{y}|_{\partial\Omega} \equiv 0 \ .$$

The 'data' y_d entering the control problem (5.1) are obtained by adding to $\widetilde{y}$ a random perturbation that is uniformly distributed in $[-\delta, \delta]$ with $\delta = 0.05\,\|\widetilde{y}\|_{L^\infty}$. If $u_d \geq 0$, and if δ and β are small then the solution $u^{\min}$ of (5.1) is expected to be close to u_d. Therefore, by letting the parameter α be slightly negative we expect to find a localized $u^{\min}$ with nontrivial active and inactive sets, as desired for testing the performance of our algorithm. In Figure 5.4 we show the solution $u^{\min}$ for $\alpha = -0.1$ and $n_0 = 64$; as it turns out, for this example about 11.5% of the constraints are inactive if $\beta = 10^{-5}$, and about 51% are inactive if $\beta = 10^{-4}$.

The data in Table 5.2 (respectively 5.3) shows the iteration counts for the two-level preconditioned CGS method (columns labeled 'A'), the multigrid method of the second kind (two levels only, columns 'B'), and for unpreconditioned CG (columns 'C'), with the regularizing parameter being $\beta = 10^{-4}$ (respectively $\beta = 10^{-5}$). For example, in Table 5.2 on the third grid ($n_2 = 256$) the SSNM converged in 4 iterations; for the first outer iteration the linear system was solved in 4 two-grid preconditioned CGS iterations, while the remaining three outer iterations required 5 CGS iterations; for the same systems the multigrid method of the second kind converged in 9 iterations, while CG required 13 iterations. First we should point out that the results shown

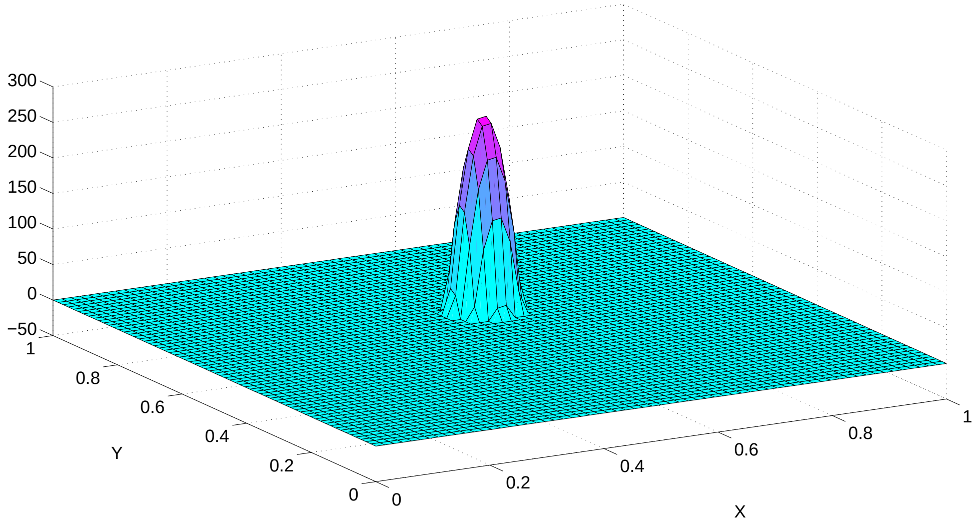

FIG. 5.3. *Target control u_d with $x_0 = [0.54, 0.62]^T$, $r = 0.06$, $\alpha = -0.1$.*

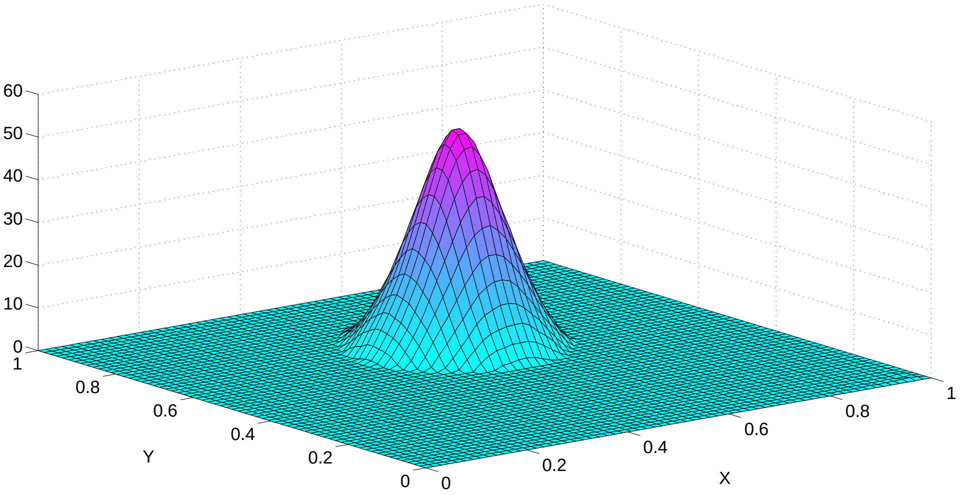

FIG. 5.4. *Solution $u^{\min}$ on coarsest grid ($n_0 = 64$) and $\beta = 10^{-5}$; about 11.5% of constraints are inactive.*

are consistent with the theoretical estimates of Theorem 2.2 and Remark 2.3, namely the number of two-grid preconditioned CGS iterations decreases with $h_j \downarrow 0$. There is one notable exception: for the very first iterate on the coarsest mesh we used as initial guess $\mathcal{I}_0^{(0)} = \{1, \ldots, N_0\}$, so the two-grid preconditioner is the same as for the unconstrained problem, thus very efficient. In order to answer the question of whether the two-grid preconditioned CGS is more efficient than CG we adopt the

Table 5.2

Iteration counts for the two-level preconditioned CGS method (A)*, Hackbusch* (B)*, and unpreconditioned CG* (C)*;* $\beta = 10^{-4}$*.*

	64			128			256			512			1024			2048		
it.	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
1	2	7	13	4	11	12	4	9	13	4	7	13	3	6	12	3	6	12
2	7	16	13	6	12	12	5	9	13	4	8	13	4	7	12	4	6	12
3	6	17	12	6	12	12	5	9	13	4	8	13	4	7	12	4	6	12
4	6	17	12	6	12	12	5	9	13	4	8	13	–	–	–	–	–	–
5	6	17	12	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

Table 5.3

Iteration counts for the two-level preconditioned CGS method (A)*, Hackbusch* (B)*, and unpreconditioned CG* (C)*;* $\beta = 10^{-5}$*; 'nc' means 'not converged'.*

	128			256			512			1024			2048		
it.	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
1	4	nc	31	6	nc	14	5	nc	14	5	18	12	4	13	11
2	11	nc	16	30	nc	12	7	nc	13	6	21	11	5	13	12
3	11	nc	15	9	nc	12	7	nc	12	6	21	11	5	13	12
4	9	nc	15	9	nc	12	7	nc	12	6	21	11	5	13	12
5	11	nc	17	–	–	–	–	–	–	–	–	–	–	–	–
5	16	nc	16	–	–	–	–	–	–	–	–	–	–	–	–
5	16	nc	16	–	–	–	–	–	–	–	–	–	–	–	–

point of view that in a truly large-scale context (three- or four-dimensional problems, or when applying the operators $\mathcal{K}_j$ requires a complicated sequence of operations) the cost of the solution process is largely given by fine-grid mat-vecs. If we accept the number of fine-space mat-vecs as a measure of efficiency we notice that the two-grid preconditioned CGS solves required a total of 11 CGS iterations, hence 22 mat-vecs at the finest level of $n_5 = 2048$, compared to 36 mat-vecs needed by CG. As can be inferred from the tables, the ratio (# multigrid mat-vecs / # CG mat-vecs) is decreasing with mesh-size. However, the above ratio would decrease much faster if $d_\sigma((\mathcal{G}_j^{\text{in}})^{-1}, \mathcal{S}_j^{\text{in}})$ would decay at a faster rate, as it does in the unconstrained case. The multigrid method of the second kind also behaves as expected, in the sense that the number of iterations decreases with $h_j \downarrow 0$. However, as the data suggest, it requires more iterations than the two-grid preconditioned CGS method. As the theory suggests both for the preconditioner introduced in this work and for the multigrid method of the second kind, the base level cannot be too coarse, or the algorithm will be inefficient or even fail to converge. As can be seen from Table 5.3, the multgrid method of the second kind only converges when $h_j \leq 2^{-10}$, while the multigrid preconditioned CGS method exhibits variable behavior at coarser levels (but does converge). Finally, we should remark that the algorithm performed surprisingly well considering that β was in fact chosen quite small compared to what the theory suggested ($\beta \approx C\sqrt{h}$).

5.3. Multigrid results. For our multigrid experiments with problem (5.1) we used CG in conjunction with the multigrid preconditioner derived from the symmetric two-grid preconditioner (2.27). This time, in order to show that our method is not restricted to triangular grids we discretize the PDE-constraints using continuous piecewise bilinear elements, and we eliminated any boundary restrictions on the controls; the coarse space is defined similarly to the one introduced in Section 2.4 (see (2.19)). Due to the problem sizes involved we solve Poisson's equation using standard multigrid – the full approximation scheme – with a base case of 256 intervals in each direction; anytime we solve the PDE at a resolution lower than 256×256 we employ a direct solver. Recall that each Hessian-vector multiplication in the SSNM solution process requires two PDE solves. Note that the multigrid solve for Poisson's equation is completely separated from the multigrid preconditioner for the optimization problem. The entire algorithm is implemented in MATLAB (version R2010a) with the exception of the operation of identifying the coarse inactive nodes, which was implemented in C (and compiled with mex); the reason behind this choice is related to the unavoidable 'for-loop' in this part of the algorithm which adds significant runtime in an artificial manner to an otherwise simple routine. The computations were performed on a computer with two 8-core 2.9 GHz Intel Xeon E5-2690 processors with 256 GB RAM.

In our experiments we measured the added wall-clock time of all the linear solves in the SSNM solution process (including all the actions related to setting up the multigrid preconditioner at each SSNM iteration), as well as the wall-clock time for all the remaining operations. We should point out that, aside from the linear solve, each SSNM iterate requires two additional mat-vecs, regardless of the method chosen for solving the linear systems. Since our efforts were focused on solving the linear systems, we compare the added runtimes for the multigrid preconditioned linear solves with those of the unpreconditioned CG solves.

Of all the experiments we conducted we show three that seem to represent the typical behavior of our method. In the first experiment we let $\beta = 10^{-4}$, and we considered uniform $n \times n$ grids on $[0, 1 \times [0, 1]$ with $n = 64, 128, 512, 1024, 2048, 4096$. In order to ensure a sufficiently large active set for the solution we imposed a slightly elevated lower constraint, namely $a \equiv 0.1$. We set the tolerance for solving the linear systems at $\tau = 10^{-8}$. The results are reported in Table 5.4. In the row marked '# cg its' we list the number of unpreconditioned CG iterations for each SSNM outer iteration. For example, for $n = 1024$ there were three SSNM iterations, and at each SSNM iteration the corresponding linear system required 11 unpreconditioned CG iterations to converge. Similarly, in the row marked '# mg its ($n_0 = 64$)', **the same** three linear systems required 8, 7, and 7 multigrid-preconditioned CG iterations; for all multigrid preconditioners in this experiment the base case is set at $n_0 = 64$. Below the rows listing the number of iterations we show the added runtimes of the linear solves; for multigrid these include the time to set up the coarse spaces, and for constructing and inverting the coarsest Hessian. Finally in the row marked '$t_{\text{mg}}/t_{\text{cg}}$' we record the speed-up of the multigrid-preconditioned solves over the unpreconditioned solves. In terms of number of iterations we note an already familiar behavior: the number of unpreconditioned CG iterations remains bounded (for this example it turned out to be constant), while the number of multigrid-preconditioned CG solves decreases slowly with increasing resolution. The speed-up follows the same pattern as the number of multigrid preconditioned CG iterations, slowly decreasing from above-unity for small-to-medium resolutions (2.87 for two-grid preconditioners at $n_1 = 128$) to 0.67 at $n_8 = 4096$. Computing solutions at resolutions higher than 4096×4096 with our current tools requires parallel programming and falls beyond the scope of this work.

For the second set of computations (see Table 5.5) we decreased β to 10^{-5} and chose the lower constraint $a \equiv 0$. We also lowered the tolerance $\tau = 10^{-10}$ in order to increase the overall number of iterations. While the behavior of unpreconditioned

Table 5.4
Comparison of iteration counts and runtimes for multigrid vs. unpreconditioned CG; $\beta = 10^{-4}$, $a = 0.1, b = 1$, tol $= 10^{-8}$.

n	128	256	512	1024	2048	4096
# cg its	11, 11, 11, 11	11, 11, 11, 11	11, 11, 11	11, 11, 11	11, 11, 11	11, 11, 11
t_{cg} (s)	4.24	25.75	104.26	490.74	2358	13553
# mg its ($n_0 = 64$)	10, 9, 9, 9	10, 10, 10, 10	9, 9, 9	8, 7, 7	8, 8, 8	7, 7, 7
t_{mg} (s)	12.17	28.03	159.22	583.19	2454	9195
$t_{\text{mg}}/t_{\text{cg}}$	2.87	1.09	1.53	1.19	1.04	0.68

CG remains essentially consistent with the earlier example (the number of iterations remains fairly constant), the number of multigrid-preconditioned CG iterations with base $n_0 = 64$ shows a significant increase with increasing resolution. This fact is consistent with the analysis, in that the estimates hold only if the base case is sufficiently fine. If we increase the base case from 64 to 128 the 'good' behavior is observed again, with the number of iterations slowly decreasing with increasing resolution. Of course, the linear systems to be solved at the new base case $n_0 = 128$ are significantly larger than for base case 64, and the runtimes are seriously affected at moderate resolutions, resulting in 'speedups' of ≈ 41 at $n = 256$ and ≈ 8.4 at $n = 512$. However, at $n = 4096$ we obtain an actual speedup of 0.63. This trend suggests that at even higher resolutions multigrid-preconditioned CG will outperform unpreconditioned CG even more.

Table 5.5
Comparison of iteration counts and runtimes for multigrid vs. unpreconditioned CG; $\beta = 10^{-5}$, $a = 0, b = 1$, tol $= 10^{-10}$.

n	256	512	1024	2048	4096
# cg its	23, 24, 23, 23	23, 24, 24	24, 24, 24	25, 24, 25	25, 25, 24
t_{cg} (s)	42.59	187.70	1034	5042	29385
# mg its ($n_0 = 64$)	20, 20, 19, 19	24, 24, 24	22, 23, 22	26, 26, 24	42, 49, 48
# mg its ($n_0 = 128$)	18, 15, 15, 15	18, 17, 17	18, 18, 18	16, 14, 17	16, 16, 16
t_{mg} (s)	1747	1577	2700	5858	18580
$t_{\text{mg}}/t_{\text{cg}}$	41.02	8.4	2.61	1.16	0.63

5.4. Further remarks. We return to the numerical results from Section 5.1, which suggest that the two-grid preconditioner is closer to being of optimal order than Theorem 2.2 predicts, in the following sense: if restricted to subspaces of functions supported away from the boundary of Ω_j^{in}, namely spaces of the form

$$\mathcal{V}_{j,H}^{\text{in}} = \text{span}\{\varphi_i^{(j)} \in \mathcal{V}_j^{\text{in}} \ : \ \text{dist}(\text{supp}(\varphi_i^{(j)}), \partial\Omega_j^{\text{in}}) \geq H\} \ ,$$

with $H > h$, then the $\widetilde{H}^{-2}(\Omega)$-approximation property of π_{j-1}^{in} is expected to be of almost optimal order. To see this, consider $u \in \mathcal{V}_{j,H}$ for sufficiently large H. Then $\pi_{j-1}^{\text{in}} u \approx 0$ on $\partial_{\text{n}}\Omega_j^{\text{in}}$, since the size of $\pi_{j-1}^{\text{in}} u$ decays exponentially fast away from supp(u), and therefore the numerical-boundary term in (3.11) from the proof of Lemma 3.3 can be almost be dropped. Since (3.9) was the source of the largest term in the estimate (3.13) we expect that

$$\|\text{Proj}_{\mathcal{V}_{j,H}^{\text{in}}}(\mathcal{G}_j^{\text{in}} - \mathcal{M}_j^{\text{in}})|_{\mathcal{V}_{j,H}^{\text{in}}}\| \leq Ch_j^2 \ .$$

While it is not difficult to formalize the above argument we should point out its main consequence: the number of generalized eigenvalues in $\sigma(\mathcal{G}_j^{\mathrm{in}}, \mathcal{M}_j^{\mathrm{in}})$ that are significantly far from 1 is proportional to the number of nodes supported near the boundary of Ω_j^{in}. If the dimension d is greater than 1, then this number normally increases with resolution, and certainly exceeds the relatively low number of iterations we found in Section 5.2. This indicates that the number of generalized eigenvalues which are $O(\sqrt{h_j})$ away from 1 dominate the computations. While the two-level preconditioner does not have optimal order, the above argument together with the analysis from Section 3 also suggest that in order to improve the presented preconditioning technique one has to tackle the diverging action of the two operators on the nodal basis functions that are supported near $\partial\Omega_j^{\mathrm{in}}$.

6. Conclusions. We have constructed multigrid preconditioners for the linear systems arising in the semismooth Newton solution process for a control problem constrained by smoothing linear operators with box-constraints on the control. The preconditioner is similar to the one constructed by Drăgănescu and Dupont in [7] for the problem without inequality control-constraints, and it maintains some of its qualitative behavior: its approximation properties improve with increasing resolution. However, even though the approximation qualities of the constructed preconditioner are of suboptimal order, the construction and the analysis form an important stepping stone towards finding optimal order preconditioners for the systems under scrutiny. The question of their practical importance and how they compare in efficiency with the similar preconditioners developed by Drăgănescu and Petra in [9] for systems arising in interior point methods requires a more involved study and forms the subject of current research.

Acknowledgements. The author is indebted to Todd Dupont for introducing him to this subject and for guiding his initial steps in multigrid research and to Michael Hintermüller for helpful discussions on semismooth Newton methods. The author also thanks the referees for their thorough reading of the manuscript and useful suggestions.

Appendix A. Some technical results regarding the spectral distance. Given a Hilbert space $(V, \langle\cdot,\cdot\rangle)$ and an operator $A \in \mathfrak{L}_+(V)$ we define the bilinear form

$$\langle u, v\rangle_A = \langle Au, v\rangle \ .$$

For $B \in \mathfrak{L}(V)$ denote

$$w_A(B) = \sup_{u\in V\setminus\{0\}} \left| \frac{\langle Bu, u\rangle_A}{\langle u, u\rangle_A} \right|$$

If A is Hermitian, then $\langle\cdot,\cdot\rangle_A$ is an inner product, case in which $w_A(B)$ is the numerical radius of B with respect to $\langle\cdot,\cdot\rangle_A$, and the power inequality

$$w_A(B^n) \le (w_A(B))^n, \quad n = 1, 2, 3, \ldots$$

holds (e.g., see Theorem 2.1-1 in [10]). The next result is a weak generalization of the inequality above for the case when A is non-symmetric and $n = 2$.

Lemma A.1. *There exists a constant $C_A = C(\|A_a^{-1}\|, \kappa(A), \|A\|)$ so that*

$$w_A(B^2) \le C_A(w_A(B))^2, \tag{A.1}$$

where A_s, A_a are the Hermitian and skew-Hermitian parts of A, and $\kappa(A) = \|A\| \cdot \|A^{-1}\|$.

Proof. First we identify two constants c_1, c_2 so that $c_1\|B\| \le w_A(B) \le c_2\|B\|$. Let $\alpha = \|A_a^{-1}\|^{-1} > 0$ be the smallest eigenvalue of A_a. We have

$$\left|\frac{\langle Bu, u\rangle_A}{\langle u, u\rangle_A}\right| = \left|\frac{\langle ABu, u\rangle}{\langle Au, u\rangle}\right| \le \alpha^{-1}\|A\| \cdot \|B\| \ , \tag{A.2}$$

since $|\langle Au, u\rangle| = |\langle A_s u, u\rangle + \langle A_a u, u\rangle| \ge \alpha\|u\|^2$, because $\langle A_a u, u\rangle \in \mathbb{R}$ and $\langle A_a u, u\rangle \in \mathbf{i}\mathbb{R}$. By taking the supremum over all $u \in V \setminus \{0\}$ we obtain

$$w_A(B) \le \alpha^{-1}\|A\| \cdot \|B\| \ . \tag{A.3}$$

From the polarization identity

$$\begin{aligned} 4\langle Bu, v\rangle_A = {} & \langle B(u+v), u+v\rangle_A - \langle B(u-v), u-v\rangle_A \\ & +\mathbf{i}\langle B(u+\mathbf{i}v), u+\mathbf{i}v\rangle_A - \mathbf{i}\langle B(u-\mathbf{i}v), u-\mathbf{i}v\rangle_A \end{aligned}$$

we obtain

$$\begin{aligned} 4|\langle Bu, v\rangle_A| &\le w_A(B)\|A\|\left(\|u+v\|^2 + \|u-v\|^2 + \|u+\mathbf{i}v\|^2 + \|u-\mathbf{i}v\|^2\right) \\ &= 4w_A(B)\|A\|\left(\|u\|^2 + \|v\|^2\right) \ . \end{aligned}$$

By taking the supremum over all unit vectors $u, v \in V$ we obtain $\|AB\| \le 2w_A(B)\|A\|$. Hence

$$\|B\| \le \|AB\|\|A^{-1}\| \le 2w_A(B)\kappa(A) \ .$$

So $c_1 = (2\kappa(A))^{-1}$ and $c_2 = \alpha^{-1}\|A\|$. Therefore

$$w_A(B^2) \le c_2\|B^2\| \le c_2\|B\|^2 \le c_2 c_1^{-2} w_A(B)^2 \ ,$$

so $C_A = c_2 c_1^{-2}$. □

We should point out that if A is Hermitian, then, cf. [10] we have $C_A = 1$. Mounting numerical evidence suggests that C_A is bounded uniformly with respect to A even for the non-Hermitian case; however, proving this claim is still part of ongoing research.

We now prove Lemma 4.1.

Proof. For two complex numbers $z_1, z_2 \in \mathbb{H}$ we have

$$\mathrm{Re}\left(\overline{(z_2 - z_1)}(\ln z_2 - \ln z_1)\right) > 0 \ . \tag{A.4}$$

To verify (A.4) consider the segment $\gamma : [0,1] \to \mathbb{C}$ given by $\gamma(t) = z_1 + t(z_2 - z_1)$. Then

$$\begin{aligned} & \mathrm{Re}\left(\overline{(z_2 - z_1)}(\ln z_2 - \ln z_1)\right) \\ & = \mathrm{Re}\left(\overline{(z_2 - z_1)}\int_\gamma \frac{dz}{z}\right) = \mathrm{Re}\left(|z_2 - z_1|^2 \int_0^1 \frac{dt}{z_1 + t(z_2 - z_1)}\right) \ , \end{aligned}$$

and (A.4) follows from $(z_1 + t(z_2 - z_1))^{-1} \in \mathbb{H}$ for $t \in [0,1]$. Given arbitrary $w_1, w_2 \in \mathbb{H}$ define the function

$$f : [0, \infty) \to \mathbb{R}, \quad f(x) = \left|\ln\frac{w_1 + x}{w_2 + x}\right|^2 \ .$$

We have

$$f'(x) = 2\,\mathrm{Re}\Big((\overline{(w_1+x)^{-1}} - \overline{(w_2+x)^{-1}})(\ln(w_1+x) - \ln(w_2+x)) \Big) < 0,$$

as follows from (A.4) with $z_i = (w_i+x)^{-1}, i = 1, 2$. Therefore f is decreasing and (4.1) follows. □

We now conclude with the proof of Lemma 4.3.

Proof. First note that

$$\mathfrak{N}_j(\mathcal{X}) - (\mathcal{G}_j^{\mathrm{in}})^{-1} = 2\mathcal{X} - \mathcal{X}\mathcal{G}_j^{\mathrm{in}}\mathcal{X} - (\mathcal{G}_j^{\mathrm{in}})^{-1} = -(\mathcal{G}_j^{\mathrm{in}})^{-1}(\mathcal{G}_j^{\mathrm{in}}\mathcal{X} - I)^2\ .$$

Since for z around 1 we have $\ln z = z - 1 + O(z-1)$ there are $k > 1$ and $\delta > 0$ so that, if $z \in \mathcal{B}_\delta(1)$ or $|\ln z| \le \delta$, then

$$k^{-1}|z-1| \le |\ln z| \le k|z-1|.$$

Denote by $A = (\mathcal{G}_j^{\mathrm{in}})^{-1}$ and assume that $d_\sigma(\mathcal{X}, A) \le \delta$ with δ to be determined. For $u \in \mathcal{V}_j^{\mathbb{C}} \setminus \{0\}$ and $A = (\mathcal{G}_j^{\mathrm{in}})^{-1}$

$$\begin{aligned}\left|\frac{\langle \mathfrak{N}_j(\mathcal{X})u, u\rangle}{\langle Au, u\rangle} - 1\right| &= \left|\frac{\big\langle A(\mathcal{G}_j^{\mathrm{in}}\mathcal{X} - I)^2 u, u\big\rangle}{\langle Au, u\rangle}\right| \le w_A\left((\mathcal{G}_j^{\mathrm{in}}\mathcal{X} - I)^2\right)\\ &\le C_A\left(w_A(\mathcal{G}_j^{\mathrm{in}}\mathcal{X} - I)\right)^2\ ,\end{aligned}$$

where C_A is the constant in from Lemma A.1 (depending on β, C). Since

$$w_A(\mathcal{G}_j^{\mathrm{in}}\mathcal{X} - I) = \sup_{u \in \mathcal{V}_j^{\mathbb{C}} \setminus \{0\}} \left|\frac{\langle \mathcal{X}u, u\rangle}{\langle Au, u\rangle} - 1\right| \le k\, d_\sigma(\mathcal{X}, A)\ ,$$

we obtain

$$\left|\frac{\langle \mathfrak{N}_j(\mathcal{X})u, u\rangle}{\langle Au, u\rangle} - 1\right| \le C_A k^2 \left(d_\sigma(\mathcal{X}, A)\right)^2 \le C_A k^2 \delta^2\ .$$

By further restricting δ so that $C_A k^2 \delta^2 < \delta$ we get

$$\sup_{u \in \mathcal{V}_j^{\mathbb{C}} \setminus \{0\}} \left|\ln \frac{\langle \mathfrak{N}_j(\mathcal{X})u, u\rangle}{\langle Au, u\rangle}\right| \le k \sup_{u \in \mathcal{V}_j^{\mathbb{C}} \setminus \{0\}} \left|\frac{\langle \mathfrak{N}_j(\mathcal{X})u, u\rangle}{\langle Au, u\rangle} - 1\right| \le C_A k^3 (d_\sigma(\mathcal{X}, A))^2\ .$$

□

REFERENCES

[1] Volkan Akçelik, George Biros, Andrei Drăgănescu, Omar Ghattas, Judith C. Hill, and Bart G. van Bloemen Waanders, *Dynamic data driven inversion for terascale simulations: real-time indentification of airborne contaminants*, in SC '05: Proceedings of the 2005 ACM/IEEE conference on Supercomputing, Washington, DC, USA, 2005, IEEE/ACM, IEEE Computer Society.

[2] George Biros and Günay Doğan, *A multilevel algorithm for inverse problems with elliptic PDE constraints*, Inverse Problems, 24 (2008), pp. 034010, 18.

[3] Alfio Borzì and Karl Kunisch, *A multigrid scheme for elliptic constrained optimal control problems*, Comput. Optim. Appl., 31 (2005), pp. 309–333.

[4] Alfio Borzi and Volker Schulz, *Multigrid methods for PDE optimization*, SIAM Rev., 51 (2009), pp. 361–395.

[5] Dietrich Braess, *Finite elements*, Cambridge University Press, Cambridge, third ed., 2007. Theory, fast solvers, and applications in elasticity theory, Translated from the German by Larry L. Schumaker.

[6] Susanne C. Brenner and L. Ridgway Scott, *The mathematical theory of finite element methods*, vol. 15 of Texts in Applied Mathematics, Springer, New York, third ed., 2008.

[7] Andrei Drăgănescu and Todd F. Dupont, *Optimal order multilevel preconditioners for regularized ill-posed problems*, Math. Comp., 77 (2008), pp. 2001–2038.

[8] Andrei Drăgănescu and Ana Maria Soane, *Multigrid solution of a distributed optimal control problem constrained by the Stokes equations*, Appl. Math. Comput., 219 (2013), pp. 5622–5634.

[9] Andrei Drăgănescu and Cosmin Petra, *Multigrid preconditioning of linear systems for interior point methods applied to a class of box-constrained optimal control problems*, SIAM Journal on Numerical Analysis, 50 (2012), pp. 328–353.

[10] Karl E. Gustafson and Duggirala K. M. Rao, *Numerical range*, Universitext, Springer-Verlag, New York, 1997.

[11] Wolfgang Hackbusch, *Die schnelle Auflösung der Fredholmschen Integralgleichung*, Beiträge zur Numerischen Mathematik, 9 (1981), pp. 47–62.

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

[13] ———, *Integral equations*, vol. 120 of International Series of Numerical Mathematics, Birkhäuser Verlag, Basel, 1995. Theory and numerical treatment, Translated and revised by the author from the 1989 German original.

[14] Martin Hanke and Curtis R. Vogel, *Two-level preconditioners for regularized inverse problems. I. Theory*, Numer. Math., 83 (1999), pp. 385–402.

[15] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim., 13 (2002), pp. 865–888 (electronic) (2003).

[16] Michael Hintermüller and Michael Ulbrich, *A mesh-independence result for semismooth Newton methods*, Math. Program., 101 (2004), pp. 151–184.

[17] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich, *Optimization with PDE constraints*, vol. 23 of Mathematical Modelling: Theory and Applications, Springer, New York, 2009.

[18] R. H. W. Hoppe and R. Kornhuber, *Adaptive multilevel methods for obstacle problems*, SIAM J. Numer. Anal., 31 (1994), pp. 301–323.

[19] Barbara Kaltenbacher, *V-cycle convergence of some multigrid methods for ill-posed problems*, Math. Comp., 72 (2003), pp. 1711–1730 (electronic).

[20] J. Thomas King, *Multilevel algorithms for ill-posed problems*, Numer. Math., 61 (1992), pp. 311–334.

[21] Andreas Rieder, *A wavelet multilevel method for ill-posed problems stabilized by Tikhonov regularization*, Numer. Math., 75 (1997), pp. 501–522.

[22] M. Stoll and A. Wathen, *Preconditioning for partial differential equation constrained optimization with control constraints*, Numerical Linear Algebra with Applications, 19 (2012), pp. 53–71.

[23] Fredi Tröltzsch, *Optimal control of partial differential equations*, vol. 112 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, 2010. Theory, methods and applications, Translated from the 2005 German original by Jürgen Sprekels.

[24] Michael Ulbrich. Habilitation, 2001. TU Munich.

[25] ———, *Semismooth Newton methods for operator equations in function spaces*, SIAM J. Optim., 13 (2002), pp. 805–842 (electronic) (2003).

[26] ———, *Semismooth Newton methods for variational inequalities and constrained optimization problems in function spaces*, vol. 11 of MOS-SIAM Series on Optimization, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011.

[27] Michael Ulbrich and Stefan Ulbrich, *Primal-dual interior-point methods for PDE-constrained optimization*, Math. Program., 117 (2009), pp. 435–485.

[28] Martin Weiser, *Interior point methods in function space*, SIAM J. Control Optim., 44 (2005), pp. 1766–1786 (electronic).