ARGONNE NATIONAL LABORATORY
9700 South Cass Avenue
Lemont, IL 60439

Convergence of Sum-Up Rounding Schemes for the Electromagnetic Cloaking Problem

Sven Leyffer, Paul Manns, and Malte Winckler

Mathematics and Computer Science Division

Preprint ANL/MCS-P9268-1019

December 10, 2019

This material was based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, under Contract DE-AC02-06CH11357.

Contents

Convergence of Sum-Up Rounding Schemes for the Electromagnetic Cloak Problem

Sven Leyffer[*] Paul Manns[*][†] Malte Winckler[‡]

December 10, 2019

Abstract

We consider the problem of designing an electromagnetic cloak from an integer programming point of view. The problem can be modeled as a PDE-constrained optimization problem with integer-valued control inputs that are distributed in the computational domain. A first-discretize-then-optimize approach results in a large-scale mixed-integer nonlinear program that is in general intractable because of the large number of integer variables that arise from the discretization of the domain. Instead, we propose an efficient algorithm that is able to approximate the local infima of the underlying non convex infinite-dimensional problem arbitrarily close without the need to solve the discretized finite-dimensional integer programs to optimality. We optimize only the continuous relaxations of the approximations for local minima and then apply the sum-up rounding methodology to obtain integer-valued controls. These controls are shown to converge and exhibit the desired approximation properties under suitable refinements of the involved discretization grids. Our results use familiar concepts arising from the analytical properties of the underlying PDE and complement previous results, derived from a topology optimization point of view.

1 Introduction

The problem of designing electromagnetic cloaks can be described as follows (see, e.g., [25]). For a given incident wave, we design a scatterer in a predefined region D_s such that an object in another region D_o is hidden (i.e. protected) from the incident wave. This means that the scatterer is designed such that the scattered wave cancels the incident wave in this region; in other words, the amplitude of the superposition of the two fields is as small as possible. A 2D scenario is sketched in Figure 1. The problem has been cast and treated as a topology optimization problem in [11] and as a mixed-integer PDE-constrained optimization (MIPDECO) problem in [27]. We analyze and investigate the latter formulation. Specifically, we optimize for a discrete-valued function, which assigns the material of the scatterer to each point in the region D_s such that it minimizes the superposition of the scattered electromagnetic field and the incident wave in the region D_o in a least-squares sense. Although we take the MIPDECO point of view, we will make use of the same

[*] Argonne National Laboratory

[†] Technische Universität Braunschweig, email: paul.manns@tu-bs.de

[‡] Universität Duisburg-Essen

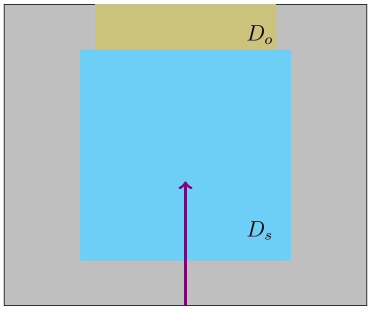


Figure 1: Illustration of our cloaking scenario. The violet arrow indicates the angle of the incident wave, the scatterer's region D_s is colored blue, and the protected object's region D_o is colored olive.

analytical properties of the Helmholtz equation as in [11] and thus derive MIPDECO counterparts of the results from [11].

From the computational point of view, the authors of [11] present a radial-basis function-based level-set method that uses a relaxed (i.e., fractional) formulation but produces designs of the scatterer such that fractional (i.e. non-physical) material placement occurs only in a small part of the domain. In contrast to their approach, we solve the relaxed problem directly and execute a computationally efficient rounding algorithm that computes a weak approximation to the relaxed control. We employ the sum-up rounding (SUR) algorithm, introduced in [22] and whose approximation properties have been analyzed in [10, 14, 18, 23]. Recently, these results have been transferred from one-dimensional problems to multidimensional problems [17, 31]. We emphasize that the runtime complexity of SUR is $\mathcal{O}(N)$, where N is the number of cells that discretize the domain of the control input function.

We show that by using compactness properties of the control-to-state operator of the Helmholtz equation, we obtain an approximation of the electromagnetic field and the optimized objective value in the norm topology. Our method, however, suffers from the drawback that the design may exhibit a lot of chattering and therefore may not be implementable in reality. We demonstrate computationally that applying a median filtering as a postprocessing step can improve the implementability of the resulting control considerably while increasing the objective value moderately. However, the filtering step is a heuristic, and we cannot guarantee that the approximation properties of the SUR approximation are conserved.

We also observe in our computational results that the objective value of the local minimizers of the continuous relaxation converges to zero, which suggests that it may be possible to cloak the incident wave entirely.

We note that the relaxation we solve is a bilinear optimal control problem, and we refer the reader to [13, 30] for general studies of bilinear optimal control problems concerned with (electromagnetic) wave phenomena.

1.1 Structure of the Paper

We introduce the Helmholtz equation and summarize its solution theory in Section 2. We state the investigated optimal problem formally and derive the first-order optimality conditions of its continuous relaxation in Section 3. We investigate the approximation algorithm in Section 4, and we present the computational results in Section 5. We summarize our conclusions in Section 6.

1.2 Notation

For an optimization problem (P), we denote its feasible set by $\mathcal{F}_{(\mathrm{P})}$. For a domain D and $k \geq 1$, $H^k(D)$ denotes the Sobolev space of all functions in $L^2(D)$ (i.e., square-integrable functions) with distributional derivatives up to order k in $L^2(D)$). The space $L^2_{loc}(D)$ is the space of locally square-integrable functions that is functions that are square integrable over compact sets. Similarly, $H^k_{loc}(D)$ is the space of functions in $L^2_{loc}(D)$ with distributional derivatives up to order k in $L^2_{loc}(D)$. The Borel sigma algebra for a set $D \subset \mathbb{R}^d$ is denoted by $\mathcal{B}(D)$. The inner product of a Hilbert space H is denoted by $(\cdot,\cdot)_H$. We denote the compact embedding of a Hilbert space H into another Hilbert space K by $H \hookrightarrow^c K$. For a Banach space X, we denote convergence of a sequence $(x_n)_n \subset X$ to a limit $x \in X$ in the norm topology by $x_n \to x$; convergence in the weak topology by $x_n \rightharpoonup x$; and convergence in the weak* topology by $x_n \rightharpoonup^* x$. For further background on concepts from PDE theory, see [21]. In the interest of a less distracting presentation, we will sometimes abbreviate $L^p(\mathbb{R}^d)$ by L^p, in particular for constant quantities.

2 State Equation

We begin by stating the main assumptions on the regularity of the domain and the given data, which we tacitly assume to hold throughout the remainder of the paper.

Assumption 2.1 (Domain and boundary)**.** *Let $D \subset \mathbb{R}^d$, $d \in \{2,3\}$, be a bounded domain with a sufficiently smooth boundary such that the following assumptions hold:*

1. *In the case of $d = 2$, the trace operator $\cdot|_{\partial D} : H^1(D) \to L^2(\partial D)$ exists and is continuous.*
2. *In the case of $d = 3$, the bounded trace operators $\cdot|_{\partial D} : H^2(D) \to H^{3/2}(\partial D)$, $\partial \cdot /\partial n|_{\partial D} : H^2(D) \to H^{1/2}(\partial D)$ as well as the bounded extension operator $E : H^2(D) \to H^2(\mathbb{R}^d)$ exist.*

Assumption 2.2 (Given data)**.** *Let $D \subset \mathbb{R}^d$, $d \in \{2,3\}$, satisfy Assumption 2.1. We assume that the scatterer $q \in L^\infty(D)$ is non-negative (i.e., $q \geq 0$ a.e. in D) and that the incident field $u_0 \in H^2_{loc}(\mathbb{R}^d)$ satisfies the homogeneous equation*

$$-\Delta u_0 - k_0^2 u_0 = 0 \quad \text{on } \mathbb{R}^d.$$

We develop our theoretical results in the context of the two state equations described next.

2.1 Helmholtz Equation with Sommerfeld Radiation Condition

The first state equation is the Helmholtz equation with a *Sommerfeld radiation condition* in dimension $d = 3$:

$$\begin{aligned} -\Delta u - k_0^2(1 + qw)u &= k_0^2 qwu_0 \text{ in } D, \\ -\Delta u - k_0^2 u &= 0 \text{ in } \mathbb{R}^3\backslash\overline{D}, \\ [u]_{\partial D} = 0 \text{ and } \left[\frac{\partial u}{\partial n}\right]_{\partial D} &= 0, \\ \lim_{r\to\infty}\sqrt{r}\sup_{\xi\in\mathbb{S}^2}\left|\frac{\partial u(r\xi)}{\partial r} - ik_0 u(r\xi)\right| &= 0. \end{aligned} \tag{2.1}$$

Here, the state variable is denoted by u, and we have an essentially bounded control input w, that is $w \in L^\infty(D)$, with $w(x) \in [w_\ell, w_u]$ for almost all (a.a.) $x \in D$ and real positive scalars $0 \le w_\ell < w_u < \infty$. The third line of (2.1) contains the so-called transmission conditions at the boundary of the domain D due to the smoothness of the incident wave u_0, where the jumps of u and $\partial u/\partial n$ for $x \in \partial D$ are defined by $[u]_{\partial D}(x) \coloneqq \lim_{\overline{D}^c\supset x^n\to x} u(x^n) - \lim_{D\ni y^n\to x} u(y^n)$ and $[\partial u/\partial n]_{\partial D}(x) \coloneqq n(x)^T\left(\lim_{\overline{D}^c\supset x^n\to x}\nabla u(x^n) - \lim_{D\ni y^n\to x}\nabla u(y^n)\right)$. The fourth line contains the so-called Sommerfeld radiation condition, which ensures that the field (wave) u is outgoing; see [7, Chap. 1].

We summarize the existence and uniqueness of solutions of the state equation (2.1) along the lines of [7, Chap. 8]. We define the operator $K : L^2(D) \to H^2(D)$ as

$$K : L^2(D) \ni \psi \mapsto k_0^2\int_D \Phi(x,\cdot)\psi(x)\mathrm{d}x \in H^2(D),$$

where Φ denotes the fundamental solution of the Helmholtz equation in $\mathbb{R}^3$; see [7, Sect. 2.1].

Proposition 2.3 ([7, Theorem 8.2]). *Let Assumption 2.2 be satisfied for $d = 3$. If $u \in H^2_{loc}(\mathbb{R}^3)$ solves (2.1), then u solves the Lippmann–Schwinger equation*

$$u - K((1 + qw)u) = K(u_0 qw). \tag{2.2}$$

Moreover if $u \in C(\mathbb{R}^3)$ solves (2.2), then $u \in H^2_{loc}(\mathbb{R}^3)$ and u solves (2.1). ■

Proposition 2.4 ([7, Theorem 8.7]). *Under Assumptions 2.1 and 2.2 for $d = 3$, there exists a unique solution $u \in H^2_{loc}(\mathbb{R}^3)$ of (2.1) for every $k_0 > 0$ and $w \in L^\infty(D)$ with $w(s) \in [w_\ell, w_u]$ a.e.* ■

We introduce the set $\mathfrak{D}$, which is the domain of the control-to-state operator of the state equation and the Lippmann–Schwinger equation:

$$\mathfrak{D} \coloneqq \{w \in L^\infty(D) : w(s) \in [w_\ell, w_u] \text{ a.e.}\}.$$

To obtain continuous dependence of the solution of (2.1) on $w \in \mathfrak{D}$, we prove the following proposition.

Proposition 2.5. *Let Assumptions 2.1 and 2.2 be satisfied for $d = 3$ and let $K_0 > 0$ such that*

$$1 - K_0^2\|\Phi\|_{L^2}(1 + \|q\|_{L^\infty} w_u) > 0$$

holds, where Φ *denotes the fundamental solution of the Helmholtz equation. Then, for every* $k_0 \in (0, K_0)$*, the control-to-state operator of the Lippmann–Schwinger equation* (2.2) $S : \mathfrak{D} \to H^2(D)$ *given by* $\omega \mapsto u|_D$ *is continuous. Furthermore, the control-to-state operator is also* L^2*-*L^∞*-Lipschitz continuous. In other words, there exists* $L > 0$ *such that*

$$\|S(w_1) - S(w_2)\|_{L^2(D)} \leq L\|w_1 - w_2\|_{L^\infty(D)}$$

for all w_1*,* $w_2 \in \mathfrak{D}$*.*

Proof. First, we show that for every $w \in \mathfrak{D}$ the corresponding solution $u = S(w)$ to (2.2) satisfies the estimate $\|u\|_{L^2(D)} \leq C\|w\|_{L^\infty(D)}$ for some constant $C > 0$ for $k_0 > 0$ sufficiently small. Inserting the definition of K into (2.2) and testing the equation with u, we obtain

$$\begin{aligned}\Big(u, u\Big)_{L^2(D)} &- k_0^2 \left(\int_D \Phi(x, \cdot)(1 + q(x)w(x))u(x)\mathrm{d}x, u\right)_{L^2(D)} \\ &= k_0^2 \left(\int_D \Phi(x, \cdot)q(x)w(x)u_0(x)\mathrm{d}x, u\right)_{L^2(D)},\end{aligned}$$

and thus by virtue of the Cauchy–Schwarz inequality and Hölder's inequality, we obtain

$$\begin{aligned}&\|u\|^2_{L^2(D)} \Big(1 - k_0^2\|\Phi\|_{L^2}(1 + \|q\|_{L^\infty}\|w\|_{L^\infty(D)})\Big) \\ &\qquad \leq k_0^2\|\Phi\|_{L^2}\|u_0\|_{L^2}\|q\|_{L^\infty}\|w\|_{L^\infty(D)}\|u\|_{L^2(D)}.\end{aligned}$$

Consequently, for k_0^2 sufficiently small, we obtain

$$\|u\|_{L^2(D)} \leq \frac{k_0^2\|\Phi\|_{L^2}\|u_0\|_{L^2}\|q\|_{L^\infty}}{1 - k_0^2\|\Phi\|_{L^2}(1 + \|q\|_{L^\infty}w_u)}\|w\|_{L^\infty(D)}.$$

Next, we show the Lipschitz continuity of $u = S(w)$, that is we show that $\|u_1 - u_2\|_{L^2(D)} \leq L\|w_1 - w_2\|_{L^\infty(D)}$ for some constant $L > 0$ for $k_0 > 0$ sufficiently small and $u_1 = S(w_1)$, $u_2 = S(w_1)$ for w_1, $w_2 \in \mathfrak{D}$. We insert the pairs u_1, w_1 and u_2, w_2 into (2.2), subtract the two equations, and employ the linearity of K to obtain

$$u_1 - u_2 - K\big((1 + qw_1)u_1 - (1 + qw_2)u_2\big) = K\big(u_0q(w_1 - w_2)\big).$$

Inserting a suitable zero gives the equivalent identity

$$u_1 - u_2 - K((1 + qw_1)(u_1 - u_2)) = K(q(w_1 - w_2)(u_0 - u_2)).$$

We test this equation with $u_1 - u_2$, which gives

$$\begin{aligned}\Big(u_1 - u_2, u_1 - u_2\Big)_{L^2(D)} &- \Big(K((1 + qw_1)(u_1 - u_2)), u_1 - u_2\Big)_{L^2(D)} \\ &= \Big(K(q(w_1 - w_2)(u_0 - u_2)), u_1 - u_2\Big)_{L^2(D)}.\end{aligned} \tag{2.3}$$

The definition of K gives the lower bound using the Cauchy–Schwarz inequality

$$\begin{aligned}&- \Big(K((1 + qw_1)(u_1 - u_2)), u_1 - u_2\Big)_{L^2(D)} \\ &\qquad \geq -k_0^2\|\Phi\|_{L^2}(1 + \|q\|_{L^\infty}\|w_1\|_{L^\infty(D)})\|u_1 - u_2\|^2_{L^2(D)}\end{aligned}$$

for the second term in (2.3). We also obtain the upper bound

$$\begin{aligned}&\Big(K(q(w_1-w_2)(u_0-u_2)), u_1-u_2\Big)_{L^2(D)}\\ &\qquad \le k_0^2\|\Phi\|_{L^2}\|q\|_{L^\infty}\|w_1-w_2\|_{L^\infty}\|(\|u_0\|_{L^2}+C\|w_2\|_{L^\infty})\end{aligned}$$

for the right hand side of (2.3), where we have used the estimate $\|u_2\|_{L^2} \le C\|w\|_{L^\infty}$ that we derived in the first step. Using the definition of $\|u_1 - u_2\|_{L^2(D)}$ and $\|w_i\|_{L^\infty} \le w_u$ for $i = 1, 2$, we obtain

$$\|u_1-u_2\|_{L^2(D)} \le \frac{k_0^2\|\Phi\|_{L^2}(\|u_0\|_{L^2}+Cw_u)}{1-k_0^2\|\Phi\|_{L^2}(1+\|q\|_{L^\infty}w_u)}\|w_1-w_2\|_{L^2(D)}$$

from (2.3) for $k_0 > 0$ sufficiently small, which proves L^2-L^2-Lipschitz continuity, which in turn gives the desired L^2-L^∞-Lipschitz continuity because the first factor of the right-hand side is bounded.

To obtain continuity with respect to the codomain $H^2(D)$, we employ a bootstrapping argument. Let $w_n \to w$ in $L^\infty(D)$ with w, $w_n \in \mathfrak{D}$ for all $n \in \mathbb{N}$. Then, we obtain $u_n \to u$ in $L^2(D)$ with $u_n = S(w_n)$ and $u = S(w)$. Insertion into the summands of (2.2) yields

$$K((1+qw_n)u_n) \to K((1+qw)u)$$

and

$$K(u_0qw_n) \to K(u_0qw)$$

in $H^2(D)$ as $n \to \infty$ because of the respective continuity of the operator K; see [7, Thm. 8.2]. Thus, from (2.2), we obtain $u_n \to \bar{u}$ in $H^2(D)$ for some $\bar{u} \in H^2(D)$. By uniqueness of the solution of (2.2) for $k_0 > 0$ sufficiently small (see, e.g., [7, Thm. 8.4]), we obtain $\bar{u} = u$, which proves the claim. □

2.2 Helmholtz Equation with Robin Boundary Condition

The radiation boundary constraints in (2.1) are difficult to implement in practice. Instead, the authors in [8] suggest a *first-order absorbing boundary condition*, which is a Robin boundary condition in our setting, and we use these boundary conditions in our practical experiments in Section 5. In particular, we consider the following approximation of (2.1) in dimension $d = 2$

$$\begin{aligned}-\Delta u - k_0^2(1+qw)u &= k_0^2 qwu_0 \text{ in } D,\\ \frac{\partial u}{\partial n} - ik_0u &= 0 \text{ on } \partial D,\end{aligned} \tag{2.4}$$

where $w(s) \in [w_\ell, w_u]$ for a.a. $s \in D$ and where we have replaced the transmission conditions $[u]_{\partial D} = 0$ and $[\partial u/\partial n]_{\partial D} = 0$, and the Sommerfeld radiation condition by Robin boundary conditions. The state equation (2.4) can be interpreted as a first-order approximation at the boundary for the *physically correct* state equation (2.1), see [11, Sect. 2] and is analyzed in [5,11]. The existence and uniqueness of solutions follow from the results in [5].

Proposition 2.6 ([5, Lemmas 2.2 and 2.4])**.** *Let Assumption 2.2 hold. Then, for every $k_0 > 0$, the state equation (2.4) admits a unique weak solution in $H^1(D)$ for all $w \in \mathfrak{D}$. Moreover, for $k_0 > 0$ sufficiently small, the control-to-state operator $S : \mathfrak{D} \to H^1(D)$ is continuous.* ■

In this paper, we consider both the Sommerfeld radiation condition and the Robin boundary conditions. We present the algorithm in the context of the simpler Robin boundary conditions, noting that extensions to more sophisticated boundary conditions are straightforward.

3 Optimal Design of Cloaks

Our aim is to design an electromagnetic cloak that is effective in the region $D_o \subset D$. As in [11,27], we restrict the domain where the scatterer is designed, namely, where material may be put to some region $D_c \subset D \backslash D_o$. We are constrained by the state equation (2.4), and our goal is to minimize the electromagnetic field in the region D_o, which corresponds to the superposition of the incidence wave and the response u. The resulting integer optimal control problem becomes

$$\begin{aligned} &\inf_{u,w} \frac{1}{2}\|u+u_0\|^2_{L^2(D_o)} \\ &\text{s.t. } -\Delta u - k_0^2(1+qw)u = k_0^2 q w u_0 \text{ in } D, \\ &\qquad \frac{\partial u}{\partial n} - i k_0 u = 0 \text{ on } \partial D, \\ &\qquad w(x) \in \{w_1, \dots, w_M\} \text{ for a.a. } x \in D_c, \\ &\qquad w(x) = 0 \text{ for a.a. } x \in D \backslash D_c \end{aligned} \tag{P}$$

with discrete material constants $w_\ell = w_1 < \ldots < w_M = w_u$, and its continuous relaxation becomes

$$\begin{aligned} &\min_{u,w} \frac{1}{2}\|u+u_0\|_{L^2(D_o)} \\ &\text{s.t. } -\Delta u - k_0^2(1+qw)u = k_0^2 q w u_0 \text{ in } D, \\ &\qquad \frac{\partial u}{\partial n} - i k_0 u = 0 \text{ on } \partial D, \\ &\qquad w(x) \in [w_\ell, w_u] \text{ for a.a. } x \in D_c, \\ &\qquad w(x) = 0 \text{ for a.a. } x \in D \backslash D_c. \end{aligned} \tag{R}$$

First-Order Optimality System

In this section we derive a first-order optimality system for the minimization problem (R) governed by (2.4). We start with the Fréchet differentiability of the control-to-state operator $S\colon L^\infty(D) \to H^1(D)$.

Lemma 3.1. *Let Assumption 2.2 be satisfied. Then, for $k_0 > 0$ sufficiently small, the control-to-state operator $S\colon \mathfrak{D} \to H^1(D)$ is continuously Fréchet differentiable, and for every $w, h \in \mathfrak{D}$ its Fréchet-derivative $\hat{u} = S'(w)h \in H^1(D)$ is the unique solution of*

$$\int_D \nabla \hat{u} \cdot \nabla \overline{v} \mathrm{d}x - k_0^2 \int_D (1+qw)\hat{u}\overline{v}\mathrm{d}x - ik_0 \int_{\partial D} \hat{u}\overline{v} ds = k_0^2 \int_D qh(u+u_0)\overline{v}\mathrm{d}x \quad \forall v \in H^1(D) \tag{3.1}$$

where $u = S(w)$ and $\overline{v}$ is the complex conjugate of v.

Proof. Fréchet differentiability of the control-to-state operator follows from standard arguments (see, e.g., [26]), together with the continuity of the control-to-state operator $S\colon L^\infty(D) \to H^1(D)$ and $S(0) = 0$. □

Thanks to Lemma 3.1 we can compute the derivative of the reduced cost-functional

$$f(w) \coloneqq J(S(w)) = \frac{1}{2}\|S(w) + u_0\|_{L^2(D_o)}$$

and introduce a corresponding adjoint state.

Lemma 3.2. *Under Assumption 2.2, we consider $k_0 > 0$ sufficiently small such that the assertions of Lemma 3.1 hold. Let $w, h \in \mathfrak{D}$ and $u = S(w) \in H^1(D)$. Then there exists a unique solution $p \in H^1(D)$ to the adjoint equation*

$$\int_D \nabla p \cdot \nabla \overline{v} \, \mathrm{d}x - k_0^2 \int_D (1 + qw) p \overline{v} \, \mathrm{d}x - ik_0 \int_{\partial D} p \overline{v} \, ds = \int_{D_o} (\overline{u + u_0}) \overline{v} \, \mathrm{d}x \quad \forall v \in H^1(D). \tag{3.2}$$

Furthermore, the derivative f' with respect to w of the reduced objective satisfies

$$f'(w)h = k_0^2 \operatorname{Re} \left(\int_D qh(u + u_0) p \, \mathrm{d}x \right),$$

where $\operatorname{Re}(\cdot)$ *is the real part.*

Proof. Because $J : L^2(D) \to \mathbb{R}$ and $S : \mathfrak{D} \to H^1(D)$ are Fréchet differentiable by Lemma 3.2, we obtain by straightforward calculations with the chain rule that

$$f'(w)h = \int_{D_o} \operatorname{Re}(u + u_0) \operatorname{Re}(\hat{u}) \, \mathrm{d}x + \int_{D_o} \operatorname{Im}(u + u_0) \operatorname{Im}(\hat{u}) \, \mathrm{d}x, \tag{3.3}$$

where $\hat{u} = S'(w)h$ solves (3.1). Now, inserting $v = \overline{p}$ into (3.1) and $\overline{v} = \hat{u}$ into (3.2) yields, after substracting the resulting equations, that

$$k_0^2 \int_D qh(u + u_0) p \, \mathrm{d}x = \int_{D_o} (\overline{u + u_0}) \hat{u} \, \mathrm{d}x. \tag{3.4}$$

We make use of the fact that $\operatorname{Re}(\overline{a}b) = \operatorname{Re} a \operatorname{Re} b + \operatorname{Im} a \operatorname{Im} b$ in (3.3) and (3.4) to deduce that

$$f'(w)h = \operatorname{Re} \left(\int_{D_o} (\overline{u + u_0}) \hat{u} \, \mathrm{d}x \right) = k_0^2 \operatorname{Re} \left(\int_D qh(u + u_0) p \, \mathrm{d}x \right),$$

which completes the proof. □

We close this section by establishing a first-order optimality system for (R) governed by (2.4).

Theorem 3.3. *Let the assumptions from Lemma 3.2 be satisfied. If $(w^*, u^*) \in \mathcal{F}_{(\mathrm{R})}$ is a local solution of* (R) *governed by* (2.4)*, then the solution of the adjoint equation $p^* \in H^1(D)$, established in Lemma 3.2, satisfies*

$$\operatorname{Re} \left(\int_D q(w - w^*)(u^* + u_0) p^* \, dx \right) \geq 0 \quad \forall w \in \mathfrak{D}.$$

Proof. Following the classical argument in [26], w^* satisfies the variational inequality

$$f'(w^*)(w - w^*) \geq 0 \quad , \forall w \in \mathfrak{D}.$$

The choice $(w, u) = (w^*, u^*)$ in Lemma 3.2 implies that this inequality is equivalent to

$$0 \leq f'(w^*)(w - w^*) = k_0^2 \operatorname{Re} \int_D q(w - w^*)(u^* + u_0) p^* \, \mathrm{d}x, \forall w \in \mathfrak{D},$$

where $p^* \in H^1(D)$ is the unique solution of the adjoint equation (3.2). □

4 Solving the MIPDECO

We begin this section by stating the algorithm to solve (P) approximately, which has been adopted from [17]. Next, we introduce the sum-up rounding algorithm and summarize its properties. We build on these properties to prove our approximation properties for (2.1) and (2.4). Then, we prove the asymptotic results of the algorithm. To obtain implementable designs, we have added a median-filtering step, and we briefly comment on the properties of this step.

4.1 Optimization Algorithm

The optimization algorithm for solving (P) is given in Algorithm 1. It tailors Algorithm 1 from [17] to our setting and extends it by a median-filtering step. We start by introducing the main components of this algorithm.

The algorithm takes three inputs: an initial discretization ($\mathrm{R}_h^{(0)}$) of the continuous relaxation (R); an initial guess w_0^R for a local minimizer of ($\mathrm{R}_h^{(0)}$); and an initial rounding grid $T^{(0)}$, which is a set of cells that decompose the domain D. The algorithm proceeds iteratively. Each iteration n consists of five steps.

First, the discretization of the continuous relaxation is refined, yielding the finite-dimensional NLP ($\mathrm{R}_h^{(n)}$). We note that we do not require discretization cells of the finite-dimensional NLP ($\mathrm{R}_h^{(n)}$) to coincide with those of the rounding grid, although this is possible and an intuitive choice. The subscript h denotes that a grid constant is associated with the approximation. Second, the NLP ($\mathrm{R}_h^{(n)}$) is solved to (local) optimality using the local minimizer of the previous iteration w_{n-1}^R as the initial guess. The resulting local minimizer is denoted by w_n^R. This is detailed in Section 4.2.

Third, the rounding grid is refined; that is the cells are decomposed into smaller ones, yielding the rounding grid $T^{(n)}$. Fourth, the SUR algorithm is executed by using the continuous control w_n^R and the rounding grid $T^{(n)}$ as inputs. It computes a discrete-valued control w_n^S, which is constant per cell of the rounding grid $T^{(n)}$. These two steps and the necessary assumptions on the rounding grid and its refinement are detailed in Section 4.3.

Fifth, a median filtering w_n^F of w_n^S is computed as a postprocessing step. It is detailed in Section 4.5.

We summarize the algorithm as Algorithm 1 below.

Algorithm 1 Solving (P) approximately

```
Input: Initial guess $w_0^R$, an initial approximation ($\mathrm{R}_h^{(0)}$) of (R), and an initial rounding grid $T^{(0)}$.
  for $n = 1, \dots$ do
      ($\mathrm{R}_h^{(n)}$) ← refine approximation ($\mathrm{R}_h^{(n-1)}$) of (R), see Section 4.2
      $w_n^R$ ← solve ($\mathrm{R}_h^{(n)}$) with initialization $w_{n-1}^R$, see Section 4.2
      $T^{(n)}$ ← refine ($T^{(n-1)}$), see Definition 4.1
      $w_n^S$ ← SUR($w_n^R, T^{(n)}$)
      $w_n^F$ ← MEDIAN-FILTER($w_n^S$)
  end for
```

4.2 Refining and Solving the Continuous Relaxation

The first and second steps of the loop in Algorithm 1 deal with the solution of (R). First, the discretization of the relaxation is refined, giving ($\mathrm{R}_h^{(n)}$). The discretization may, for example, be obtained by using piecewise affine globally continuous finite elements to discretize the state equation (2.4) and undergo a uniform grid refinement strategy during the iterations. Second, the finite-dimensional problem ($\mathrm{R}_h^{(n)}$) is solved to (local) optimality. The optimization of the resulting finite-dimensional NLP can be executed with a bound-constrained nonlinear optimization algorithm. We highlight that the problem is not convex, since the control-to-state operator is not convex, and thus, we can expect only local optimality.

Our algorithm and its convergence results are agnostic to the specific choice or implementation of the discretization and the first and second steps. Their analysis is not part of this work because we deal with the integer optimal control aspect here. However, under the assumptions on the rounding grid refinement that are introduced in Section 4.3 for all weak* accumulation points $w^{R,*}$ of $(w_n^R)_n$ with approximating subsequences $w_{n_k}^R \rightharpoonup^* w^{R,*}$, we obtain $w_{n_k}^S \rightharpoonup^* w^{R,*}$ with the help of the results from [17].

If an optimal consistency principle holds, that is, if the sequence of local minimizers of the refined discretizations converges at least weakly* to a local minimizer of the continuous relaxation (R), this also yields optimality of the sequence $(w_n^S)_n$. For background and the proofs of convergence of solutions of such discretized optimization problems to the minimizer of the approximated infinite-dimensional one, we refer the reader to [1, 3, 12, 19] and the references therein. Moreover, for bilinear optimal control problems, similar results can be found in [16, 29]. To complement these contributions, which focus mainly on the a priori analysis, we cite [15, 28] for studies regarding a posteriori error estimators for optimal control problems with control constraints. We note that [28] also establishes the strong convergence of the resulting AFEM-algorithm.

4.3 Multidimensional Sum-Up Rounding

The refinement of the rounding grid in the third step of the loop in Algorithm 1 produces a rounding grid $T^{(n)}$ from the previous one $T^{(n-1)}$. We require a refinement that yields a weak* approximation of relaxed control by the discrete-valued control produced in the fourth step. This result depends both on a suitable refinement of the grid cells and on the ordering of the grid cells. We state a set of sufficient conditions below.

Definition 4.1 ([17, Definition 4.9]). *We call a sequence* $\left(\left\{T_1^{(n)}, \dots, T_{N^{(n)}}^{(n)}\right\}\right)_n \subset 2^{\mathcal{B}(D)}$ *of finite partitions of D an* order-conserving domain dissection *of D if*

1. $N^{(0)} = 1$, $T_1^{(0)} = D$;

2. $\left\{T_1^{(n)}, \dots, T_{N^{(n)}}^{(n)}\right\}$ *is a finite partition of D for all n;*

3. *for all n and for all $i \in \{1, \dots, N^{(n-1)}\}$, there exist $1 \le j < k \le N^{(n)}$ such that*

$$\bigcup_{\ell=j}^{k} T_\ell^{(n)} = T_i^{(n-1)},$$

i.e. the order of the grid cells is preserved from $n-1$ to n;

4. $\max_{i\in\{1,\dots,N^{(n)}\}} \lambda(T_i^{(n)}) \to 0$*, where* λ *denotes the Lebesgue measure in* $\mathbb{R}^d$*, that is, the maximum cell volume tends to zero; and*

5. the σ*-algebra generated by* $\bigcup_{n=1}^{\infty}\left\{T_1^{(n)},\dots,T_{N^{(n)}}^{(n)}\right\}$ *is* $\mathcal{B}(D)$*, the Borel* σ*-algebra of the set* D*.*

These properties can, in particular, be verified for the grids induced by the approximating sequences the approximating sequences of space-filling curves; see [17]. A space-filling curve is a surjective and continuous mapping of the unit interval to a unit hypercube of a higher dimension. Space-filling curves yield a decomposition of the domain into grid cells and induce an ordering of them. The approximants of space-filling curves induce a uniform refinement of the grid in every iteration, which yields properties 1, 2, 3 and 5 in Definition 4.1. Furthermore, in every iteration the approximating curve traverses the grid cells of the previous grid iterations in the same order as the previous approximating curves did. This implies is the order-conservation property 3 in Definition 4.1; see [17].

Remark 4.2. *The treatment of multidimensional domains by using order-conserving domain dissections is a recent development put forward in [17, 31]. Before, the SUR algorithm and other rounding schemes were applied to time-dependent problems; see, for example, [14, 23, 24]. The one-dimensional results apply when considering a time horizon* $[0,T]$ *as the domain and decomposing it into intervals. Using the intuitive consecutive ordering of the intervals along the time axis gives us the setting of [14, 23, 24]. In this sense, space-filling curves may be interpretated as a way to map a decomposition of the domain to consecutive intervals discretizing a time horizon.*

The fourth step of Algorithm 1 is the application of the SUR algorithm (see [17, 23]) to the control w_R^n using the rounding grid $T^{(n)}$. The algorithm makes use of a special ordered set of type 1 (SOS1) encoding of the discrete control realizations; see [23]. To make this formal, we introduce the notions of binary and relaxed controls.

Definition 4.3 (Binary and relaxed control [17, Definition 2.2])**.** *Let* $D \subset \mathbb{R}^d$ *be a bounded domain. A measurable function* $\omega : D \to \{0,1\}^M$ *such that* $\sum_{i=1}^M \omega_i = 1$ *a.e. in* D *holds is called binary control and a measurable function* $\alpha : D \to [0,1]^M$ *such that* $\sum_{i=1}^M \alpha_i = 1$ *a.e. in* D *holds is called relaxed control.*

SUR computes binary controls ω, which are constant per grid cell from relaxed controls α. The relaxed control α can be obtained from a feasible continuous control w_n^R by solving $w_n^R(x) = \sum_{i=1}^M \alpha_i(x) w_i$ for a.a. $x \in D$. Once the binary control is computed, the discrete-valued control w_n^S, which is feasible for (P), can be reconstructed with $w_n^S(x) \coloneqq \sum_{i=1}^M \omega_i(x) w_i$. We note that α is in general nonunique and a consistent way to compute α from w_n^R has to be chosen in Algorithm 1, that is, we have to make sure that if the w_n^R converge to w^*, then we also require that the corresponding α_n converge to α^*. This will be a requirement of Theorem 4.10 to prove convergence of Algorithm 1. One option to achieve this (see [17]) is to represent $w_n^S(x)$ for $x \in D$ by a convex combination of its neighboring points in $\{w_1,\dots,w_M\}$. To this end we choose i, $\alpha_i(x)$ such that $w_n^R(x) = \alpha_i(x)w_i + (1-\alpha_i(x))w_{i+1}$, we set $\alpha_{i+1}(x) \coloneqq 1-\alpha_i(x)$, and we set $\alpha_j(x) \coloneqq 0$ for all $j \in \{1,\dots,i-1,i+2,\dots,M\}$.

SUR computes a binary control ω that is constant per grid cell. To this end, it iterates over the grid cells $1,\dots,N$ and assigns a value to ω on the respective grid cell. In each iteration, it first computes the integrated signed difference between the relaxed control α and the resulting binary control ω over the grid cells, where ω is already defined. Second, the weighted mean of α over the current grid cell is added to this quantity. Third and last, ω is set to 1 on this grid cell in the entry where this sum has its maximum value and zero in the others. If a tie arises with respect to the maximizing entry, the smallest applicable index is chosen. We summarize these steps in Algorithm 2, where χ_A denotes the indicator function of the set A.

Algorithm 2 SUR (Multidimensional sum-up rounding)

```
Input: Rounding grid $T = \{S_1, \ldots, S_N\}$, and relaxed control $\alpha$.
Output: Binary control $\omega$.
  for $i = 1, \ldots, N$ do
    $\Phi_i := \int_{\bigcup_{\ell=1}^{i-1} S_\ell} \alpha(x) - \omega(x) \mathrm{d}x$
    $\gamma_i := \Phi_i + \int_{S_i} \alpha(x) \mathrm{d}x$
    $\tilde{\omega}_{i,j} = \begin{cases} 1 & : \ j = \arg\max_{k \in \{1,\ldots,M\}} \gamma_{i,k}, \\ 0 & : \ \text{else}, \end{cases} \quad \text{for all } j \in \{1, \ldots, M\}$
  end for
  return $\omega = \sum_{i=1}^{N} \chi_{S_i} \tilde{\omega}_i$
```

4.4 Properties of Sum-Up Rounding

The multidimensional SUR algorithm given in Algorithm 2 allows us to approximate a relaxed control, α, with a binary control, ω. This approximation does not happen in the norm topology, because the no binary control can approximate any relaxed control in norm. Instead, by executing SUR on a sequence of refined rounding grids, which constitute an order-conserving domain dissection we obtain weak* convergence of the corresponding sequence of ω^n to α in $L^\infty(D)$, denoted by $\omega_n \rightharpoonup^* \alpha$ in $L^\infty(D)$, namely,

$$\int_D f(x)\omega_n(x)\mathrm{d}x \to \int_D f(x)\alpha(x)\mathrm{d}x \text{ for all } f \in L^1(D).$$

Proposition 4.4. *Let α be a relaxed control and $w_\alpha = \sum_{i=1}^{M} \alpha_i(s) w_i$ for $w \in \mathfrak{D}$. Then, the multidimensional SUR algorithm applied to a sequence of finite partitions of D that constitute an order-conserving domain dissection produces a sequence of binary controls $(\omega_n)_n$ with corresponding $w_{\omega_n} = \sum_{i=1}^{M} (\omega_n)_i w_i$ such that*

1. *$\omega_n(s) \in \{0, 1\}$ for all $s \in D$ and for all $n \in \mathbb{N}$;*
2. *$\sup_{i \in \{1,\ldots,N^{(n)}\}} \left\| \int_{\bigcup_{j=1}^{i} T_j^{(n)}} \alpha(s) - \omega_n(s) \mathrm{d}s \right\|_\infty \leq C \max_{i \in \{1,\ldots,N^{(n)}\}} \lambda(S_i^{(n)})$ for some $C > 0$ independent of the sequence of partitions, its indexing, and α;*
3. *$\omega_n \rightharpoonup^* \alpha$ in $L^\infty(D, \mathbb{R}^M)$; and*
4. *$w_{\omega_n} \rightharpoonup^* w_\alpha$ in $L^\infty(D)$.*

Proof. This follows immediately from [17, Theorem 4.7]. □

4.5 Median Filtering

The fifth step of Algorithm 1 consists of executing the algorithm MEDIAN-FILTER. The algorithm takes a set of cells $S_1, \dots, S_N$ and a function w that is constant per cell (i.e., it can be represented as $w \in \mathbb{R}^N$) as inputs. The algorithm iterates through all cells of a rounding grid and computes the median value of all cells neighboring the current cell. Here, a cell is neighboring another one if the distance between them is below a certain threshold r. Since we consider cells as sets, we need to define the distance between two sets S, $T \in \mathbb{R}^d$ which we do by means of the Pompeiu-Hausdorff distance. That is, we define

$$d(S,T) \coloneqq \max \left\{ \sup_{s \in S} \inf_{t \in T} \|s - t\|_{\mathbb{R}^D}, \sup_{t \in T} \inf_{s \in S} \|s - t\|_{\mathbb{R}^D} \right\}.$$

We summarize the algorithm as Algorithm 3.

Algorithm 3 MEDIAN-FILTER

Input: Piecewise constant function w on a grid consisting of cells $S_1, \dots, S_N$ with vector representation $w \in \mathbb{R}^N$ and radius r of the filter.

Output: Piecewise constant function w^F on the same grid on which w is defined in vector representation $w^F \in \mathbb{R}^N$

```
for $n = 1, \dots, N$ do
    $w_i^F \leftarrow \arg\min_{w_j : d_{PH}(S_i, S_j) < r} |w_j - w_i|$
end for
```

4.6 Approximation Result for the Sommerfeld Radiation Condition

Before stating our main theorem, we prove the following preparatory lemma.

Lemma 4.5. *Let Assumption 2.2 hold. For $k_0 > 0$ sufficiently small, the control-to-state operator $S : \mathfrak{D} \to H^2(D)$ is completely continuous in the sense that $w_n \rightharpoonup^* w$ in $L^\infty(D)$ with w, $w_n \in \mathfrak{D}$ for all $n \in \mathbb{N}$ yields $S(w_n) \to S(w)$ in $H^2(D)$.*

Proof. By virtue of Propositions 2.3 and 2.5, we have a continuous control-to-state operator $S : \mathfrak{D} \to H^2(D)$ of the Lippmann–Schwinger equation (2.2) and the Helmholtz equation (2.1). Assumption 2.1 yields the compact embedding $H^2(D) \hookrightarrow^c L^2(D)$; see, for example, [21, Chap. 7].

We take the point of view of the Lippmann–Schwinger equation (2.1) and execute a bootstrapping argument. Let $w_n \rightharpoonup^* w$ with w, $w_n \in \mathfrak{D}$ for all $n \in \mathbb{N}$. First, Proposition 2.5 and $S(0) = 0$ yield that the $u_n = S(w_n)$ are uniformly bounded in $L^2(D)$ for $k_0 > 0$ sufficiently small. Thus, there exists $\bar{u}$ such that $u_n \rightharpoonup \bar{u}$, where u_n is a subsequence that we again denote by u_n for ease of notation. Here, $\rightharpoonup$ indicates convergence in the weak topology. Furthermore, this implies the boundedness of the sequence $((1+qw_n)u_n)_n$, and we obtain $(1+qw_n)u_n \rightharpoonup v$ for some $v \in L^2(D)$ by again passing to a subsubsequence, which we again denote by the same symbol for ease of notation.

The compactness of K yields that

$$u_n \to K(v) + K(u_0 q w) =: \bar{u} \text{ in } H^2(D).$$

Thus, we conclude that $(1+qw_n)u_n \rightharpoonup (1+qw)\bar{u}$. Using the compactness of K, we obtain that

$$\bar{u} - K((1+q)w)\bar{u}) = K(u_0 qw).$$

By the uniqueness of the solution of (2.2) for $k_0 > 0$ sufficiently small (see [7, Thm 8.4]), we have $\bar{u} = S(w)$. Because the argument holds for all subsequences of the original sequences $(u_n)_n$ and $(w_n)_n$, the claim follows. □

Next, we state and prove our main theorem on the approximation relationship between (R) and (P) when they are constrained by the state equation (2.1) instead of the state equation (2.4).

Theorem 4.6. *Consider* (P) *and* (R) *being constrained by* (2.1) *instead of* (2.4)*. Let* $k_0 > 0$ *be sufficiently small. Then* (R) *admits a minimizer* $(u,w) \in H^2(D) \times \mathfrak{D}$*. Furthermore,*

$$\inf_{(u,w)\in\mathcal{F}_{(\mathrm{P})}} \frac{1}{2}\|u+u_0\|^2_{L^2(D_o)} = \min_{(u,w)\in\mathcal{F}_{(\mathrm{R})}} \frac{1}{2}\|u+u_0\|^2_{L^2(D_o)},$$

where $\mathcal{F}_{(*)}$ *is the feasible set of* $(*)$*.*

Proof. The objective of the optimization problem (R) is bounded from below. Furthermore, the set $\mathfrak{D}$ is nonempty, and (2.1) is uniquely solvable for all $\alpha \in \mathfrak{D}$ for $k_0 > 0$ sufficiently small. Therefore, (R) admits an infimum. The set $\mathfrak{D}$ is convex and weak*-closed in $L^\infty(D)$, which means that for all weakly* convergent sequences of functions in $\mathfrak{D}$, their limit is contained in $\mathfrak{D}$ as well. We consider a minimizing sequence of pairs $(u_n, w_n^R)_n \subset \mathcal{F}_{(\mathrm{R})}$, in particular $u_n = S(w_n^R)$. Because $(w_n^R)_n$ is bounded, it admits weak* accumulation points $\bar{w}^R$ as well as corresponding subsequences $w_k^R \rightharpoonup^* \bar{w}^R$ in $L^\infty(D)$. For $k_0 > 0$ sufficiently small, we employ Lemma 4.5 and obtain $u_k \to S(\bar{w}^R) =: \bar{u}$. Because the norm is continuous, we obtain that

$$\frac{1}{2}\|\bar{u}+u_0\|^2_{L^2(D_o)} = \min_{(u,w)\in\mathcal{F}_{(\mathrm{R})}} \frac{1}{2}\|u+u_0\|^2_{L^2(D_o)},$$

which establishes the first claim.

Now, let (u, w^R) be a minimizer of (R). From Proposition 4.4, we obtain that there exists a sequence $w_n^S \rightharpoonup^* w^R$ in $L^\infty(D)$ such that $(w_n^S, S(w_n^S))_n \subset \mathcal{F}_{(\mathrm{P})}$. Thus, Lemma 4.5 and continuity of the norm yield that

$$\frac{1}{2}\|S(w_n^S)+u_0\|^2_{L^2(D_o)} \to \frac{1}{2}\|\bar{u}+u_0\|^2_{L^2(D_o)} = \min_{(u,w)\in\mathcal{F}_{(\mathrm{R})}} \frac{1}{2}\|u+u_0\|^2_{L^2(D_o)}.$$

Thus, $(w_n^S, S(w_n^S))_n$ is a minimizing sequence for (P), which proves the second claim. □

Remark 4.7. *The sequence* $(w_n^S)_n$ *in the proof of Theorem 4.6 consists of discrete-valued controls that converge weakly* to a continuously valued control* w*. If they converge in the norm topology, the continuously valued control is also discrete-valued since the limits in the norm and weak* topologies coincide. If* (R) *admits only continuously valued minimizers, then* (P) *does not admit any minimizing control, but its infimum coincides with the one of* (R)*.*

4.7 Approximation Result for Robin Boundary Conditions

Before turning to the main statement, we again start with a preparatory lemma that establishes complete continuity of S.

Lemma 4.8. *For $k_0 > 0$ sufficiently small, the control-to-state operator $S : \mathfrak{D} \to H^1(D)$ is completely continuous in the sense that $w_n \rightharpoonup^* w$ yields $S(w_n) \to S(w)$.*

Proof. Proposition 2.6 establishes a continuous control-to-state operator $S : L^\infty(D) \to H^1(D)$, and Assumption 2.1 establishes the compact embedding $H^1(D) \hookrightarrow^c L^2(D)$.

From the proof of [5, Lem. 2.2], it follows that for $k_0 > 0$ sufficiently small,

$$\|S(w)\|_{H^1(D)} \le C\|w\|_{L^\infty(D)}$$

for some $C > 0$ and all $w \in \mathfrak{D}$. Thus, $w_n \rightharpoonup^* w$ yields $(S(w_n))_n$ being bounded in $H^1(D)$. From the compact embedding $H^1(D) \hookrightarrow^c L^2(D)$, we have that $S(w_k) \rightharpoonup \bar{u}$ in $H^1(D)$ and that $S(w_k) \to \bar{u}$ in $L^2(D)$ for a subsequence and some $\bar{u} \in H^1(D)$.

Consider the following weak formulation of the state equation (2.4):

$$(\nabla u, \nabla \phi)_{L^2(D)} - ik_0^2(u, \phi)_{L^2(\partial D)} - k_0^2((1 + qw)u, \phi)_{L^2(D)} = k_0^2(qwu_0, \phi)_{L^2(D)} \text{ for all } \phi \in H^1(D).$$

Inserting the subsequence $(w_k)_k$ from above, we immediately obtain that

$$\begin{aligned} (\nabla u_n, \nabla \phi)_{L^2(D)} &\to (\nabla \bar{u}, \nabla \phi)_{L^2(D)}, \\ k_0^2(qw_n u_0, \phi)_{L^2(D)} &\to k_0^2(qwu_0, \phi)_{L^2(D)}. \end{aligned}$$

Because the trace operator is linear and bounded (see Assumption 2.1), it is also weak-weak continuous; in other words, it maps weakly convergent sequences to weakly convergent sequences. We thus obtain

$$-ik_0^2(u_n, \phi)_{L^2(\partial D)} \to -ik_0^2(\bar{u}, \phi)_{L^2(\partial D)}.$$

Because the product of a norm-convergent and a weakly convergent sequence in $L^2(D)$ weakly converges to the product of the limits, it follows that

$$-k_0^2((1 + qw_n)u_n, \phi)_{L^2(D)} \to -k_0^2((1 + qw)\bar{u}, \phi)_{L^2(D)}.$$

These observations give the identity

$$(\nabla \bar{u}, \nabla \phi)_{L^2(D)} - ik_0^2(\bar{u}, \phi)_{L^2(\partial D)} - k_0^2((1 + qw)\bar{u}, \phi)_{L^2(D)} = k_0^2(q\alpha u_0, \phi)_{L^2(D)} \text{ for all } \phi \in H^1(D).$$

The weak solution of (2.4) is unique by Proposition 2.6. Thus, $\bar{u} = S(w)$, and by passing to subsubsequences the claim follows. □

Next, we state the main theorem on the approximation relationship between (R) and (P). The proof is analogous to the one of Theorem 4.6 and therefore is omitted.

Theorem 4.9. *Let* (P) *and* (R) *be constrained by* (2.1)*. Let $k_0 > 0$ sufficiently small. Then* (R) *admits a minimizer $(u, w) \in H^1(D) \times \mathfrak{D}$. Furthermore,*

$$\inf_{(u,w)\in\mathcal{F}_{(\mathrm{P})}} \frac{1}{2}\|u + u_0\|^2_{L^2(D_o)} = \min_{(u,w)\in\mathcal{F}_{(\mathrm{R})}} \frac{1}{2}\|u + u_0\|^2_{L^2(D_o)}.$$

■

4.8 Asymptotics of the Optimization Algorithm

We combine our considerations from the preceding subsections with the results in [17] to obtain the following convergence result for Algorithm 1. In particular, this theorem shows that for a sequence of refined rounding grids and a sequence of controls that converge to a (local) minimizer of the continuous relaxation (R), we obtain a sequence of controls feasible for the integer control problem (P) that converges weakly* to the (local) minimizer of the continuous relaxation and approximates the (locally optimal) objective value arbitrarily well.

Theorem 4.10. *Let the assumptions of Proposition 2.6 hold in the case of the state equation* (2.4)*, and let the assumptions of Proposition 2.5 hold in the case of the state equation* (2.1)*. Let the sequence of rounding grids $(\mathcal{T}^{(n)})_n$ be an order-conserving domain dissection. Then, for every weak* accumulation point $w^{R,*}$ of the iterates $(w_n^R)_n$ with approximating iterates $w_{n_k}^R \rightharpoonup^* w^{R,*}$ in $L^\infty(D)$ and corresponding consistent sequence of relaxed controls $\alpha_{n_k} \rightharpoonup^* \alpha^*$ with $\sum_{i=1}^M w_i(\alpha_{n_k})_i = w_{n_k}^R$ and $\sum_{i=1}^M w_i\alpha_i^* = w^{R,*}$ produced by Algorithm 1, we obtain that*

$$J(S(w_{n_k}^S)) \to J(S(w^{R,*})).$$

Proof. Our results from the preceding subsections and the assumption that the rounding grids constitute an order-conserving domain dissection enable us to apply Theorem 4.10 of [17] to our setting and the claim follows. □

Theorem 4.10 implies that if the iterates produced by Algorithm 1 or a subsequence of iterates converges to a local minimum of the relaxation, then we have convergence of the integer-valued approximations computed by means of SUR to the same local minimum in terms of the state vector and the objective value.

SUR provides an efficient way to compute weak* approximations of relaxed controls. The resulting binary control is not optimal with respect to the number of switches for the provable bound (see [14,23]) on the approximation. Optimizing with respect to switching cost while preserving the weak* approximation may require us to solve a different, potentially expensive, optimization problem; see [6]. However, applying a heuristic that consists of a median filtering reduces the switching and may not impact the resulting approximation too severely. Unfortunately, it is straightforward to construct a counterexample of a sequence of chattering functions with sum-up rounding that converges to a nonzero constant function such that the median at every grid cell evaluates to zero, thereby destroying the approximation property.

5 Computational Results

We consider two discrete realizations for the material constant w in (P), namely, $w_\ell = w_1 = 0$ and $w_u = w_2 = 1$. All our results use the Robin boundary setting (2.4). To solve discretizations of the relaxation of the MIPDECO for local minimizers, we rely on standard optimization techniques for bound-constrained optimization and employ a limited-memory quasi-Newton method with BFGS updates.

We have obtained our numerical results on a laptop computer with Intel Core i7-6820HQ CPU clocked at 2.70 GHz. To compute the results, we have employed the open-source libraries

FEniCS [2] for the discretization of the state equation, DOLFIN-ADJOINT [9] for the computation of the reduced objective and its derivative by means of adjoint calculus, and PETSc [4] for the optimization of the reduced objective functional with the PETSc Tao solver [20].

The finest discretization grid we consider for the approximation of the state as well as the control vector is a uniform decomposition of the square domain into $2^9 \times 2^9$ square cells, which consist of two triangles each. We have optimized the relaxed problem both with piecewise constant Ansatz functions for the control and with piecewise affine, globally continuous Ansatz functions for the state. The results are similar, and we present only the latter ones. For the integer-valued controls (i.e., the output of SUR) we have used piecewise constant (discontinuous Galerkin of order 0) Ansatz functions.

In contrast to [11], we use Robin boundary conditions in the computational experiments. Although physically more realistic approximations of the Sommerfeld radiation boundary condition exist (see [11, Sect. 5]), we have chosen to use the Robin boundary conditions for ease of the computational setup. This gives us consistency between the computational setup and our problem analysis and reduces unnecessary implementation overhead.

However, this comes at the cost that the results cannot be compared one to one with the results from [11], which is underlined by the following observation. We digitized the scatterer design plotted in [11, Fig. 7], solved our discretizations of (2.4) with this control input, and evaluated the objective. The resulting objective value we obtained for this discrete (binary) control input was 1.103×10^{-1} for a uniform triangular discretization of the domain with mesh size $h = 1.10 \times 10^{-2}$ and 1.085×10^{-1} with mesh size $h = 5.52 \times 10^{-3}$. This is significantly higher than the value 3.76×10^{-3} for mesh size $h = 6.67 \times 10^{-3}$ reported in [11]; and using this design as an initialization for the optimization of (R) yields a significantly improved design in our computational setup, which is not close to the relaxation in [11, Fig. 6]. We highlight that the main point of this paper is to show the effectiveness of SUR, and we do not believe that using more cumbersome boundary conditions to approximate the Sommerfeld radiation condition adds significantly to our observations.

5.1 State Vector and Objective Approximation

We set up several experiments to validate our theoretical results. In all of them, we choose the parameter values $q = 0.75$ and $k_0 = 6\pi$, which are the settings from [11,27]. Furthermore, we chose the angle of the incident wave as $\pi/2$ as in [11].

We perform nine iterations of Algorithm 1; that is, we compute a sequence of (locally optimal) controls w_n^R by solving $(\mathrm{R}_h^{(n)})$ for each grid $n = 1, \dots, 9$, which requires about 12 hours on our laptop. We compute the corresponding relaxed controls α_n in a consistent manner, apply the SUR algorithm on the nth grid to compute ω_n from α_n, and reconstruct the discrete-valued control w_n^S from ω_n as described in Section 4.3. We assess the convergence of the corresponding state vectors by inserting w_n^R and w_n^S into the numerical approximation of the control-to-state operator S on the nth grid. Furthermore, we project the w_n^R onto the finest triangulation and show self-convergence of the state vectors $\|S(w_n^R) - S(w_9^R)\|_{L^2}$. We use the same projections to evaluate $J(S(w_n^R))$ on the finest grid in order to be able to compare the values. The mesh size (grid cell diameter) is $h^n = 2\sqrt{2}\frac{1}{2^n}$ for the cells both of the rounding grid and of the finite-element discretization.

The results are summarized in Table 1 and validate Theorem 4.10. In particular, we observe that the difference between the state vector corresponding to the relaxed control and the binary control

Table 1: Convergence of state vector and objective for refined grids on which both w_n^R and w_n^S are computed, $\|\cdot\| = \|\cdot\|_{L^2}$.

n	# binary vars	h^n	$\|S(w_n^S) - S(w_n^R)\|$	$\|S(w_n^R) - S(w_9^R)\|$	$J(S(w_n^R))$
3	32	3.54×10^{-1}	1.752×10^{-1}	1.795×10^{0}	2.263×10^{-1}
4	128	0.18×10^{-1}	3.062×10^{-1}	1.939×10^{0}	1.991×10^{-1}
5	504	8.84×10^{-2}	2.400×10^{-1}	2.022×10^{0}	1.624×10^{-1}
6	2008	4.42×10^{-2}	2.151×10^{-1}	9.538×10^{-1}	2.188×10^{-2}
7	8028	2.21×10^{-2}	5.386×10^{-2}	2.136×10^{-1}	1.999×10^{-3}
8	32112	1.10×10^{-2}	2.096×10^{-2}	3.947×10^{-2}	9.852×10^{-5}
9	128452	5.52×10^{-3}	4.125×10^{-3}	0	3.533×10^{-6}

produced by SUR converges to zero. Furthermore, we observe self-convergence of the relaxed states $S(w_n^R)$. The objective $J(S(w_n^R))$ of the objective also appears to converge to zero, which suggests that cloaks may exist such that the electromagnetic field vanishes entirely in D_o. The real and imaginary parts of u_0, $S(w_9^R)$, $S(w_9^S)$ and $S(w_9^F)$ are plotted in Figure 2. The qualitative impression of the achieved cloaking is similar to the results in [11].

As a second experiment, we take the (locally optimal) control $w^R := w_9^R$ and its corresponding relaxed control α from the previously described experiment. Then, we execute SUR on the same refined rounding grids to obtain a sequence of discrete-valued controls $(w_n^S)_n$ until the rounding grid matches the grid of the relaxation. Furthermore, for every rounding grid, we compute a median filtering w_n^F of w_n^S as a postprocessing step after SUR to remove jittering in w_n^S and improve the implementability. We assess the convergence of the corresponding state vectors and the corresponding objectives by inserting the w_n^S and w_n^F for $n = 1, \ldots, 8$ into the numerical approximation of S and J on the finest grid $n = 9$. The results are stated in Table 2, which shows the relative errors. The relative error between the solution for the relaxed control $S(w^R)$ and the solutions for the binary controls $S(w_n^S)$ converges to zero. The relative error of the objective is not convincing at first because the convergence ends at a value of 10^{-1}. But the objective value of the relaxation is very small, and in absolute values we note that $J(S(w_9^S)) = 3.933 \times 10^{-6}$ and thus the absolute gap in the objective between the integer and the relaxed control problem is only $J(S(w_9^S)) - J(S(w_9^R)) = 4.000 \times 10^{-7}$. Similarly, we have an absolute objective value $J(S(w_9^F)) = 3.671 \times 10^{-3}$ for the filtered control.

The weak* convergence of the controls as well as the norm convergence of the state vector can be observed, and in Figure 3, we plot the iterates $(w_n^S)_n$, $(S(w_n^S))_n$ as well as the limits w^R and $S(w^R)$. The density of the refined binary structure of the scatterers w_S^n converges to the relaxation w^R as expected by Proposition 4.4. The convergence of the corresponding scattered electromagnetic fields is clearly visible. The amplitude fields $|S(\omega_6^S)|$,...,$|S(\omega_9^S)|$ exhibit hardly any visual difference from $S(w^R)$.

The effect of the filtering on the control can be observed in Figure 4. A lot of the scattered switching between material and no material placement is removed in the filtered control inputs in the bottom row. The objective is considerably higher; but with $J(w_9^F) = 3.671 \times 10^{-3}$, it is still

Table 2: Convergence of state and objective for w^R with $u^R = S(w^R)$ computed on a fine grid and refined rounding grids for the execution of SUR (w_n^S) and SUR with median-filter post-processing step (w_n^F).

n	h^n	$\frac{\|S(w_n^S)-u^R\|_{L^2}}{\|u^R\|_{L^2}}$	$\frac{J(S(w_n^S))-J(u^R)}{J(u^R)}$	$\frac{\|S(w_n^F)-u^R\|_{L^2}}{\|u^R\|_{L^2}}$	$\frac{J(S(w_n^F))-J(u^R)}{J(u^R)}$
1	1.41×10^{0}	9.646×10^{-1}	6.177×10^{4}	7.028×10^{-1}	5.021×10^{4}
2	7.07×10^{-1}	9.332×10^{-1}	5.299×10^{4}	7.028×10^{-1}	5.021×10^{4}
3	3.54×10^{-1}	9.843×10^{-1}	4.131×10^{4}	6.548×10^{-1}	4.813×10^{4}
4	1.77×10^{-1}	6.507×10^{-1}	2.364×10^{4}	5.649×10^{-1}	4.607×10^{4}
5	8.84×10^{-2}	3.646×10^{-1}	7.888×10^{3}	4.188×10^{-1}	1.754×10^{4}
6	4.42×10^{-2}	1.173×10^{-1}	7.719×10^{2}	1.339×10^{-1}	1.685×10^{3}
7	2.21×10^{-2}	3.042×10^{-2}	1.767×10^{1}	1.046×10^{-1}	1.036×10^{3}
8	1.10×10^{-2}	8.843×10^{-3}	1.898×10^{0}	1.020×10^{-1}	1.097×10^{3}
9	5.52×10^{-3}	2.071×10^{-3}	1.132×10^{-1}	1.000×10^{-1}	1.038×10^{3}

small, and the electromagnetic field is still damped heavily in the region D_o (see Figure 2).

5.2 Effect of Increased the Wave Number

The proofs in Section 4 are valid for the wave number $k_0 > 0$ sufficiently small. Thus, we consider an experiment that evaluates the approximation quality of the state vector for an increasing wave number k_0. We consider the finest grid of our computations and vary the wave number from 0.24π to 750π by scaling it by five from one run to the next.

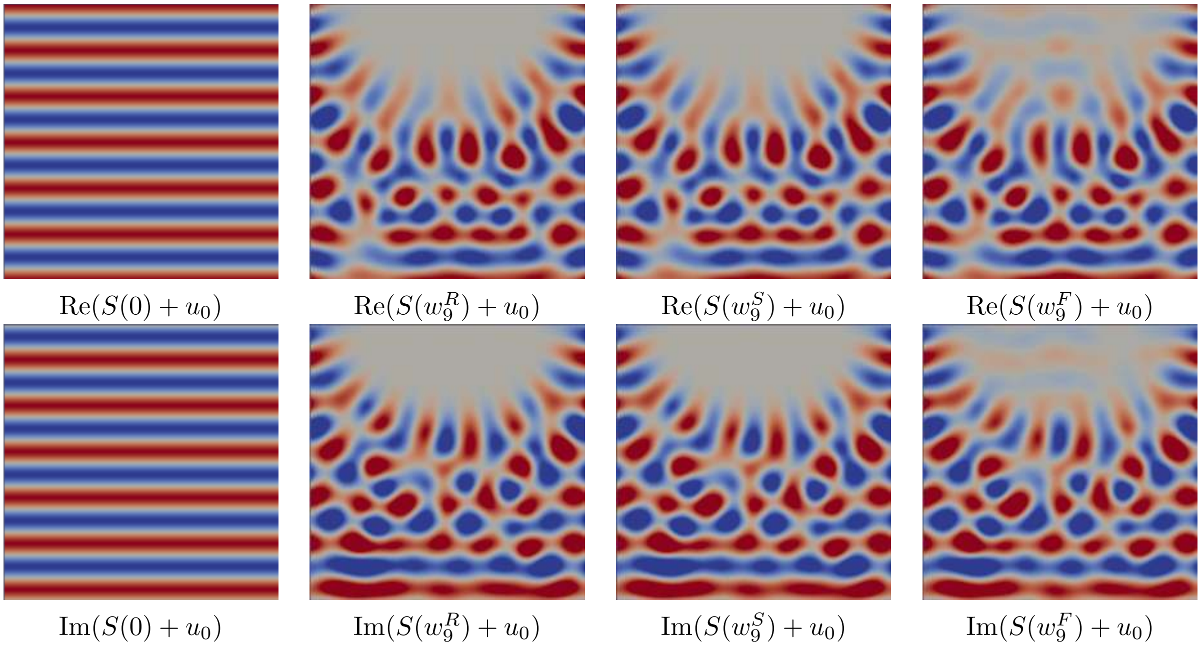
In every run (i.e., for every wave number k_0), we do the following. Similar to the second experiment of Section 5.1, we use the same fixed fractional-valued w^R on the finest grid and refine the rounding grid until it matches the finest grid to compute the approximating controls $(w_n^S)_n$ by means of SUR. We evaluate the state vectors $S(w^R)$ and $(S(w_n^S))_n$ and compute the relative errors $\|S(w^R) - S(w_n^S)\|_{L^2}/\|S(w^R)\|_{L^2}$. The convergence appears to be linear in the volume of the square grid cells of the refined rounding grids, which can be seen by comparing the (scaled) volume of the grid cells indicated by the three dashed lines in Fig. 5 with the approximation errors.

Over the 9 iterations, we observe convergence only for the small wave numbers $k_0 = 0.24\pi$, 1.2π, 6π. We potentially see the start of convergence for the wave number $k_0 = 30\pi$ and no convergence for the wave numbers $k_0 = 150\pi$ and $k_0 = 750\pi$. The corresponding plots are given in Figure 5.

6 Conclusion

We have successfully applied the SUR algorithm and the underlying approximation methodology to approximate the local infimum of the electromagnetic cloaking problem by taking the point of view of a nonconvex MIPDECO. The computational results confirm our theoretical results.

When implementing the design of the scatterer given by the final iterate of Algorithm 1, one should use improved approximations of the boundary conditions as suggested in [11] or extend the computational domain in all directions so that the areas D_o and D_s are only minimally affected


$\mathrm{Re}(S(0)+u_0)$ $\mathrm{Re}(S(w_9^R)+u_0)$ $\mathrm{Re}(S(w_9^S)+u_0)$ $\mathrm{Re}(S(w_9^F)+u_0)$

$\mathrm{Im}(S(0)+u_0)$ $\mathrm{Im}(S(w_9^R)+u_0)$ $\mathrm{Im}(S(w_9^S)+u_0)$ $\mathrm{Im}(S(w_9^F)+u_0)$

Figure 2: Real (top) and imaginary (bottom) part of the unscattered wave ($\omega \equiv 0$), the incident wave u_0, and the optimally scattered wave $S(w_9^R)$ for the relaxation and the scattered waves due to SUR and the median filtering $S(w_9^S)$ and $S(w_9^F)$, respectively.

by approximation error at the corners. Because doing so does not change the number of controls that are optimized, we assume that the additional computational costs do not grow too large in this case.

Acknowledgments

The authors thank Dirk Lorenz, TU Braunschweig, for pointing at median filterings for reduction of chattering.

References

[1] A. A. Ali, K. Deckelnick, and M. Hinze. Error analysis for global minima of semilinear optimal control problems. *arXiv preprint arXiv:1705.01201*, 2017.

[2] M. S. Alnæs, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. E. Rognes, and G. N. Wells. The FEniCS Project Version 1.5. *Archive of Numerical Software*, 3(100), 2015. doi:10.11588/ans.2015.100.20553.

[3] N. Arada, E. Casas, and F. Tröltzsch. Error estimates for the numerical approximation of a semilinear elliptic control problem. *Computational Optimization and Applications*, 23(2):201–229, Nov 2002.

[4] S. Balay, S. Abhyankar, M. F. Adams, J. B., P. Brune, K. Buschelman, L. Dalcin, V. Eijkhout, D. Kaushik, M. G. Knepley, L. Curfman McInnes, W. D. Gropp, K. Rupp, B. F. Smith, S. Zampini, and H. Zhang. PETSc Users Manual. Technical Report ANL-95/11 - Revision 3.7, Argonne National Laboratory, 2016.

[5] G. Bao and P. Li. Inverse medium scattering for the Helmholtz equation at fixed frequency. *Inverse Problems*, 21(5):1621, 2005.

[6] F. Bestehorn, C. Hansknecht, C. Kirches, and P. Manns. A switching cost aware rounding method for relaxations of mixed-integer optimal control problems. In *Proceedings of the IEEE Conference on Decision and Control*, 2019. To appear.

[7] D. Colton and R. Kress. *Inverse Acoustic and Electromagnetic Scattering Theory.* Springer, 2013.

[8] B. Engquist and A. Majda. Radiation boundary conditions for acoustic and elastic wave calculations. *Communications on Pure and Applied Mathematics*, 32(3):313–357, 1979.

[9] P. E. Farrell, D. A. Ham, S. W. Funke, and M. E. Rognes. Automated derivation of the adjoint of high-level transient finite element programs. *SIAM Journal on Scientific Computing*, 35(4):C369–C393, 2013.

[10] F. M. Hante and S. Sager. Relaxation methods for mixed-integer optimal control of partial differential equations. *Computational Optimization and Applications*, 55(1):197–225, 2013. doi:10.1007/s10589-012-9518-3.

[11] J. Haslinger and R. A. E. Mäkinen. On a topology optimization problem governed by two-dimensional Helmholtz equation. *Computational Optimization and Applications*, 62(2):517–544, 2015. doi:10.1007/s10589-015-9746-4.

[12] M. Hinze. A variational discretization concept in control constrained optimization: the linear-quadratic case. *Computational Optimization and Applications*, 30(1):45–61, 2005. doi:10.1007/s10589-005-4559-5.

[13] K. Ito and K. Kunisch. Optimal bilinear control of an abstract Schrödinger equation. *SIAM Journal on Control and Optimization*, 46(1):274–287, 2007.

[14] C. Kirches, F. Lenders, and P. Manns. Approximation properties and tight bounds for constrained mixed-integer optimal control. *Optimization Online Preprint*, (5404), 2019. submitted.

[15] Kristina. Kohls, Arnd. Rösch, and Kunibert G. Siebert. A posteriori error analysis of optimal control problems with control constraints. *SIAM Journal on Control and Optimization*, 52(3):1832–1861, 2014.

[16] A. Kröner and B. Vexler. A priori error estimates for elliptic optimal control problems with a bilinear state equation. *Journal of Computational and Applied Mathematics*, 230(2):781 – 802, 2009.

[17] P. Manns and C. Kirches. Multi-dimensional sum-up rounding for elliptic control systems. *Submitted*, 2018. `https://spp1962.wias-berlin.de/preprints/080r.pdf`.

[18] P. Manns and C. Kirches. Improved regularity assumptions for partial outer convexification of mixed-integer PDE-constrained optimization problems. *ESAIM: Control, Optimisation and Calculus of Variations*, 2019. To appear, doi:10.1051/cocv/2019016.

[19] C. Meyer and A. Rösch. Superconvergence properties of optimal control problems. *SIAM Journal on Control and Optimization*, 43(3):970–985, 2004.

[20] T. Munson, J. Sarich, S. Wild, S. Benson, and L. C. McInnes. TAO 3.5 Users Manual. Technical Report ANL/MCS-TM-322, Argonne National Laboratory, Mathematics and Computer Science Division, 2015.

[21] M. Renardy and R. C. Rogers. *An Introduction to Partial Differential Equations*, volume 13. Springer Science & Business Media, 2006. doi:10.1007/b97427.

[22] S. Sager. *Numerical methods for mixed-integer optimal control problems.* Der andere Verlag, Tönning, Lübeck, Marburg, 2005.

[23] S. Sager, H. G. Bock, and M. Diehl. The integer approximation error in mixed-integer optimal control. *Mathematical Programming, Series A*, 133(1–2):1–23, 2012. doi:10.1007/s10107-010-0405-3.

[24] S. Sager, M. Jung, and C. Kirches. Combinatorial integral approximation. *Mathematical Methods of Operations Research*, 73(3):363–380, 2011. doi:10.1007/s00186-011-0355-4.

[25] D. Schurig, J. J. Mock, B. J. Justice, S. A. Cummer, J. B. Pendry, A. F. Starr, and D. R. Smith. Metamaterial electromagnetic cloak at microwave frequencies. *Science*, 314(5801):977–980, 2006.

[26] F. Tröltzsch. *Optimal Control of Partial Differential Equations.* American Mathematical Society, 2010.

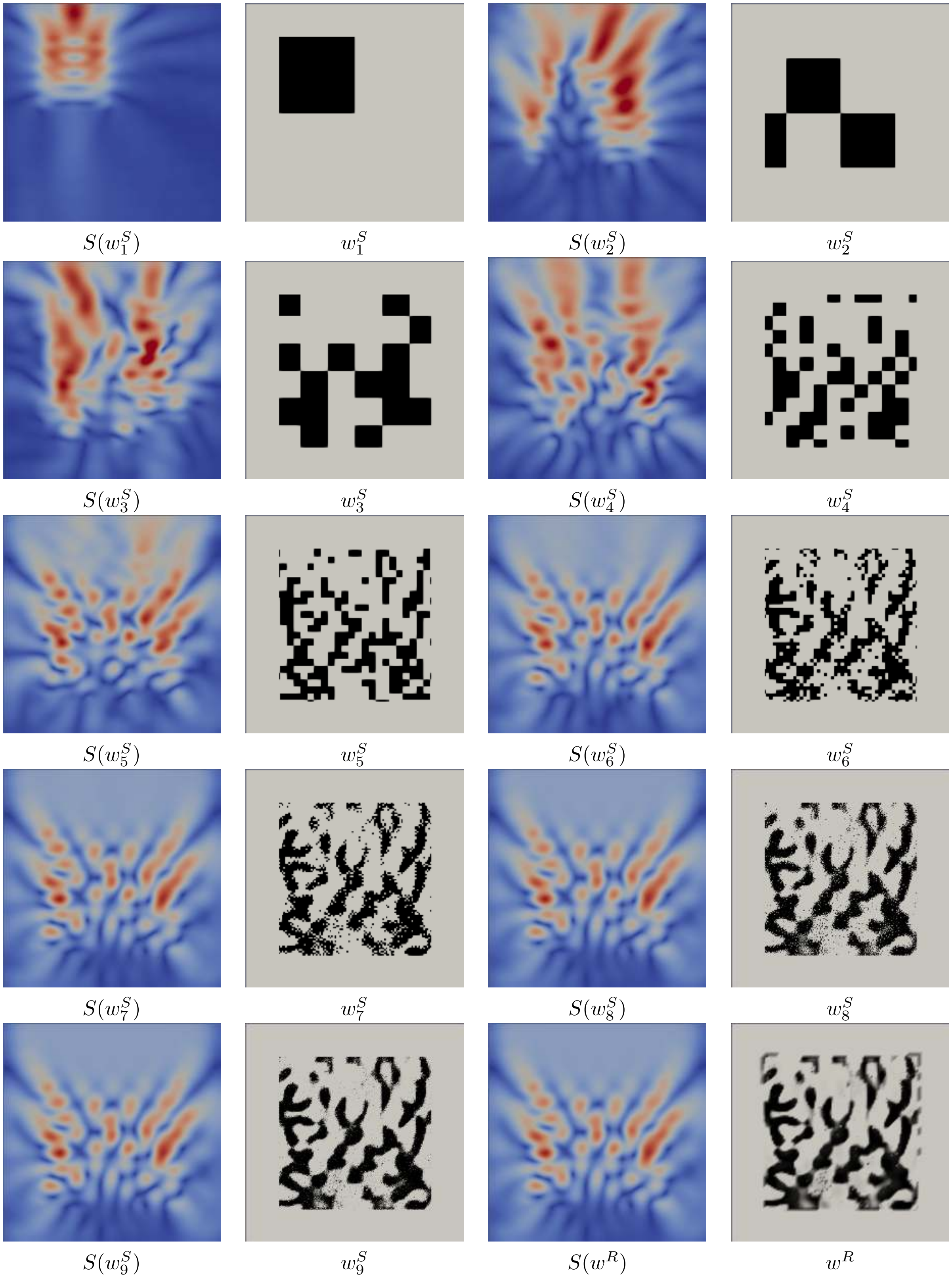
[27] R. H. Vogt, S. Leyffer, and T. Munson. A mixed-integer PDE-constrained formulation for electromagnetic cloaking. *Submitted*, 2019.

[28] Xu, Yifeng and Zou, Jun. A convergent adaptive edge element method for an optimal control problem in magnetostatics. *ESAIM: M2AN*, 51(2):615–640, 2017.

[29] I. Yousept. Finite element analysis of an optimal control problem in the coefficients of time-harmonic eddy current equations. *Journal of Optimization Theory and Applications*, 154(3):879–903, Sept 2012.

[30] I. Yousept. Optimal bilinear control of eddy current equations with grad-div regularization. *Journal of Numerical Mathematics*, 23(1):81–98, 2015. doi:10.1515/jnma-2015-0007.

[31] J. Yu and M. Anitescu. Multidimensional sum-up rounding for integer programming in optimal experimental design. *Mathematical Programming*, pages 1–40, 2019.



$S(w_1^S)$ w_1^S $S(w_2^S)$ w_2^S

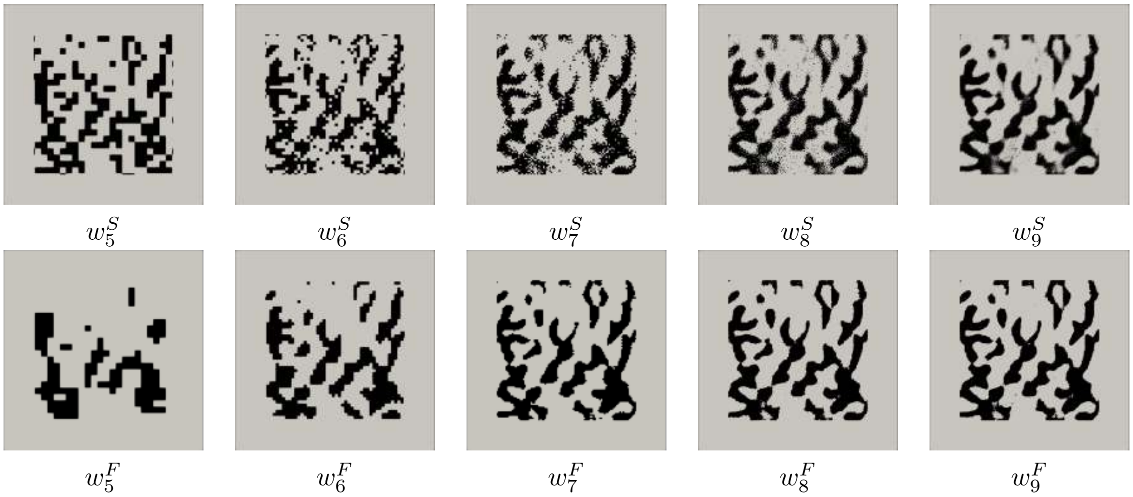
$S(w_3^S)$ w_3^S $S(w_4^S)$ w_4^S

$S(w_5^S)$ w_5^S $S(w_6^S)$ w_6^S

$S(w_7^S)$ w_7^S $S(w_8^S)$ w_8^S

$S(w_9^S)$ w_9^S $S(w^R)$ w^R

Figure 3: State (magnitude plotted) and weak* control vector convergence.



w_5^S w_6^S w_7^S w_8^S w_9^S

w_5^F w_6^F w_7^F w_8^F w_9^F

Figure 4: Control vector iterates (top) and their filterings (bottom).

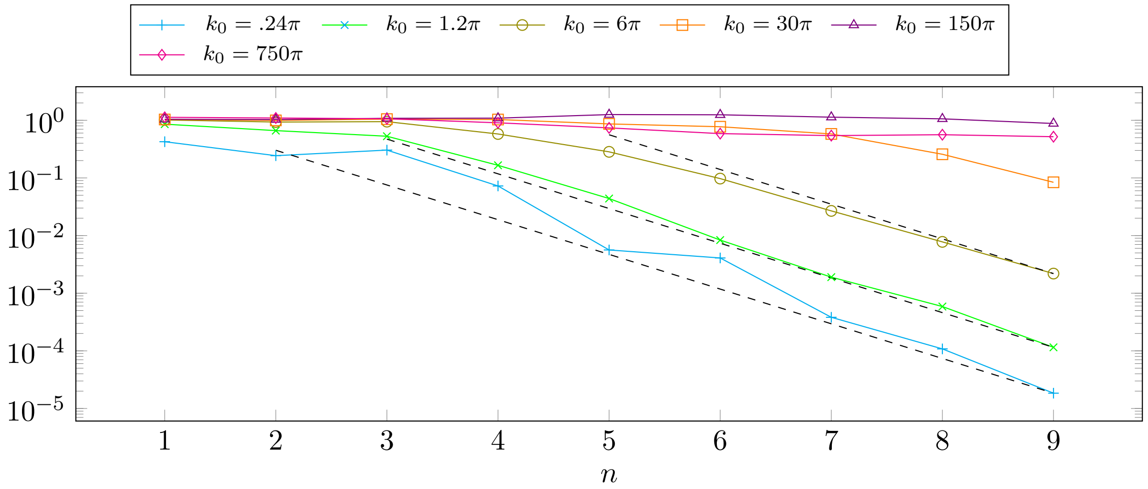


Figure 5: Sequences of the relative state vector approximation errors $\|S(w_n^S) - S(w^R)\|_{L^2}/\|S(w^R)\|_{L^2}$ of SUR approximations w_n^S for a fixed w^R and refined rounding grids indexed by $n = 1, \dots, 9$ for different choices of the wave number k_0.