State Elimination for Mixed-Integer Optimal Control of PDEs by Semigroup Theory*

Anna Thünen[†], Sven Leyffer[‡], and Sebastian Sager[§]

Aachen, May 17, 2021

Abstract

Mixed-integer optimal control problems governed by PDEs (MIPDECOs) are powerful modeling tools but also challenging in terms of theory and computation. We propose a highly efficient state elimination approach for MIPDECOs that are governed by PDEs that have the structure of an abstract ODE in function space. This allows us to avoid repeated calculations of the states for all time steps, and our approach is applied only once before starting the optimization. The presentation of theoretical results is complemented by numerical experiments.

1 Introduction

Mixed-integer partial differential equation constrained optimization (MIPDECO) is highly important because it has a broad range of applications, such as topology optimization (see, e.g., [12]) and PDEs on networks with discrete decisions, in particular traffic flow with traffic lights [8], the operation of transmission lines [6], and gas networks [9]. MIPDECO combines two key classes of optimization: mixed-integer nonlinear optimization (MINLP) and partial differential equation (PDE) constrained optimization. In MINLP the feasible set and the objective function are quantified by nonlinear functions. In addition to real-valued decision variables, it includes integer variables leading to combinatorial difficulties; see, for example, [1]. On the other hand, many applications in optimization involve complex systems modeled by differential equations. PDE-constrained optimization poses different challenges since discretizations lead to a large number of variables and numerical complexity (see [13, 22]). Each of the problem classes presents big challenges by itself. With MIPDECO combining the two classes, approaches are needed that overcome the issues of *both* integer variables and the number of variables. This work

Techniques for MIPDECO are often derived from methods that have been proven efficient for mixed-integer control of ordinary differential equations (ODEs). Pioneering work on partial outer convexification, presented in [19], provides strong theoretical results. Extensions to problems governed by PDEs include [11] for parabolic and [10] for hyperbolic PDEs. The usually strong regularity assumptions are weakened in [17]. Also, extensions of rounding approaches to spatial distributed controls have been investigated, for example, in [16].

[*] Argonne National Laboratory MCS Division Preprint ANL/MCS-P9481-0521

[†] Anna Thünen is with the Institute of Geometry and Practical Mathematics, RWTH Aachen University, 52062 Aachen, Germany thuenen@igpm.rwth-aachen.de

[‡] Sven Leyffer is with the Mathematics and Computer Science Division at Argonne National Laboratory, Lemont, IL 60439, USA leyffer@anl.gov

[§] Sebastian Sager is with the Faculty of Mathematics, Otto-von-Guericke University, 30106 Magedeburg, Germany sager@ovgu.de

Combinatorical constraints coupling over time can be handled by combinatorial integer approximation problems; see, for example, [20]. Extensions to this approach include reductions of unrealistic frequent switching [21], by constrained total variation of the control [23], and by minimum dwell time constraints. Implementations include the open-source software package pycombina [2].

In applications, tailored branch-and-bound algorithms have been applied [7]. Recently, in [5] a penalty method was proposed that relies on a combination of tailored basin hopping and interior-point method.

In this paper, we reduce the complexity of the MIPDECO by exploiting the structure of the PDE using semigroup theory. This approach allows us to split the state solution into parts that are then combined with a smart use of convolution. The resulting explicit control-to-state-map is plugged into the objective and additional constraints and replaces the PDE, which no longer appears in the problem formulation. Computationally this approach is performed before handing the optimization problem, which then can be solved with much less computational effort. We illustrate this approach by applying it to the time-dependent 2D heat equation.

1.1 Problem Formulation

The MIPDECO problem studied here is formulated as a minimization of an objective functional $J:\ \mathcal{U}_{\text{ad}} \times \mathcal{V}_{\text{ad}} \times \mathcal{W}_{\text{ad}} \to \mathbb{R}$, where $\mathcal{U}_{\text{ad}} \subset \mathcal{U}$ is the state space , $\mathcal{V}_{\text{ad}} \subset \mathcal{V}$ is the space of admissible real valued controls, and $\mathcal{W}_{\text{ad}} \subset \mathcal{W}$ is the space of admissible discrete controls for which we assume that they take values of a finite set $\hat{W} = \{\hat{w}^1, \dots, \hat{w}^{|\hat{W}|}\}$.

The problem is constrained by a special type of PDE that may be formulated as an operator differential equation where $A:\ \mathcal{D}(A) \to \mathcal{U}$ is the infinitesimal generator of a strongly continuous semigroup $\{T(t)\}_{t\geq 0}$ on $\mathcal{U}$.

With that, we can write the MIPDECO problem as follows:

$$\begin{aligned}
\min_{u,v,w} \quad & J(u,v,w) = \phi(u(t_f)) + \int_0^{t_f} \psi(u(t), v(t), w(t))\,\mathrm{d}t \\
\text{s.t.} \quad & \dot{u}(t) = Au(t) + F(t, u(t), v(t), w(t)), \quad t \in [0, t_f] \\
& u(0) = u_0 \in \mathcal{D}(A) \\
& u \in \mathcal{U}_{\text{ad}}, v \in \mathcal{V}_{\text{ad}}, w \in \mathcal{W}_{\text{ad}},
\end{aligned} \tag{1}$$

where $t_f > 0$ is the final time and with the sufficiently regular functions $\phi:\ \mathcal{U} \to \mathbb{R}$ and $\psi:\ \mathcal{U} \times \mathcal{V} \times \mathcal{W} \to \mathbb{R}$ specifying the objective and the source functional $F:\ T \times \mathcal{U} \times \mathcal{V} \times \mathcal{W} \to \mathcal{U}$. Furthermore, we assume that $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}$ are normed linear spaces.

The theory of the existence of solutions of the dynamic problem and of the control problem depends on the operator A and associated spaces. Because we are considering mild solutions, we require that $\mathcal{U}$, $\mathcal{V}$, and $\mathcal{W}$ be Banach spaces and that at least $F \in L^1(T \times \mathcal{U} \times \mathcal{V} \times \mathcal{W}; \mathcal{U})$.

1.2 Outline

We begin with some preliminaries in Section 2, where we cover basic concepts of semigruop theory for PDEs and define convolution. In Section 3 we propose a reduction method that eliminates the PDE in the MIPDECO (1) for continuous and discrete time. Then, in Sec. 4 we apply this theory to a MIPDECO problem governed by the heat equation. In Section 5 we derive a discrete representation of the heat equation, which also illustrates the low computational effort of the presented elimination scheme. In Section 6 we conclude with numerical experiments.

2 Preliminaries: Semigroups of Linear Operators and Convolution

We start by reviewing the required concepts from semigroup theory, which allow us to view the PDE as an abstract ODE in function space. Thereafter, we provide definitions of convolution.

2.1 Uniformly Continuous Semigroups of Bounded Linear Operators

We begin with the definition of a semigroup and its generators as in [18, Chapter 1].

Definition 2.1 (Semigroup of bounded linear operators)**.** *Let $\mathcal{U}$ be a Banach space. A one-parameter family $T(t)$, $0 \leq t < \infty$, of bounded linear operators from $\mathcal{U}$ into $\mathcal{U}$ is a semigroup of bounded linear operators on $\mathcal{U}$ if*

(i) $T(0) = I$, (I is the identity operator on $\mathcal{U}$).

(ii) $T(t+s) = T(t)T(s)$ for every $t, s \geq 0$ (the semigroup property).

A semigroup of bounded linear operators, $T(t)$, is uniformly continuous if

$$\lim_{t\to 0} \|T(t) - I\| = 0.$$

The linear operator A defined by

$$\mathcal{D}(A) = \left\{ u \in \mathcal{U} \,\middle|\, \lim_{t\to 0} \frac{T(t)u - u}{t} \text{ exists} \right\}$$

and

$$Au = \lim_{t\to 0} \frac{T(t)u - u}{t} = \left.\frac{\mathrm{d}^+ T(t)u}{\mathrm{d}t}\right|_{t=0} \quad \text{for } u \in \mathcal{D}(A)$$

is the infinitesimal generator of the semigroup $T(t)$, and $\mathcal{D}(A)$ is the domain of A.

We are interested in PDEs that can be formulated as an abstract ODE in the form of the following initial-value problem:

$$\dot{u}(t) = Au(t) + f(t), \; u(0) = u_0. \tag{2}$$

As in classical PDE theory, semigroup theory has multiple solution concepts. Here, we consider mild solutions only.

Definition 2.2 (Mild solution [18, Definition 2.3])**.** *Assume that $u_0 \in \mathcal{D}(A)$ and $f \in L^1([0, t_f]; \mathcal{U})$. Then a solution $u(t)$ of* (2) *is given by*

$$u(t) = T(t)u_0 + \int_0^t T(t-s)f(s)\,\mathrm{d}s, \tag{3}$$

and $u(t)$ is called a mild solution of (2)*.*

Provided a continuous differentiable source term f, mild solutions to (2) exist for any initial value $u_0 \in \mathcal{D}(A)$; see [18, Corollary 2.5].

We conclude this brief introduction by drawing the connection of mild solutions to solutions in the classical sense [18, Theorem 3.2].

Theorem 2.1 (Mild solutions are classical solutions)**.** *Let A be a infinitesimal generator of a semigroup $T(t)$. Let $f \in L^1([0,t_f];\mathcal{U})$, and assume that for every $0 < t < T_f$ there is a $\delta_t > 0$ and a continuous real-valued function $W_t, \mathbb{R}_+ \to \mathbb{R}_+$, such that*

$$\|f(t) - f(s)\| \leq W_t(|t-s|),$$

and

$$\int\limits_0^{\delta_t} \frac{W_t(\tau)}{\tau} \,\mathrm{d}\tau < \infty.$$

Then for every $u_0 \in \mathcal{U}$ the mild solution of (2) *is a classical solution.*

Note that choosing $W_t(\tau) = c\tau$ with $c > 0$ yields the theorem for Lipschitz continuous f.

Remark 2.1. *This theory of semigroups of bounded linear operators allows us to write solutions explicitly in terms of $T(t)$. However, since $T(t)$ is not available explicitly in general, the theory does not provide explicit solutions. Still, we make use of the structure of mild solutions* (3) *in Section 3.*

2.2 Convolution

We define the term convolution.

Definition 2.3. *Let $h_1, h_2 : \mathbb{R} \to \mathbb{R}$ be integrable functions. Then their convolution is given by*

$$(h_1 * h_2)(t) = \int\limits_{\mathbb{R}} h_1(s) h_2(t-s) \,\mathrm{d}s.$$

If the functions are defined on a subset of $\mathbb{R}$, the functions are extended by zero.

Continuous convolution can be transferred to discrete domains. Therefore it is applicable for the discrete counterparts emerging in the MINLP formulation of a MIPDECO problem.

Definition 2.4. *Let $h_1, h_2 : \mathbb{Z} \to \mathbb{R}$ be sequences. Then their convolution is defined by*

$$(h_1 * h_2)[n] = \sum_{m=-\infty}^{\infty} h_1[m] h_2[n-m].$$

If the sequences are defined on a subset of $\mathbb{Z}$, the sequences are extended with zero.

3 Elimination of the PDE

In this section we present the derivation of the technique that allows us to reduce significantly the computational effort of the MIPDECO. Together with structural assumptions, the main result is stated and proved. Subsequently, this result is transferred to the discretized-in-time problem.

3.1 Explicit Representation of the Solution Operator

Now we return to the model in (1). We assume that A is a linear operator. Furthermore, we assume that the source term $F(t, u(t), v(t), w(t))$ depends linearly on the controls v and w only; in other words, there is no explicit dependence on t or on the state u:

$$F(t, u(t), v(t), w(t)) = F(v(t), w(t)).$$

Further, we assume that

$$\hat{W} = \left\{ w(t) \in \{0,1\}^L \,\middle|\, \sum_{l=1}^{L} w_l(t) = \bar{w} \right\}$$

and that the other control v takes values of the same dimension, namely, $v : \; [0, t_f] \to \mathbb{R}^L$. Therefore, the following representation of the source term exists:

$$F(v(t), w(t)) = \sum_{l=1}^{L} w_l(t) v_l(t) \tilde{f}_l, \tag{4}$$

where $\tilde{f}_l \in \mathcal{U}$ is constant for $l = 1, \ldots, L$ and we can write (1) as

$$\begin{aligned} \min_{u,v,\alpha} \quad & J(u, v, w) = \phi(u(t_f)) + \int\limits_0^{t_f} \psi(u(t), v(t), w(t)) \; \mathrm{d}t \\ \text{s.t.} \quad & \dot{u}(t) = Au(t) + \sum_{l=1}^{L} w_l(t) v_l(t) \tilde{f}_l \quad t \in [0, t_f] \\ & u(0) = u_0 \in \mathcal{D}(A) \\ & u \in \mathcal{U}_{\text{ad}}, v \in \mathcal{V}_{\text{ad}}, w \in \mathcal{W}_{\text{ad}}, \end{aligned} \tag{5}$$

where $\mathcal{W}_{\text{ad}} = \{w \in L^2([0, t_f]; \mathbb{R}) | w(t) \in \hat{W}\}$.

Before exploiting the structure with the aid of semigroups in (5), we summarize the required assumptions.

Assumption 3.1. *We consider controlled dynamics that have representations as abstract ODEs of the form*

$$\dot{u}(t) = Au(t) + \sum_{l=1}^{L} w_l(t) v_l(t) \tilde{f}_l, t \in [0, t_f], \quad u(0) = u_0, \tag{6}$$

which fulfill the following properties:

- $A : \mathcal{D}(A) \to \mathcal{U}$ *linear, infinitesimal generator of a strongly continuous semigroup* $T(t)$
- $\tilde{f}_l \in \mathcal{U}$ *constant for* $l = 1, \ldots, L$
- $v, w \in L^2([0, t_f]; \mathbb{R}^L)$
- $u_0 \in \mathcal{D}(A)$

Note that the assumption of $\tilde{f}_l$ being constant is in the sense of the space $\mathcal{U}$; that is, it is constant in time, but it may vary if spatial coordinates are considered in $\mathcal{U}$. An example is given in Section 4.

Lemma 3.1 (Control-to-state map)**.** *Let Assumption 3.1 hold. Then the solution of the dynamical system* (6) *is given in terms of the controls* $(v(t), w(t))$*:*

$$u(t) = T(t)u_0 + \sum_{l=1}^{L} ((T\tilde{f}_l) * (w_l v_l))(t). \tag{7}$$

Proof. Since A is a linear operator, the solution $u(t)$ to the dynamics can be split as follows:

$$u(t) = \bar{u}^{\mathrm{h}}(t) + \bar{u}^{\mathrm{inh}}(t),$$

with $\bar{u}^{\mathrm{h}}(t) = T(t)u_0$ the homogeneous part and $\bar{u}^{\mathrm{inh}}$ the inhomogeneous part of the solution. Therefore, the function $\bar{u}^{\mathrm{h}}(t)$ solves the initial value problem:

$$\dot{u}(t) = Au(t),\ u(0) = u_0. \tag{8}$$

The inhomogeneous part of the solution is derived with the solution formula in (3):

$$\bar{u}^{\mathrm{inh}}(t) = \int_0^t T(t-s)f(s)\,\mathrm{d}s,$$

where the source term (4) is plugged in for f and we obtain by linearity

$$\bar{u}^{\mathrm{inh}}(t) = \sum_{l=1}^{L} \int_0^t T(t-s)\tilde{f}_l\ w_l(s)v_l(s)\,\mathrm{d}s.$$

Define $\bar{u}^l(t) = T(t)\tilde{f}_l$, or in other words let $\bar{u}^l(t)$ solve the initial value problem:

$$\dot{u}(t) = Au(t),\ u(0) = \tilde{f}_l, \tag{9}$$

for $l = 1, \ldots, L$. Combining the $\bar{u}^l(t)$ solutions, we get

$$\bar{u}^{\mathrm{inh}}(t) = \sum_{l=1}^{L} \int_0^t \bar{u}^l(t-s)\ w_l(s)v_l(s)\,\mathrm{d}s.$$

This may be written as convolution (Def. 2.3) as follows:

$$\bar{u}^{\mathrm{inh}}(t) = \sum_{l=1}^{L} (\bar{u}^l * (w_l v_l))(t).$$

Adding the homogeneous and the inhomogeneous parts of the solution completes the proof. □

This lemma provides an explicit solution of the PDE and permits us to formulate an optimal control problem from which the dynamics have been eliminated.

Theorem 3.1 (PDE-free MIPDECO)**.** *The optimal control problem* (5) *is equivalent to*

$$\begin{aligned} \min_{u,v,w}\ & J(u,v,w) = \phi(u(t_f)) + \int_0^{t_f} \psi(u(t), v(t), w(t))\ \mathrm{d}t \\ s.t.\ & v \in \mathcal{V}_{\mathrm{ad}}, w \in \mathcal{W}_{\mathrm{ad}}, \end{aligned} \tag{10}$$

where $u(t) = u(t,v,w) = T(t)u_0 + \sum_{l=1}^{L} ((T\tilde{f}_l) * (w_l v_l))(t)$*.*

This reduced problem can save significant computational effort since PDE optimization problems are typically discretized. The size of discrete representation grows with the size of the discretization mesh and therefore the number of discretized state variables, that is, $\mathcal{O}(n^4)$ for three space dimensions and time and n grid points in every dimension.

Remark 3.1 (Cost of Deriving (10))**.** *With our approach, we eliminate the PDE once before starting the optimization algorithm. Thus, the additional effort of computing the reduced problem is independent of the iteration number of the optimization algorithm, which can grow exponentially with the discretization. Here, the effort depends only linearly on L, since $L+1$ initial value problems are solved: once for* (8) *and L times for* (9)*.*

Also note that the existence of $T(t)$ depends on the operator A; we refer readers to [18] for details.

3.2 Solution Operator in Discrete Time

Because of the presence of integer variables and the lack of suitable first-order optimality conditions, the first-discretize-then-optimize approach is used mostly for MIPDECO. The discretization of the reduced problem needs careful treatment, because the convolution cannot be discretized with standard quadrature rules; see, for example, [15].

Therefore, we aim to extend the explicit representation of $u(t)$ as in Theorem 3.1 for the discretized state. We discretize time as follows. Let $0 = t_0 < \cdots < t_{T_n} = t_f$ be a time discretiziation, and denote by $v_{k,l} = v_l(t_k)$ and $w_{k,l} = w_l(t_k)$ the discrete values of the controls. Denote by

$$\mathbf{v} = \begin{bmatrix} v_{0,1} & \cdots & v_{0,L} \\ \vdots & \ddots & \vdots \\ v_{T_n,1} & \cdots & v_{T_n,L} \end{bmatrix}, \mathbf{w} = \begin{bmatrix} w_{0,1} & \cdots & w_{0,L} \\ \vdots & \ddots & \vdots \\ w_{T_n,1} & \cdots & w_{T_n,L} \end{bmatrix},$$

the matrices of the discretized controls. We use $\mathbf{v}_{k,:}$ or $\mathbf{w}_{k,:}$ to refer to the kth column. The vector of the discretized state is denoted by $\mathbf{u}$, respectively.

Lemma 3.2 (Control-to-state map in discrete time)**.** *Let Assumption 3.1 hold. Then the discretized solution of the dynamical system* (6) *is given in terms of the controls $\mathbf{v}, \mathbf{w}$, and we have*

$$u_k = \bar{u}_k^{\mathrm{h}} + \sum_{l=1}^{L} \sum_{m=0}^{k} \bar{u}_{k-m}^{l} w_{m,l} v_{m,l}, \tag{11}$$

where $u_k = u(t_k)$ and homogeneous and inhomogeneous parts of the solution are $\bar{u}_k^{\mathrm{h}} = \bar{u}^{\mathrm{h}}(t_k)$ and $\bar{u}_k^l = \bar{u}^l(t_k)$, respectively, for $k = 0, \ldots, T_n$.

Proof. The solution in (7) is expressed as a (continuous) convolution that can directly be rewritten as a discrete convolution of the sequence of the discrete inhomogeneous solution part $(\bar{u}_k^l)_k$ with the sequence of the product of the controls $(v_{k,l} w_{k,l})_k$:

$$u_k = \bar{u}_k^{\mathrm{h}} + \sum_{l=1}^{L} (\bar{u}_{\cdot}^{l} * (w_{\cdot,l} v_{\cdot,l}))[k].$$

With Def. 2.4 it follows immediately that

$$u_k = \bar{u}_k^{\mathrm{h}} + \sum_{l=1}^{L} \sum_{m=0}^{k} \bar{u}_{k-m}^{l} w_{m,l} v_{m,l}.$$

□

With this explicit representation of the discretized solution, according to Theorem 3.1 we state the time-discrete MIPDECO without PDE below.

Theorem 3.2 (PDE-free MIPDECO in discrete time)**.** *The optimal control problem in* (10) *can be rewritten as*

$$\begin{aligned}\min_{\mathbf{u},\mathbf{v},\mathbf{w}} \quad & \mathbf{J}(\mathbf{u},\mathbf{v},\mathbf{w}) \\ & = \phi(u_{T_n}) + \sum_{k=0}^{T_n} a_k \psi(u_k, \mathbf{v}_{k,:}, \mathbf{w}_{k,:}) \\ s.t. \quad & \mathbf{v} \in \mathbf{V}_{\mathrm{ad}}, \mathbf{w} \in \mathbf{W}_{\mathrm{ad}},\end{aligned} \tag{12}$$

where $u_k = \bar{u}^{\mathrm{h}}_k + \sum\limits_{l=1}^{L} \sum\limits_{m=0}^{k} \bar{u}^l_{k-m} w_{m,l} v_{m,l}$. *In the discretized objective* $\mathbf{J}$*, the integral is replaced by a suitable quadrature with the weights* a_k *for* $k = 0, \ldots, T_n$*. Further, let* $\mathbf{V}_{\mathrm{ad}} \subseteq \mathbb{R}^{T_n+1\times L}$ *and* $\mathbf{W}_{\mathrm{ad}} \subseteq \mathbb{Z}^{T_n+1\times L}$ *denote the discretization of the admissible control sets* $\mathcal{V}_{\mathrm{ad}}$ *and* $\mathcal{W}_{\mathrm{ad}}$*.*

Note that the representation in (12) is semi-discrete; that is, the spatial dimension is somewhat hidden in the structure of the state variable $\mathbf{u}$.

However, this discretized representation allows us to quantify the computational effort needed to derive the problem formulation (12), which is computed prior to the optimization.

Remark 3.2 (Computational Cost of Deriving (12))**.** *In particular, we need to compute the discrete representations of* $\bar{u}^{\mathrm{h}}$ *and* $\bar{u}^l$ *for* $l = 1, \ldots, L$*. Each of these objects is uniquely defined by a system of equations that are due to the chosen spatial discretization. More precisely, in the case of linear systems, the coefficient matrices of these systems are identical, and only the right-hand side of the equation varies. This makes it relatively cheap since a once-computed LU decomposition can be used for all systems. Of course the effort to compute, for example, an LU decomposition depends on the chosen discretization and the mesh size, but we want to highlight that it is needed only once before the optimization.*

These advantageous properties are explained in more detail for an example problem in Section 5.

4 Control of Heat Equation with Optimal Actuator Placement

In this section a MIPDECO problem governed by the heat equation is introduced. The problem is adapted from [14]. The goal is to place and operate a small and fixed number of actuators (e.g., one or two) over time in a given domain. The possible locations are given as a finite set of coordinates in space. An example of a possible actuator distribution is given in Figure 1. First, we present a problem with binary and real-valued controls that also model placement and intensity control.

4.1 Model

We consider a rectangle $\Omega = [0,1] \times [0,2]$ and the time horizon $[0, t_f]$. The objective (13a) is quadratic: its first term is of tracking type and captures the desired final state u_f, the second term regularizes the state, and the third term regularizes the real-valued control with the regularization parameters $\beta, \gamma \in \mathbb{R}_+$. The constraints are a source budget (13f), which limits the quantity of placed

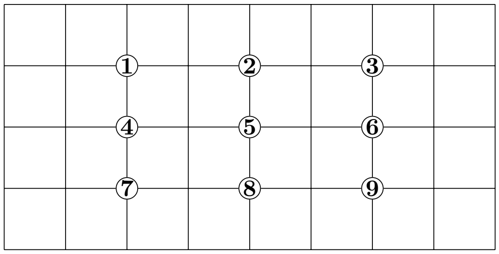

Figure 1: Domain Ω with actuator locations

actuators, and the two-dimensional heat equation (13b) with some source term. Additionally, we assume Dirichlet boundary (13c) and initial conditions (13d). These can be written as follows.

$$\begin{aligned} \min_{u,v,w} \quad & J(u,v) = \|u(t_f,x) - u_f(x)\|_{2,\Omega}^2 \\ & + \beta \|u(t,x)\|_{2,[0,t_f]\times\Omega}^2 + \gamma \sum_{l=1}^{L} \|v_l(t)\|_{2,[0,t_f]}^2 && \text{(13a)} \\ \text{s.t.} \quad & \frac{\partial u}{\partial t}(t,x) - \kappa \Delta u(t,x) = \sum_{l=1}^{L} v_l(t) f_l(x) \text{ in } (0,t_f] \times \Omega && \text{(13b)} \\ & u(t,x) = 0 \text{ in } [0,t_f] \times \partial\Omega && \text{(13c)} \\ & u(0,x) = u_0(x) \text{ in } \Omega && \text{(13d)} \\ & -\mathbf{M} w_l(t) \le v_l(t) \le \mathbf{M} w_l(t) \\ & \quad \text{for all } l \in \{1,\dots,L\} \text{ in } [0,t_f] && \text{(13e)} \\ & \sum_{l=1}^{L} w_l(t) = \bar{w} \text{ in } [0,t_f] && \text{(13f)} \\ & w_l(t) \in \{0,1\} \text{ for all } l \in \{1,\dots,L\} \text{ in } [0,t_f]. && \text{(13g)} \end{aligned}$$

The variables are the state $u : [0,t_f] \times \Omega \to \mathbb{R}$, the binary controls $w_l : [0,t_f] \to \{0,1\}$, and the real-valued controls $v_l : [0,t_f] \to \mathbb{R}$ for $l = 1,\dots,L$. The nonnegative integer $\bar{w}$ denotes the quantity of actuators and L the quantity of their possible locations. The thermal diffusivity κ can either be constant $\kappa \in \mathbb{R}_+$ or vary in space $\kappa = \kappa(x,y) \in \mathbb{R}_+$ representing a certain material or a distribution of various materials. We define the source term for all locations $l \in \{1,\dots,L\}$ and a fixed parameter $\varepsilon \in \mathbb{R}_+$ as

$$f_l(x) = \frac{1}{\sqrt{2\pi\varepsilon}} \exp\left(\frac{-\left\|x^l - x\right\|^2}{2\varepsilon} \right), \tag{14}$$

where x^l is the coordinate of the mesh point of actuator location l.

Remark 4.1. *Originally, the problem formulation in [14] included a nonconvex right-hand side of the heat equation:*

$$\frac{\partial u}{\partial t}(t,x) - \kappa \Delta u(t,x) = \sum_{l=1}^{L} v(t) w_l(t) f_l(x).$$

We overcome this potential issue by substitution of $v(t)w_l(t)$ by $v_l(t)$ in (13b)*. Furthermore, we introduce a bound $\mathbf{M} \gg 1$ on the real-valued controls v_l in* (13e)*, and we limit the amount of* *actuators by the source budget constraint* (13f). *This formulation is more general in the sense that we can allow more than one actuator in the model.*

We note that we can easily generalize this problem, for example, by using L_1 regularization or other regularizations.

4.2 Solution Space and Existence of Solutions

For $w_l \in L^2(0, [0, t_f])$, $v_l \in L^2(0, [0, t_f])$, $f_l \in L^2(0, [0, t_f])$, $u_0 \in L^2(\Omega)$, and constant κ, we would expect the solution u in the weak sense (13b–13d) to be in $W_2^{1,0}([0, t_f] \times \Omega)$, the linear space of all $u \in L^2([0, t_f] \times \Omega)$ having a weak first-order partial derivative with respect to (x, y) in $L^2([0, t_f] \times \Omega)$, which is discussed in more detail [22, Chapter 3]. In particular, with these assumptions the objective J is then well defined.

However, existence of optimal solutions is in general unclear because of the integrality constraints (13g). If this integrality is relaxed, the constraints (13e) and (13f) become redundant and thus also the variables w_l for $l = 1, \ldots, L$. The remaining PDE optimization problem (13a-13d) has a strictly convex objective and linear constraints. Therefore it is a unique optimal control for the relaxation.

4.3 Formulation as an Abstract ODE

The heat equation in the actuator placement and operation problem in (13) is a parabolic PDE that can be formulated as an abstract ODE as in (1). Thus, instead of (13), we can equivalently write the following:

$$\begin{aligned}
\min_{u,v,w} \quad & J(u, w) = \|u(t_f) - u_f\|_2^2 \\
& + \beta \|u(t)\|_{2,[0,t_f]}^2 + \gamma \sum_{l=1}^{L} \|v_l(t)\|_{2,[0,t_f]}^2 && (15a) \\
\text{s.t.} \quad & \dot{u}(t) = Au(t) + \sum_{l=1}^{L} v_l(t) f_l \text{ in } [0, t_f] && (15b) \\
& u(0) = u_0 && (15c) \\
& u \in \mathcal{U}_{\text{ad}} && (15d) \\
& v \in \mathcal{V}_{\text{ad}} = \{v | -\mathbf{M} w_l(t) \le v_l(t) \le \mathbf{M} w_l(t), \\
& \qquad\qquad l = 1, \ldots, L\} && (15e) \\
& w \in \mathcal{W}_{\text{ad}} = \{w | \textstyle\sum_{l=1}^{L} w_l(t) = \bar{w}\}. && (15f)
\end{aligned}$$

The linear infinitesimal generator of the strongly continuous semigroup is

$$(Au)(x) = \kappa \sum_{n=1}^{2} \frac{\partial^2 u}{\partial x_n^2},$$

where its domain is $\mathcal{D}(A) = H^2(\Omega) \cap H_0^1(\Omega)$ and the strongly continuous semigroup of contractions is $\{T(t)\}_{t \ge 0}$ on $\mathcal{U}$. We choose the admissible sets $\mathcal{U}_{\text{ad}} = C([0, t_f]; U)$, $\mathcal{V}_{\text{ad}} = \mathcal{V} = C_{\text{pw}}\left([0, t_f]; \mathbb{R}^L\right)$, and $\mathcal{W}_{\text{ad}} \subset \mathcal{W} = L^\infty\left([0, t_f]; \{0, 1\}^L\right)$.

With the formulation of the problem (15) together with the choice of the appropriate spaces, the conditions in Assumption 3.1 are satisfied, and Theorem 3.2 is applicable.

To derive the ingredients of the control-to-state-map as in reduced problem in Theorem 3.1, one must solve $L+1$ initial boundary value problems; see Remark 3.1. The homogeneous solution $\bar{u}^{\mathrm{h}}$ represents thermal diffusion of the initial state u_0 without any control application; that is, all controls are fixed to zero.

$$\begin{aligned} \frac{\partial u}{\partial t}(t,x) - \kappa \Delta u(t,x) &= 0 && \text{in } [0,t_f] \times \Omega \\ u(t,x) &= 0 && \text{in } [0,t_f] \times \partial\Omega \\ u(t,x) &= u_0(x) && \text{in } \Omega \end{aligned}$$

The inhomogeneous parts of the solution $\bar{u}^l$ are the solutions of the heat equation with $f_l(x)$ as initial state, which can be interpreted as the control v_l applied in time $t = 0$. Thus, we have for $l = 1, \ldots, L$

$$\begin{aligned} \frac{\partial u}{\partial t}(t,x) - \kappa \Delta u(t,x) &= 0 && \text{in } [0,t_f] \times \Omega, \\ u(t,x) &= 0 && \text{in } [0,t_f] \times \partial\Omega, \\ u(0,x) &= f_l(x) && \text{in } \Omega. \end{aligned}$$

Also the discrete version of the approach, Theorem 3.2, applies to a time discretization of (15). We explain how the advantages of our approach become clear in the computation in the following section.

5 Spatial Discretization via Finite Differences

Now we explain in detail how the results in Section 3 reduce significantly the optimization of MIPDECO. For this purpose we apply the method to the discretized version of the model in Section 4. We conclude with the statement of the reduced MINLP.

5.1 Discretization of the Heat Equation

Since we are studying the interaction of MINLP and PDE constrained optimization, we use simple discretization schemes only. However, our results can be generalized to other methods and meshes.

We consider the uniform step size in space and time and define approximate values of the states and controls for $k = 0, \ldots, T_n$, $i = 0, \ldots, N$, and $j = 0, \ldots, M$:

$$u_{k,i,j} \approx u(kh_t, (ih_{x_1}, jh_{x_1})), v_{k,l} \approx v_l(kh_t), w_{k,l} \approx w_l(kh_t).$$

We discretize the PDE in (13) using a central difference operator in space and backward difference operator in time, yielding the linear system

$$GU = BV + d, \tag{16}$$

where $G \in \mathbb{R}^{T_nNM \times T_nNM}$ and $B \in \mathbb{R}^{T_nNM \times T_nL}$ contain coefficients, $d \in \mathbb{R}^{T_nNM}$ contains initial and boundary conditions, and $U \in \mathbb{R}^{T_nNM}$ and $V \in \mathbb{R}^{T_nL}$ are the unknown states and controls, written as vectors:

$$U = \begin{bmatrix} \mathrm{vec}(u_{1,\cdot,\cdot}) \\ \vdots \\ \mathrm{vec}(u_{T_n,\cdot,\cdot}) \end{bmatrix}, \quad V = \begin{bmatrix} \mathrm{vec}(v_{1,\cdot}) \\ \vdots \\ \mathrm{vec}(v_{T_n,\cdot}) \end{bmatrix}. \tag{17}$$

The matrix G may be written as the sum of two Kronecker products:

$$G = C \otimes I_{NM} + I_{T_n} \otimes \kappa K, \tag{18}$$

where I_{NM} and I_{T_n} denote identity matrices of dimension NM and T_n, respectively. The matrix $C \in \mathbb{R}^{T_n \times T_n}$ is an implicit Euler matrix:

$$C = \frac{1}{h_t} \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix},$$

where $h_t = \frac{t_f}{T_n}$ is the time step size and $K \in \mathbb{R}^{NM \times NM}$ is the coefficient matrix of the five-point stencil discretization of the Laplace operator:

$$K = \frac{1}{h_{x_1} h_{x_2}} \begin{bmatrix} D & -I_N & & & \\ -I_N & D & -I_N & & \\ & \ddots & \ddots & \ddots & \\ & & -I_N & D & -I_N \\ & & & -I_N & D \end{bmatrix},$$

where $h_{x_1} = \frac{1}{N}$ and $h_{x_2} = \frac{2}{M}$ denote the space step sizes discretizing in domain $\Omega = [0, 1] \times [0, 2]$. The matrix I_N is the identity of dimension N, and $D \in \mathbb{R}^{N \times N}$ is a tridiagonal matrix:

$$D = \begin{bmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{bmatrix}.$$

The right-hand side of the linear system (16) consists of the right-hand side of the heat equation (13b) written as

$$B = I_{T_n} \otimes F, \tag{19}$$

with I_{T_n} the identity matrix of dimension T_n, the source term (14) $F = [\text{vec}(f_1), \ldots, \text{vec}(f_L)] \in \mathbb{R}^{NM \times L}$, and the initial (13d) and boundary conditions (13c): $d = [\text{vec}(u_{0,\cdot,\cdot})^T, 0, \ldots, 0]^T \in \mathbb{R}^{NMT_n}$.

5.2 Alternative Derivation of the Reduction Approach

To solve the linear system (16), we use a key property of V that we can write as a linear combination of unit vectors of $\mathbb{R}^{T_n L}$ and reformulate (16):

$$GU = B \sum_{i=1}^{LT_n} V_i e_i + d, \tag{20}$$

where $\{e_i\}_{i=1}^{T_n L}$ denotes the standard basis of $\mathbb{R}^{T_n L}$ and $V_i \in \mathbb{R}$ corresponds to a $v_{k,l}$; see (17). Since G is regular, we use linearity and get

$$U = \sum_{i=1}^{LT_n} V_i \underbrace{G^{-1} B e_i}_{\text{inhomogeneous part}} + \underbrace{G^{-1} d}_{\text{homogeneous part}}. \tag{21}$$

In this discrete formulation we need to solve for the initial values d and for every column of B, thus $T_n L + 1$ linear systems.

However, since B is the Kronecker product of an identity with the matrix F (see (19)), it is a block diagonal matrix with identical blocks, namely, F. In addition, the matrix G is also special with regard to the structure, being the finite difference discretization; see Equation (18).

Thanks to these structural properties, we need to solve only for $i = 1, \dots, L$ in Equation (21) corresponding to the L columns of F and get the subsolutions $(\bar{u}^l_k)_{k=0}^{T_n}$. In this manner we obtain the inhomogeneous part of the solution yielding a total of $L + 1$ linear systems to solve, as stated in Remark 3.2.

Hence, instead of solving the linear system for every time step and every location, we need to solve for every location only once to obtain the inhomogeneous part of the solution. Then, the partial solutions are shifted in time with the corresponding control in the convolution formula (11).

Note also that the coefficient matrix G is the same for all $L + 1$ linear equation systems. Thus, if a direct solver is used, the LU composition of G can be used for all systems which makes computations cheap.

5.3 MINLP Formulation of the MIPDECO

Because the discretized PDE is a system of linear equations, we can eliminate the PDE by solving a linear system for every pair of control variables $(v_{k,l}, w_{k,l})$. This idea leads to a mixed-integer quadratic program (MIQP) formulation of the problem with reduced size that is solvable within less computational time than the original MIQP formulation requires.

The reduced objective $\mathbf{J}$ in (23) of this problem, in combination with the bounds (22b) and source budget (22c), form the reduced MIQP formulation of the problem:

$$\begin{aligned}
\min_{\mathbf{v},\mathbf{w}} \quad & \mathbf{J}(\mathbf{v}) && \text{(22a)}\\
\text{s.t.} \quad & -\mathbf{M} w_{k,l} \le v_{k,l} \le \mathbf{M} w_{k,l} && \text{(22b)}\\
& \sum_{l=1}^{L} w_{k,l} = \bar{w} && \text{(22c)}\\
& w_{k,l} \in \{0,1\} && \text{(22d)}\\
& \text{for } l = 1, \dots, L \text{ and } k = 1, \dots, T_n.
\end{aligned}$$

This problem consists of LT_n binary and LT_n continuous variables, while NMT_n state variables are eliminated.

6 Numerical Results

In this section we present results of our numerical experiments for the actuator placement problem in (13).

The problem is discretized as in Section 5 and implemented in the modeling language AMPL [4]. The resulting MIQP is high dimensional in terms of variables and constraints. The number of real and binary variables is stated depending on the size of the number of grid points in space and time $N = 0.5M = T_n = N_c$ in context with the computational time in Table 1. Note that N_c is the number of discretization points of the control that may coincide with T_n or be independent. The CPU time of CPLEX [3] is compared with the CPU time of CLPEX with prior state elimination. A dash indicates that no result could be obtained within the time limit. The presented state

Mesh size	8	16	32	64
Variables				
real	1449	9681	71073	545601
binary	72	144	288	576
CPU (CPLEX)	7.3	5567.3	–	–
State Elim.				
CPU (Elim.)	0.004	0.228	3.604	–
CPU (CPLEX)	1.056	3.224	13.280	–

Table 1: Problem size and CPU time for the discretization of the actuator placement problem in (13).

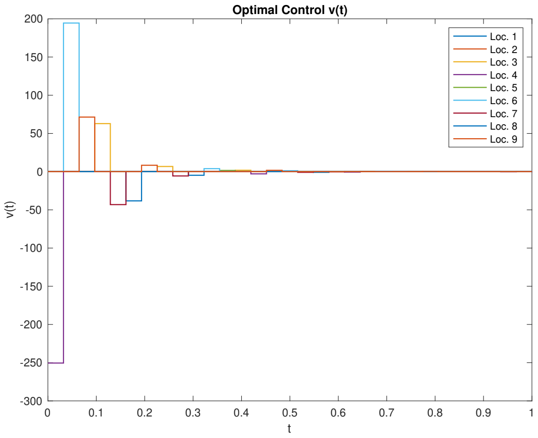

Figure 2: Control: The control in the different actuator positions is plotted over time.

elimination approach reduces the number of real variables of instances to the same as binary variables.

In the implementation we used the following parameters. The domain is $\Omega = [0,1] \times [0,2]$ with the actuator locations as in Figure 1, namely, $L = 9$. The final time is $t_f = 1$. The desired final state is chosen to be $u_f(x) = 0$, and the initial condition is $u_0(x) = 100\sin(\pi x_1)\sin(\pi x_2)$ for all $x \in \Omega$. The regularization parameters are $\beta = 2$ and $\gamma = 2 \cdot 10^{-3}$. The thermal diffusivity is constant $\kappa = 0.01$ in the domain Ω. and a single actuator is considered, namely, $\bar{w} = 1$. The parameter of the Gaussian source term is chosen as $\varepsilon = 0.01$.

As the implementation parameter we set the bound to $\mathbf{M} = 2500$, choose the size of the mesh to be $N = 0.5M = T_n = 32$, and vary the control grid $N_c \in \{4, 8, 16, 32\}$.

The graphs in Figure 2 show the location and the intensity of the placed actuator for the optimal control v. The decrease of the intensity until $t \approx 0.5$ is due to the fact that the state is driven quickly close to the desired state $u_f \equiv 0$ (see Figure 3) and due to the regularization of the control in the objective.

In Figure 3 the optimal state u is shown for $t \in \{0, 0.125, 1\}$ and compared with the state at $t = 1$ without any control application. Because of the control application, the norm of the state can be brought to the same order of magnitude as without control application in less than $\Delta t = 0.125$.

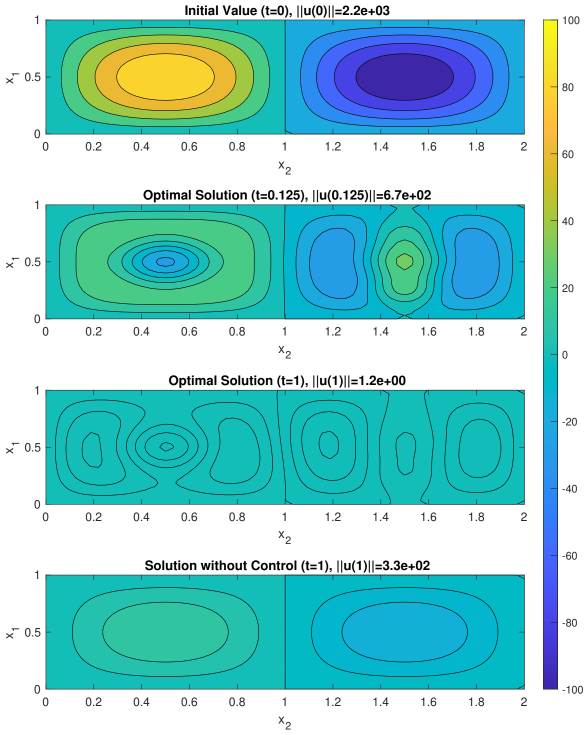

Figure 3: State: The initial state is plotted on the top, the final state of the optimal solution below, and the final state of the heat equation without control application on the bottom.

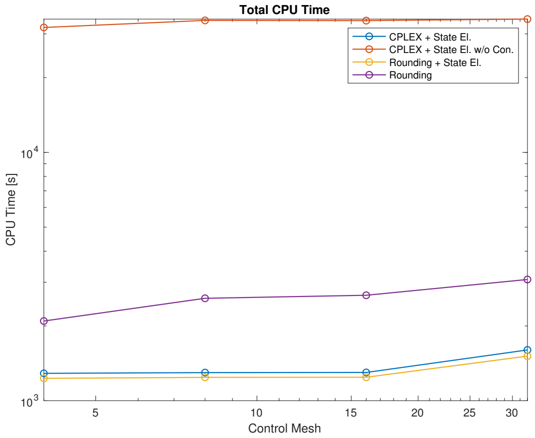

Figure 4: CPU Time: Comparison of the CPU time of the state elimination, with an approach without the smart use of the convolution, and with a rounding approach. The discretization of the states is $N = T_n = 32$, and the discretization of the control varies. Note that the application of CPLEX without a state elimination does not terminate with results in a one-day time limit and is therefore not presented in this figure.

The effect of the control makes the norm of the state at final time $t_f = 1$ significantly smaller and more homogeneous than without control in the bottom picture.

In Figure 4 we compare the state elimination method presented here in blue with a state elimination method that does not apply the convolution in red. We also compare with sum-up rounding (see [19]) that requires solving to full integer relaxations of the problem and provides high-quality approximations to the optimal control. This illustrates the potential of the saved computational effort during the optimization because our method outperforms also this relatively cheap rounding scheme. Note that the state elimination is also beneficial combined with rounding because it simplifies the solution of the integer relaxations significantly; see the yellow curve in Fig. 4. We also observe that all curves rise relatively flat. The reason is that we have fixed the space and time discretization of the states to $N = T_n = 32$ and vary only the number of grid points of the controls N_c. This also illustrates in particular that the bottleneck is the high-dimensional state variable.

7 Conclusion

We introduced a class of mixed-integer PDE-constrained optimal control problems whose computational complexity can be reduced by elimination of the state variables. This elimination method is derived by semigroup theory and clever convolution of solutions parts for continuous and discrete time. The relevance of this class of problems is illustrated by the actuator placement problem that is governed by the time-dependent 2D heat equation. For this example, we conclude with numerical results. Because of the efficient reduction of the problem size, the proposed method outperforms simple rounding schemes in terms of computational time.

Appendix

The reduced objective is discretized by the trapezoidal rule:

$$\begin{aligned}
&\mathbf{J}(\mathbf{v}) = \\
&h_{x_1} h_{x_2} \sum_{i=1}^{N-1} \sum_{j=1}^{M-1} \Bigg(\bar{u}^{\mathrm{h}}_{Tn,i,j} + \sum_{t=1}^{T_n} \sum_{l=1}^{L} v_{t,l} \; \bar{u}^{l}_{Tn-t,i,j} - u_{f,i,j} \Bigg)^2 \\
&+ \beta \Bigg(\frac{1}{2} h_{x_1} h_{x_2} h_t \sum_{i=1}^{N-1} \sum_{j=1}^{M-1} \left(\bar{u}^{\mathrm{h}}_{0,i,j} \right)^2 \\
&+ h_{x_1} h_{x_2} h_t \sum_{k=1}^{T_n-1} \sum_{i=1}^{N-1} \sum_{j=1}^{M-1} \Bigg(\bar{u}^{\mathrm{h}}_{k,i,j} + \sum_{t=1}^{k} \sum_{l=1}^{L} v_{t,l} \; \bar{u}^{l}_{k-t,i,j} \Bigg)^2 \\
&+ \frac{1}{2} h_{x_1} h_{x_2} h_t \sum_{i=1}^{N-1} \sum_{j=1}^{M-1} \Bigg(\bar{u}^{\mathrm{h}}_{Tn,i,j} + \sum_{t=1}^{T_n} \sum_{l=1}^{L} v_{t,l} \; \bar{u}^{l}_{T_n-t,i,j} \Bigg)^2 \Bigg) \\
&+ \gamma h_t \sum_{l=1}^{L} \Bigg(\frac{1}{2} \left(v_{0,l} \right)^2 + \sum_{t=1}^{T_n-1} \left(v_{t,l} \right)^2 + \frac{1}{2} \left(v_{Tn,l} \right)^2 \Bigg),
\end{aligned} \tag{23}$$

where $u_{f,i,j} = u_f((ih_{x_1}, jh_{x_2}))$.

Acknowledgment

We thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for financial support through 314838170, GRK 2297 MathCoRe, SPP 1962, STE2063/2-1, and HE5386/19-1. This work was supported by the Applied Mathematics activity within the U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research program, under Contract DE-AC02-06CH11357.

References

[1] Pietro Belotti, Christian Kirches, Sven Leyffer, Jeff Linderoth, James Luedtke, and Ashutosh Mahajan. Mixed-integer nonlinear optimization. *Acta Numerica*, 22:1–131, apr 2013.

[2] A. Bürger, C. Zeile, M. Hahn, A. Altmann-Dieses, S. Sager, and M. Diehl. pycombina: An open-source tool for solving combinatorial approximation problems arising in mixed-integer optimal control. 2020. accepted.

[3] IBM ILOG Cplex. IBM ILOG CPLEX Optimization Studio CPLEX User's Manual. *International Business Machines Corporation*, 2017.

[4] Robert Fourer, David M. Gay, and Brian W. Kernighan. *Ampl: A Modeling Language for Mathematical Programming.* DUXBURY, 2002.

[5] Dominik Garmatter, Margherita Porcelli, Francesco Rinaldi, and Martin Stoll. Improved penalty algorithm for mixed integer PDE constrained optimization problems. 2019.

[6] S. Göttlich, A. Potschka, and C. Teuber. A partial outer convexification approach to control transmission lines. *Computational Optimization and Applications*, 72(2):431–456, nov 2018.

[7] Simone Göttlich, Michael Herty, and Ute Ziegler. Modeling and optimizing traffic light settings in road networks. *Computers & Operations Research*, 55:36–51, mar 2015.

[8] Simone Göttlich, Andreas Potschka, and Ute Ziegler. Partial outer convexification for traffic light optimization in road networks. *SIAM Journal on Scientific Computing*, 39(1):B53–B75, jan 2017.

[9] Mirko Hahn, Sven Leyffer, and Victor M. Zavala. Mixed-integer PDE-constrained optimal control of gas networks. *Preprint ANL*, aug 2017.

[10] Falk M. Hante. Relaxation methods for hyperbolic PDE mixed-integer optimal control problems. *Optimal Control Applications and Methods*, 38(6):1103–1110, mar 2017.

[11] Falk M. Hante and Sebastian Sager. Relaxation methods for mixed-integer optimal control of partial differential equations. *Computational Optimization and Applications*, 55(1):197–225, nov 2012.

[12] Jaroslav Haslinger and Raino A. E. Mäkinen. On a topology optimization problem governed by two-dimensional Helmholtz equation. *Computational Optimization and Applications*, 62(2):517–544, apr 2015.

[13] Michael Hinze, Michael Ulbrich, and René Pinnau. *Optimization with PDE Constraints.* Springer-Verlag GmbH, 2008.

[14] Orest V. Iftime and Michael A. Demetriou. Optimal control of switched distributed parameter systems with spatially scheduled actuators. *Automatica*, 45(2):312–323, feb 2009.

[15] C. Lubich. Convolution quadrature and discretized operational calculus, I. *Numerische Mathematik*, 52(2):129–145, jan, 1988.

[16] Paul Manns and Christian Kirches. Multi-dimensional sum-up rounding using hilbert curve iterates. *PAMM*, 19(1), nov 2019.

[17] Paul Manns and Christian Kirches. Improved regularity assumptions for partial outer convexification of mixed-integer PDE-constrained optimization problems. *ESAIM: Control, Optimisation and Calculus of Variations*, 26:32, 2020.

[18] A. Pazy. *Semigroups of Linear Operators and Applications to Partial Differential Equations.* Springer New York, 1983.

[19] Sebastian Sager, Hans Georg Bock, and Moritz Diehl. The integer approximation error in mixed-integer optimal control. *Mathematical Programming*, 133(1-2):1–23, sep 2010.

[20] Sebastian Sager, Michael Jung, and Christian Kirches. Combinatorial integral approximation. *Mathematical Methods of Operations Research*, 73(3):363–380, apr 2011.

[21] Sebastian Sager and Clemens Zeile. On mixed-integer optimal control with constrained total variation of the integer control. *Computational Optimization and Applications*, dec 2020.

[22] Fredi Tröltzsch. *Optimale Steuerung partieller Differentialgleichungen.* Vieweg+Teubner Verlag, 2009.

[23] Clemens Zeile, Nicolò Robuschi, and Sebastian Sager. Mixed-integer optimal control under minimum dwell time constraints. *Mathematical Programming*, jul 2020.