Relaxation methods for mixed-integer optimal control of partial differential equations

Falk M. Hante · Sebastian Sager

Abstract We consider integer-restricted optimal control of systems governed by abstract semilinear evolution equations. This includes the problem of optimal control design for certain distributed parameter systems endowed with multiple actuators, where the task is to minimize costs associated with the dynamics of the system by choosing, for each instant in time, one of the actuators together with ordinary controls. We consider relaxation techniques that are already used successfully for mixed-integer optimal control of ordinary differential equations. Our analysis yields sufficient conditions such that the optimal value and the optimal state of the relaxed problem can be approximated with arbitrary precision by a control satisfying the integer restrictions. The results are obtained by semigroup theory methods. The approach is constructive and gives rise to a numerical method. We supplement the analysis with numerical experiments.

1 Introduction and problem formulation

The factoring of decision processes interacting with continuous evolution plays an important role in model-based optimization for many applications. For example,

F.M. Hante (✉)
Department of Mathematics, University of Erlangen-Nuremberg, Cauerstraße 11, 91058 Erlangen, Germany
e-mail: hante@math.fau.de

S. Sager
Institute of Mathematical Optimization, Otto-von-Guericke University, Universitätsplatz 2, 39106 Magdeburg, Germany
e-mail: sager@ovgu.de

when optimally controlling chemical processes there are often both continuous decisions such as inlet and outlet flows, as well as discrete decisions, such as the operation of on–off valves and pumps that may redirect flows within the reactor, [14, 21]. Such mixed-integer optimal control problems are therefore studied in different communities with different approaches. Most of these approaches address problems that are governed by systems of ordinary differential equations in Euclidean spaces, see [22] for a survey on this topic.

Total discretization of the underlying system obviously leads to typically large mixed-integer nonlinear programs. Hence, relaxation techniques have become an integral part of efficient mixed-integer optimal control algorithms, either in the context of branch-and-bound type methods or, more directly, by means of nonlinear optimal control methods combined with suitable rounding strategies. An important result is that the solution of the relaxed problem can be approximated with arbitrary precision by a solution fulfilling the integer requirements [23].

In this paper, we extend such relaxation techniques to problems that are governed by certain systems of partial differential equations. Motivating applications are for example to switch between reductive and oxidative conditions in order to maximize the performance in a monolithic catalyst [27], port switching in chromatographic separation processes [5, 14], or to optimize switching control within photochemical reactions [26]. Our problem setting also includes the switching control design in the sense that systems are equipped with multiple actuators and the optimizer has to choose one of these together with ordinary controls for each instant in time.

Concerning systems involving partial differential equations, such switching control design has already been studied using several techniques: In [18, Chap. 8] optimal switching controls are constructed for systems governed by abstract semilinear evolution equations by combining ideas from dynamic programming and approximations of the value function using viscosity solutions of the Hamilton-Jacobi-Bellman equations. Switching boundary control for linear transport equations using switching time sensitivities has been studied in [9]. Exemplary for the heat equation and based on variational methods, the controllability in case of switching among several actuators has been considered in [29] and null-controllability for the one-dimensional wave equation with switching boundary control has been considered in [7]. Based on linear quadratic regulator optimal control techniques and enumeration of the integer values for a fixed time discretization, optimal switching control of abstract linear systems has been considered in [13].

Our approach is complementary to the above, as we break the computationally very expensive combinatorial complexity of the problem by relaxation. This comes at the downside of providing only a suboptimal solution and possibly at the price of fast switching but, as we will see, with arbitrary small integer-optimality gap, depending on discretization, and extensions to limit the number of switching.

We will be concerned with the following problem of mixed-integer optimal control: Minimize a cost functional

$$J = \phi\big(z(t_f)\big) + \int_0^{t_f} \psi\big(z(t), u(t)\big)\, dt \tag{1}$$

over trajectories $z \in X_{[0,t_f]} \subset \{z : [0, t_f] \to X\}$ and control functions $u \in U_{[0,t_f]} \subset \{u : [0, t_f] \to U\}$ and $v \in V_{[0,t_f]} \subset \{v : [0, t_f] \to V\}$ subject to the constraints that z is a mild solution of the operator differential equation

$$\begin{cases} \dot{z}(t) = Az(t) + f\big(t, z(t), u(t), v(t)\big), & t \in (0, t_f] \\ z(0) = z_0 \in X \end{cases} \tag{2}$$

and that the control functions satisfy

$$u(t) \in U_{\text{ad}} \subset U, \quad v(t) \in V_{\text{ad}} \subset V, \quad t \in [0, t_f] \tag{3}$$

where X, U and V are Banach spaces, $X_{[0,t_f]}$, $U_{[0,t_f]}$ and $V_{[0,t_f]}$ are normed linear spaces, $A : D(A) \to X$ is the infinitesimal generator of a strongly continuous semigroup $\{T(t)\}_{t \geq 0}$ on X, $t_f \geq 0$ is a fixed real number, $f : [0, t_f] \times X \times U \times V \to X$, $\phi : X \to \mathbb{R}$ and $\psi : X \times U \to \mathbb{R}$ are given functions, U_{ad} is some subset of U and V_{ad} is a finite subset of V. This setting is for the most part classical, except for the assumptions on V_{ad}.

We will refer to the above infinite-dimensional dynamic optimization problem as mixed-integer optimal control problem, short (MIOCP), and to the control function $[u, v]$ as a mixed-integer control. This accounts for the fact that we do not impose restrictions on the set $U_{\text{ad}} \subset U$ while we can always identify the finite set $V_{\text{ad}} \subset V$ of the feasible control values for v with a finite number of integers

$$V_{\text{ad}} = \big\{v^1, \ldots, v^N\big\} \simeq \{1, \ldots, N\}. \tag{4}$$

Moreover, the operator differential equation (2) is an abstract representation of certain initial-boundary value problems governed by linear and semilinear partial differential equations, see, e.g., [20].

The existence of an optimal solution of the problem (MIOCP) depends, inter alia, on the spaces $X_{[0,t_f]}$, $U_{[0,t_f]}$ and $V_{[0,t_f]}$ where we seek $z : [0, t_f] \to X$, $u : [0, t_f] \to U$ and $v : [0, t_f] \to V$, respectively. Common choices are, for $U_{[0,t_f]}$, the spaces of square integrable (L^2), piecewise k-times differentiable (C_{pw}^k) or (piecewise) k-times weakly differentiable (H_{pw}^k) functions $u : [0, t_f] \to U$ and, for $V_{[0,t_f]}$, the spaces of essentially bounded (L^∞) or piecewise constant (PC) functions $v : [0, t_f] \to V$. We defer these considerations by assuming later that there exists an optimal solution of a related (to a certain extent convexified and relaxed) optimal control problem and present sufficient conditions guaranteeing that the solution of the relaxed problem can be approximated with arbitrary precision by a solution satisfying the integer restrictions. We do only assume that $X_{[0,t_f]} \subset C([0, t_f]; X)$, $U_{[0,t_f]} \subset L^1(0, t_f; U)$ and $V_{[0,t_f]} \subset L^1(0, t_f; V)$ to ensure that certain quantities in problem (MIOCP) are well-defined.

This relaxation method becomes most easily evident from writing problem (MIOCP) using a differential inclusion, that is, minimize (1) subject to the constraints that z is a solution of

$$\begin{cases} \dot{z}(t) \in Az(t) + \big\{f\big(t, z(t), u(t), v^i\big) : v^i \in V_{\text{ad}}\big\}, & t \in (0, t_f] \\ z(0) = z_0. \end{cases} \tag{5}$$

and u satisfies

$$u(t) \in U_{\mathrm{ad}}, \quad t \in [0, t_f].$$

It is well known that, under certain technical assumptions, the solution set of (5) is dense in the solution set of the convexified differential inclusion

$$\begin{cases} \dot{z}(t) \in Az(t) + \overline{\mathrm{co}}\{f(t, z(t), u(t), v^i) : v^i \in V_{\mathrm{ad}}\}, \quad t \in (0, t_f] \\ z(0) = z_0, \end{cases} \tag{6}$$

where $\overline{\mathrm{co}}$ denotes the closure of the convex hull. This is proved in [6] for the case when X is a separable Banach space and in [2] for non-separable Banach spaces. While these results rely on powerful selection theorems, our main contribution is a constructive proof based on discretization, giving rise to a numerical method at the prize of additional regularity assumptions.

We will see that the advantage of such a relaxation method is that the convexified problem, using a particular representation of (6), falls into the class of optimal control problems with partial differential equations without integer-restrictions. The already known theory, in particular concerning existence, uniqueness and regularity of optimal solutions as well as numerical considerations such as sensitivities, error analysis for finite element approximations, etc., can thus be carried over to the mixed-integer problem under consideration here. We also discuss the possibility to include state-constraints for example enforcing time-periodicity constraints

$$z(t_f) = z(0), \tag{7}$$

as occurring in chromatographic separation processes [5, 14]. The disadvantage of this approach is that we target at a solution that is only suboptimal (though with arbitrary precision) and that switching costs, a standard regularization of mixed-integer problems to prevent chattering solutions, or additional combinatorial constraints can lead to larger optimality/feasibility gaps. Nevertheless, we will show how a-priori bounds for such a gap can be obtained when constraints on the number of switches are incorporated.

The framework we use for the analysis here will be semigroup theory. Recall that for given $z_0 \in X$ and given control functions u, v, the mild solution of the state equation (2) is given by a function $z \in C(0, t_f; X)$ satisfying the variation of constants formula

$$z(t) = T(t)z_0 + \int_0^t T(t - s) f(s, z(s), u(s), v(s)) \, ds, \quad 0 \le t \le t_f \tag{8}$$

in the Lebesgue-Bochner sense. This abstract setting covers in particular the usual setup for weak solutions of linear parabolic partial differential equations with distributed control on reflexive Banach spaces where A arises from a time-invariant variational problem, see [1, Sect. 1.3].

We include in our analysis explicitly the possibility to approximate the state equation (2) and say that $\tilde{z} \in X_{[0,t_f]}$ is an ε-accurate solution of (2) if z is the mild solution

and $\|\tilde{z}(t) - z(t)\|_X \leq \varepsilon$ for all $t \in [0, t_f]$. Accordingly, we define for a given $\varepsilon \geq 0$ the set of ε-admissible solutions for (MIOCP) as

$$\Xi_\varepsilon := \big\{ (u, v, z) \in U_{[0,t_f]} \times V_{[0,t_f]} \times X_{[0,t_f]} : u, v \text{ satisfy (3)}$$

$$\text{and } z = z(u, v) \text{ is an } \varepsilon\text{-accurate solution of (2)} \big\}. \qquad (9)$$

Further, we denote throughout the paper by $H^1_{\mathrm{pw}}(0, t_f; X)$ the space of X-valued functions defined on the interval $[0, t_f]$ and being piecewise once-weakly differentiable with a piecewise defined weak derivative that is square-integrable in the Lebesgue-Bochner sense. Consistently, we denote by $C^{0,\vartheta}_{\mathrm{pw}}(0, t_f; X)$ the space of X-valued functions defined on the interval $[0, t_f]$ being piecewise Hölder-continuous with a Hölder-constant ϑ. In both constructions, piecewise means that there exists a finite partition of the interval $[0, t_f]$

$$0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_{K_{t_f}} < \tau_{K_{t_f}+1} = t_f \qquad (10)$$

so that the function has the respective regularity on all intervals $[\tau_k, \tau_{k+1})$, $k = 0, \ldots, K_{t_f}$. We denote by $\| \cdot \|_X$ the norm on X and by $\| \cdot \|_{\mathcal{L}(X)}$ the operator norm induced by $\| \cdot \|_X$. Further, we denote by $\| \cdot \|_{U_{[0,t_f]}}$ and $\| \cdot \|_{V_{[0,t_f]}}$ the norm of $U_{[0,t_f]}$ and $V_{[0,t_f]}$, respectively. For simplicity of notation, we also define $T(-t) = \mathrm{Id}$ for all $t > 0$, Id denoting the identity on X.

The paper is organized as follows. In Sect. 2, we present details of the relaxation method and the main results concerning estimates of the approximation error. In Sect. 3, we discuss extensions of the method to incorporate certain combinatorial constraints. In Sect. 4, we discuss applications for linear and semilinear equations and present numerical results for the heat equation with spatial scheduling of different actuators and a semilinear reaction-diffusion system with an on-off type control. In Sect. 5, we conclude with some additional remarks and point out open problems.

2 Relaxation method

Consider the following problem involving a particular representation of the convexified differential inclusion (6)

$$\left\{ \begin{array}{ll} \displaystyle\min_{[\omega,\alpha,y] \in U_{[0,t_f]} \times \tilde{V}_{[0,t_f]} \times X_{[0,t_f]}} J = \phi\big(y(t_f)\big) + \int_0^{t_f} \psi\big(y(t), \omega(t)\big)\, dt \quad \text{s.t.} & (11a) \\[4mm] \dot{y}(t) = Ay(t) + \displaystyle\sum_{i=1}^N \alpha_i(t) f\big(t, y(t), \omega(t), v^i\big), \quad t \in (0, t_f] & (11b) \\[4mm] y(0) = y_0 := z_0 & (11c) \\[1mm] \omega(t) \in U_{\mathrm{ad}}, \quad t \in [0, t_f] & (11d) \\[1mm] \alpha(t) = \big(\alpha_1(t), \ldots, \alpha_N(t)\big) \in [0,1]^N, \quad t \in [0, t_f] & (11e) \\[4mm] \displaystyle\sum_{i=1}^N \alpha_i(t) = 1, \quad t \in [0, t_f], & (11f) \end{array} \right.$$

Algorithm 1 Relaxation method

1: Choose a time discretization grid $\mathcal{G}^0 = \{0 = t_0^0 < t_1^0 < \cdots < t_{n^0}^0 = t_f\}$, a sequence of non-negative accuracies $\{\varepsilon^k\}_{k \in \mathbb{N}}$ and some termination tolerance $\varepsilon > 0$. Set $k = 0$.
2: LOOP
3:    Find an ε^k-accurate optimal solution $[\omega^k, \alpha^k, y^k]$ of the relaxed problem (11a)–(11f) and set $J_{\text{rel}}^k := J(\omega^k, \alpha^k, y^k)$.
4:    If $\alpha^k \in PC(0, t_f; \{0, 1\}^N)$ and $\varepsilon^k \leq \varepsilon$, then set $v^k(t) = \sum_{i=1}^N \alpha_i^k(t) v^i$ and $z^k(t) = y^k(t)$ for $t \in [0, t_f]$ and STOP.
5:    Using $\mathcal{G}^k$ and α^k, define a piecewise constant function $\beta^k = (\beta_1^k, \ldots, \beta_N^k) : [0, t_f] \to \{0, 1\}^N$ by

$$\beta_i^k(t) = p_{i,j}^k, \quad t \in [t_j^k, t_{j+1}^k), \quad i = 1, \ldots, N, \quad j = 0, \ldots, n^k - 1 \quad (A_1)$$

   where for all $i = 1, \ldots, N, j = 0, \ldots, n^k - 1$

$$p_{i,j}^k = \begin{cases} 1 & \text{if } (\hat{p}_{i,j}^k \geq \hat{p}_{l,j}^k \ \forall l \in \{1, \ldots, N\} \setminus \{i\}) \quad \text{and} \\ & (i < l \ \forall l \in \{1, \ldots, N\} \setminus \{i\} : \hat{p}_{i,j}^k = \hat{p}_{l,j}^k) \\ 0 & \text{else} \end{cases} \quad (A_2)$$

$$\hat{p}_{i,j}^k = \int_0^{t_{j+1}^k} \alpha_i^k(\tau) \, d\tau - \sum_{l=0}^{j-1} p_{i,l}^k (t_{l+1}^k - t_l^k).$$

6:    Set $u^k(t) = \omega^k(t)$, $v^k(t) = \sum_{i=1}^N \beta_i^k(t) v^i$ for $t \in [0, t_f]$ and $J^k = \phi(z^k(t_f)) + \int_0^{t_f} \psi(z^k(t), \omega^k(t)) \, dt$ where z^k is an ε^k-accurate solution of

$$\dot{z}(t) = Az(t) + \sum_{i=1}^N \beta_i^k(t) f\big(t, z(t), \omega^k(t), v^i\big), \quad t \in (0, t_f], \quad z(0) = z_0. \quad (A_3)$$

7:    If $|J_{\text{rel}}^k - J^k| \leq \frac{\varepsilon}{2}$ and $\varepsilon^k \leq \frac{\varepsilon}{2}$ then STOP.
8:    Choose $\mathcal{G}^{k+1} = \{0 = t_0^{k+1} < t_1^{k+1} < \cdots < t_{n^{k+1}}^{k+1} = t_f\}$ such that $\mathcal{G}^k \subset \mathcal{G}^{k+1}$ and set $k = k + 1$.
9: END LOOP
10: Set $u^*(t) = \omega^k(t)$, $v^*(t) = v^k(t)$ and $z^*(t) = z^k(t)$ for $t \in [0, t_f]$.

where we write ω, α, y for u, v, z, respectively, to emphasize the relaxation of the original problem (MIOCP). Here, $\tilde{V}_{[0,t_f]} = \{\alpha : [0, t_f] \to \mathbb{R}^N : \|\alpha\|_{V_{[0,t_f]}} < \infty\}$.

Observe that the control functions α_i take values on the full interval $[0, 1]$, but that any optimal controls $[\omega^*, \alpha^*]$ of (11a)–(11f) yields an optimal mixed-integer control

$$[u^*, v^*] := \left[\omega^*, \sum_{i=1}^N \alpha_i^* v^i\right] \quad (12)$$

of problem (MIOCP) if $\alpha^*(t) \in \{0, 1\}^N$ for almost every $t \in (0, t_f)$. However, it is not very difficult to construct examples where $\alpha^*(t) \in (0, 1)^N$ for t on some interval of positive measure. It is only in some special cases where optimality of $\alpha^*(t) \in [0, 1]^N$ implies that α^* takes only values on the boundary of its feasible set. For examples where this property, known as the bang-bang principle, can be verified in the context of partial differential equations, see [28, Sect. 3.2.4] and the references therein. Moreover, the relaxed problem (11a)–(11f) can in most applications only be solved approximately.

Therefore, consider the following hypothesis.

(H_0) Problem (11a)–(11f) has an optimal solution in $U_{[0,t_f]} \times \tilde{V}_{[0,t_f]} \times X_{[0,t_f]}$.

Under this assumption, we will accept ε-accurate optimal solutions of the relaxed problem (11a)–(11f) and propose in Algorithm 1 an iterative procedure to obtain from these solutions a mixed-integer control taking values in $U_{\mathrm{ad}} \times \{0, 1\}^N$. We then show in Theorem 1 under certain technical assumptions that the optimal value and the optimal state of problem (11a)–(11f) can be approximated by the proposed procedure with arbitrary precision. To make our terminology precise, we include the following definition.

Definition 1 Given some $\varepsilon \geq 0$, we say that $[\omega^\varepsilon, \alpha^\varepsilon, y^\varepsilon] \in U_{[0,t_f]} \times \tilde{V}_{[0,t_f]} \times X_{[0,t_f]}$ is an ε-*accurate optimal solution of* (11a)–(11f) if y^ε is an ε-accurate solution of (11b)–(11c) with $\omega = \omega^\varepsilon$ and $\alpha = \alpha^\varepsilon$, the constraints (11d)–(11f) hold with $\omega = \omega^\varepsilon$ and $\alpha = \alpha^\varepsilon$ and

$$J\left(\omega^\varepsilon, \alpha^\varepsilon, y^\varepsilon\right) \leq \inf_{\omega,\alpha,y \text{ s.t. } (11b)–(11f)} J(\omega, \alpha, y) + \varepsilon. \tag{13}$$

So any ε-accurate optimal solution of the relaxed problem (11a)–(11f) is admissible with respect to the control constraints on α and ω, ε-close admissible with respect to the state variable y and the corresponding value of the cost function is ε-close to the optimal value of problem (11a)–(11f).

Now, consider Algorithm 1 on p. 202 to obtain a mixed-integer control for problem (MIOCP). The main result is the following.

Theorem 1 *Assuming* (H_0), *let* $[\omega^*, \alpha^*, y^*]$ *denote an optimal solution of the relaxed problem* (11a)–(11f) *and assume that the following assumptions hold true.*

(H_1) *The functions* ϕ, ψ, *and* f *satisfy the Lipschitz-estimates*

$$\left|\phi(y_1) - \phi(y_2)\right| \leq \eta \|y_1 - y_2\|_X,$$

$$\left|\psi(y_1, \omega_1) - \psi(y_2, \omega_2)\right| \leq \xi\left(\|y_1 - y_2\|_X + \|\omega_1 - \omega_2\|_U\right),$$

$$\left\|f\left(t, y_1, \omega_1, v^i\right) - f\left(t, y_2, \omega_2, v^i\right)\right\|_X \leq L\left(\|y_1 - y_2\|_X + \|\omega_1 - \omega_2\|_U\right),$$

for all $y_1, y_2 \in X, \omega_1, \omega_2 \in U_{\mathrm{ad}}, t \in [0, t_f]$ *and* $i = 1, \ldots, N$ *with positive constants* η, ξ *and* L.

(H$_2$) *For all $i = 1, \ldots, N$ and $t \in [0, t_f]$ the function*

$$s \mapsto T(t - s) f\left(s, y^*(s), \omega^*(s), v^i\right)$$

is in $H^1_{pw}(0, t_f; X)$ and there exists a positive constant C_i such that

$$\left\| \frac{d}{ds} \left(T(t - s) f\left(s, y^*(s), \omega^*(s), v^i\right) \right) \right\|_X \leq C_i \quad a.e. \text{ in } 0 < s < t < t_f.$$

Let $C = \sum_{i=1}^N C_i$.

(H$_3$) *For all $i = 1, \ldots, N$, there exists a positive constant M_i such that*

$$\sup_{t \in [0, t_f]} \left\| f\left(t, y^*(t), \omega^*(t), v^i\right) \right\|_X \leq M_i.$$

Let $M = \sum_{i=1}^N M_i$.

(H$_4$) *The solution $[\omega^*, \alpha^*, y^*]$ is stable in the sense that there exists a positive constant C_J such that*

$$\left\| \omega^*(t) - \omega(t) \right\|_U + \left\| \alpha^*(t) - \alpha(t) \right\|_{\mathbb{R}^N}$$
$$\leq C_J \left| J\left(\omega^*, \alpha^*, y\left(\omega^*, \alpha^*\right)\right) - J\left(\omega, \alpha, y(\omega, \alpha)\right) \right| \quad a.e. \text{ in } (0, t_f) \quad (14)$$

for $[\omega, \alpha]$ in some neighborhood of $[u^*, \alpha^*]$ in $U_{[0,t_f]} \times \tilde{V}_{[0,t_f]}$.

Moreover, assume that $\varepsilon^k \to 0$ and that the sequence $\{\mathcal{G}^k\}_k$ in Algorithm 1 is such that $\Delta t^k \to 0$ with

$$\Delta t^k = \max_{i=1,\ldots,n^k} \left\{ t_i^k - t_{i-1}^k \right\}. \quad (15)$$

Define the constants $C_1 = (C_J(M + 1)(1 + \eta + \xi)e^{t_f \bar{M} L} + 1)$, $C_2 = (M + t_f C)e^{t_f \bar{M} L}$, $C_3 := (\eta + t_f \xi)C_1 + t_f \xi C_J$ and $C_4 := (\eta + t_f \xi)C_2$, where $\bar{M} = \sup_{t \in [0, t_f]} \|T(t)\|_{\mathcal{L}(X)}$ and the constants η, ξ, M, C, L and C_J are given by hypothesis (H$_1$)–(H$_4$). Then, $[u^k, v^k, z^k]$ defined by Algorithm 1 is in Ξ_{ε^k} for all $k = 0, 1, 2, \ldots$ and satisfies the estimates

$$\left\| y^*(t) - z^k(t) \right\|_X \leq C_1 \varepsilon^k + C_2(N - 1)\Delta t^k, \quad t \in [0, t_f], \quad (16)$$

and

$$\left| J\left(\omega^*, \alpha^*, y^*\right) - J\left(u^k, v^k, z^k\right) \right| \leq C_3 \varepsilon^k + C_4(N - 1)\Delta t^k. \quad (17)$$

In particular, Algorithm 1 terminates in a finite number of steps with an ε-feasible mixed-integer solution $[u^*, v^*, z^*] \in \Xi_\varepsilon$ of Problem (MIOCP) satisfying the estimate

$$\left| J\left(\omega^*, \alpha^*, y^*\right) - J\left(u^*, v^*, z^*\right) \right| \leq \varepsilon, \quad (18)$$

where $\varepsilon > 0$ is chosen arbitrarily in step 1.

Before we prove Theorem 1, we first prove a result saying that the deviation of two mild solutions in $X_{[0,t_f]}$ equipped with the uniform norm can be estimated in terms of the absolute value of the integrated difference of two linearly entering control functions. This estimate is non-standard and generalizes the result in [23] to a Banach space setting, noting that the absolute value of the integrated difference is not a norm and in particular not comparable to the $V_{[0,t_f]}$-norm as a natural choice. This estimate, together with an approximation result for this integrated difference is the key ingredient in order to prove all a-priori estimates needed for the proof of Theorem 1.

Lemma 1 *Let $\varepsilon > 0$ and $\bar{M} = \sup_{t \in [0,t_f]} \|T(t)\|_{\mathcal{L}(X)}$. Suppose that $[\omega^*, \alpha^*, y^*]$ is a feasible solution of the relaxed problem* (11a)–(11f) *and assume that the hypotheses* (H$_1$)–(H$_3$) *of Theorem* 1 *hold true. Let $\beta = (\beta_1, \dots, \beta_N) \in L^\infty(0, t_f; [0, 1]^N)$ be such that*

$$\max_{i=1,\dots,N} \sup_{t \in [0,t_f]} \left| \int_0^t \alpha_i^*(\tau) - \beta_i(\tau) \, d\tau \right| \leq \varepsilon \tag{19}$$

and let z be the mild solution of (A$_3$) *in Algorithm* 1 *with $\beta_i^k = \beta_i$, $i = 1, \dots, N$, and $\omega^k = \omega^*$. Then*

$$\|y^*(t) - z(t)\|_X \leq \left((M + Ct)e^{\bar{M}Lt} \right) \varepsilon, \quad t \in [0, t_f]. \tag{20}$$

Proof Fix $t \in [0, t_f]$ and set, for the sake of brevity, $\delta(t) = \|y^*(t) - z(t)\|_X$ and $f^i(t, y(t)) = f(t, y(t), \omega^*(t), v^i)$. Recalling (10) and using hypothesis (H$_2$), let $\{\tau_0, \tau_1, \dots, \tau_{K+1}\}$ be the set of partition points of the functions $s \mapsto T(t - s)f^i(s, y^*(s))$ as objects in $H_{\mathrm{pw}}^1(0, t; X)$, $i = 1, \dots, N$, so that $\tau_0 = 0$ and $\tau_{K+1} = t$. From the definition of the mild solutions for (11a)–(11f) and (A$_3$), we have

$$\delta(t) = \left\| \sum_{i=1}^N \int_0^t T(t-s)f^i(s, y^*(s))\alpha_i^*(s) - T(t-s)f^i(s, z(s))\beta_i(s) \, ds \right\|_X.$$

Adding $0 = T(t-s)f^i(s, y^*(s))\beta_i(s) - T(t-s)f^i(s, y^*(s))\beta_i(s)$ under the integral, applying the triangular inequality and rearranging terms this yields

$$\delta(t) \leq \sum_{i=1}^N \left\| \int_0^t T(t-s)\left[f^i(s, y^*(s)) - f^i(s, z(s))\right]\beta_i(s) \, ds \right\|_X$$

$$+ \sum_{i=1}^N \left\| \sum_{k=0}^K \int_{\tau_k}^{\tau_{k+1}} T(t-s)f^i(s, y^*(s))[\alpha_i^*(s) - \beta_i(s)] ds \right\|_X.$$

Now using integration by parts in the second part, we obtain

$$\delta(t) \leq \sum_{i=1}^N \left\| \int_0^t T(t-s)\left[f^i(s, y^*(s)) - f^i(s, z(s))\right]\beta_i(s) \, ds \right\|_X$$

$$+ \sum_{i=1}^{N} \left\| \sum_{k=0}^{K} (T(t - \tau_{k+1}) f^i (\tau_{k+1}, y^*(\tau_{k+1})) \int_0^{\tau_{k+1}} \alpha_i^*(s) - \beta_i(s) \, ds \right.$$

$$- T(t - \tau_k) f^i (\tau_k, y^*(\tau_k)) \int_0^{\tau_k} \alpha_i^*(s) - \beta_i(s) \, ds$$

$$\left. - \int_{\tau_k}^{\tau_{k+1}} \frac{d}{ds} (T(t - s) f^i (s, y^*(s))) \int_0^s \alpha_i^*(\vartheta) - \beta_i(\vartheta) \, d\vartheta \, ds) \right\|_X .$$

Then by rearranging terms, noting that the appearing telescopic sum evaluates as

$$\sum_{k=0}^{K} (T(t - \tau_{k+1}) f^i (\tau_{k+1}, y^*(\tau_{k+1})) \int_0^{\tau_{k+1}} \alpha_i^*(s) - \beta_i(s) \, ds$$

$$- T(t - \tau_k) f^i (\tau_k, y^*(\tau_k)) \int_0^{\tau_k} \alpha_i^*(s) - \beta_i(s) \, ds)$$

$$= f^i (t, y^*(t)) \int_0^t \alpha_i^*(s) - \beta_i(s) \, ds$$

because of $\tau_0 = 0$, $\tau_{K+1} = t$, $T(t - t) = \text{Id}$ and $\int_0^0 \alpha_i^*(\vartheta) - \beta_i(\vartheta) \, d\vartheta = 0$, and by applying the triangular inequality this estimate simplifies to

$$\delta(t) \leq \sum_{i=1}^{N} \int_0^t \|T(t - s)\|_{\mathcal{L}(X)} \|f^i (s, y^*(s)) - f^i (s, z(s))\|_X |\beta_i(s)| \, ds$$

$$+ \sum_{i=1}^{N} \|f^i (t, y^*(t))\|_X \left| \int_0^t \alpha_i^*(s) - \beta_i(s) \right| ds$$

$$+ \sum_{i=1}^{N} \int_0^t \left\| \frac{d}{ds} (T(t - s) f^i (s, y^*(s))) \right\|_X \left| \int_0^s \alpha_i^*(\vartheta) - \beta_i(\vartheta) \, d\vartheta \right| ds.$$

Then, by definition of δ, f^i and the constant $\bar{M}$, the definition of the constants L, C and M in hypotheses (H$_1$)–(H$_3$), assumption (19) and the fact that $\beta_i(t) \leq 1$, this yields

$$\delta(t) \leq \bar{M} L \int_0^t \delta(s) \, ds + M\varepsilon + Ct\varepsilon.$$

Finally, using the Gronwall lemma and rearranging terms, we obtain the desired estimate

$$\delta(t) \leq ((M + Ct)e^{\bar{M}Lt})\varepsilon. \qquad \square$$

Next, we recall from [23] the following result on integral approximations.

Lemma 2 *Let $\alpha = (\alpha_1, \ldots, \alpha_N) : [0, t_f] \to [0, 1]^N$ be a measurable function satisfying $\sum_{i=1}^{N} \alpha_i(t) = 1$ for all $t \in [0, t_f]$. Define a piecewise constant function*

$\beta : [0, t_f] \to \{0, 1\}^N$ by

$$\beta_i(t) = p_{i,j}, \quad t \in [t_j, t_{j+1}), \quad i = 1, \ldots, N, \quad j = 0, \ldots, n-1 \quad (21)$$

where for all $i = 1, \ldots, N$, $j = 0, \ldots, n-1$, $p_{i,j}$ is defined by (A_2) in Algorithm 1 with $p_{i,j}^k = p_{i,j}$. Then it holds for $\Delta t = \max_{i=1,\ldots,n}\{t_i - t_{i-1}\}$

1. $\max_{i=1,\ldots,N} |\int_0^t \alpha_i(\tau) - \beta_i(\tau)\,d\tau| \le (N-1)\Delta t$ for all $t \in [0, t_f]$,
2. $\sum_{i=1}^N \beta_i(t) = 1$ for all $t \in [0, t_f]$.

Proof See Theorem 5 of [23]. $\qquad\square$

With the above two results we are now in the position to prove Theorem 1.

Proof of Theorem 1 Let the assumptions of Theorem 1 hold true. First we show that the sequence $[u^k, v^k, z^k]$ obtained by Algorithm 1 is ε^k-feasible for the problem (MIOCP). We have $u^k = \omega^k \in U_{\mathrm{ad}}$ for all k by construction of ω^k in step 3, $v^k(t) = \sum_{i=1}^N \beta_i^k(t)v^i, t \in [0, t_f]$, by construction in step 6 so $v^k(t) \in V_{\mathrm{ad}}, t \in [0, t_f]$, because $\beta_i^k(t) \in \{0, 1\}$ for all i, k and $t \in [0, t_f]$ as seen from (A_1) and (A_2). By construction in step 6, z^k is an ε^k-accurate solution of (A_3). Thus, recalling (9), $[u^k, v^k, z^k] \in \Xi_{\varepsilon^k}$ for all $k = 0, 1, 2, \ldots$.

Next, we show (16). To this end, let $y(\cdot; \omega^k, \beta^k)$, $z(\cdot; \omega^*, \beta^k) = y(\cdot; \omega^*, \beta^k)$ and $y(\cdot; \omega^*, \alpha^k)$ denote the mild solutions of (11b), (11c) with the respective controls. The stability assumption (H_4) and the continuity assumption (H_1) implies that

$$\begin{aligned}
&\left\| \alpha^*(t) - \alpha^k(t) \right\|_{\mathbb{R}^N} + \left\| \omega^*(t) - \omega^k(t) \right\|_U \\
&\le C_J \left| J(\omega^*, \alpha^*, y(\omega^*, \alpha^*)) - J(\omega^k, \alpha^k, y(\omega^k, \alpha^k)) \right| \\
&\le C_J \big(\left| J(\omega^*, \alpha^*, y(\omega^*, \alpha^*)) - J(\omega^k, \alpha^k, y^k) \right| \\
&\qquad + \left| J(\omega^k, \alpha^k, y^k) - J(\omega^k, \alpha^k, y(\omega^k, \alpha^k)) \right| \big) \\
&\le C_J \big(\varepsilon^k + \left| \phi(y^k(t_f)) - \phi(y(t_f; \omega^k, \alpha^k)) \right| \\
&\qquad + \int_0^{t_f} \left| \psi(y^k(t), \omega^k(t)) - \psi(y(t; \omega^k, \alpha^k), \omega^k(t)) \right| dt \big) \\
&\le C_J (1 + \eta + t_f \xi) \varepsilon^k
\end{aligned} \quad (22)$$

for a.e. $t \in (0, t_f)$, where we added $0 = -J(\omega^k, \alpha^k, y^k) + J(\omega^k, \alpha^k, y^k)$, used the triangular inequality and that $[\omega^k, \alpha^k, y^k]$ is an ε-accurate optimal solution of (11a)–(11f). Moreover, by fixing $t \in [0, t_f]$, adding $0 = -y(t; \omega^k, \beta^k) + y(t; \omega^k, \beta^k)$, $0 = -z(t; \omega^*, \beta^k) + y(t; \omega^*, \beta^k)$ and $0 = -y(t; \omega^*, \alpha^k) + y(t; \omega^*, \alpha^k)$ and using the triangular inequality, we see that

$$\left\| z^k(t) - y^*(t) \right\|_X \le \delta_1(t) + \delta_2(t) + \delta_3(t) + \delta_4(t) \quad (23)$$

with

$$\delta_1(t) = \left\| z^k(t) - y\big(t; \omega^k, \beta^k\big) \right\|_X, \qquad \delta_2(t) = \left\| y\big(t; \omega^k, \beta^k\big) - z\big(t; \omega^*, \beta^k\big) \right\|_X$$

$$\delta_3(t) = \left\| z\big(t; \omega^*, \beta^k\big) - y\big(t; \omega^*, \alpha^k\big) \right\|_X, \qquad \delta_4(t) = \left\| y\big(t; \omega^*, \alpha^k\big) - y^*(t) \right\|_X.$$

Observe that $\delta_1(t) \le \varepsilon^k$, because z^k is an ε^k-accurate solution of (A$_3$) and thus of (11b), (11c) with controls ω^k, α^k. By definition of the mild solution and using (H$_1$), we have

$$\delta_2(t) \le \sum_{i=1}^{N} \int_0^t \left\| T(t-s) \right\|_{\mathcal{L}(X)} \left\| \beta_i^k(s) \right\| \left\| \big(f\big(s, y(s; \omega^k, \beta^k), \omega^k(s), v^i \big) \right.$$

$$\left. - f\big(s, z\big(s; \omega^*, \beta^k\big), \omega^*(s), v^i\big) \big) \right\|_X ds$$

$$\le \bar{M}L \int_0^t \delta_2(s) + \left\| \omega^k(s) - \omega^*(s) \right\|_U ds. \tag{24}$$

Using (22), the Gronwall inequality implies that

$$\delta_2(t) \le C_J (1 + \eta + \xi) e^{\bar{M}Lt} \varepsilon^k. \tag{25}$$

From Lemma 2 with $\alpha = \alpha^k$, $\beta = \beta^k$ and $\Delta t = \Delta t^k$ with Δt^k from (15), we get that

$$\max_{i=1,\dots,N} \left| \int_0^t \alpha_i^k(\tau) - \beta_i^k(\tau) \, d\tau \right| \le (N-1)\Delta t^k, \quad t \in [0, t_f]. \tag{26}$$

Moreover, Lemma 1 used with $y^* = y(\cdot; \omega^*, \alpha^k)$, $\alpha^* = \alpha^k$, $\beta = \beta^k$ and $\varepsilon := (N-1)\Delta t^k$ implies that

$$\delta_3(t) = \left\| z\big(t; \omega^*, \beta^k\big) - y\big(t; \omega^*, \alpha^k\big) \right\|_X \le \big((M + Ct) e^{\bar{M}Lt} \big)(N-1)\Delta t^k. \tag{27}$$

Again by definition of the mild solution we have

$$\delta_4(t) \le \sum_{i=1}^{N} \int_0^t \left\| T(t-s) \right\|_{\mathcal{L}(X)} \left\| \alpha_i^k(s) f\big(s, y\big(s, \omega^*, \alpha^k\big), \omega^*(s), v^i \big) \right.$$

$$\left. - \alpha_i^*(s) f\big(s, y\big(s; \omega^*, \alpha^*\big), \omega^*(s), v^i\big) \right\|_X ds. \tag{28}$$

Adding $0 = \alpha^k(s)(-f(s, y(s; \omega^*, \alpha^*), \omega^*(s), v^i) + f(s, y(s; \omega^*, \alpha^*), \omega^*(s), v^i))$ under the integral, we obtain

$$\delta_4(t) \le \bar{M} \sum_{i=1}^{N} \int_0^t \left\| \alpha_i^k(s) \big[f\big(s, y\big(s; \omega^*, \alpha^k\big), \omega^*(s), v^i \big) \right.$$

$$\left. - f\big(s, y\big(s, \omega^*, \alpha^*\big), \omega^*(s), v^i\big) \big] \right.$$

$$+ \left[\alpha_i^k(s) - \alpha_i^*(s) \right] f\left(s, y\left(s; \omega^*, \alpha^*\right), \omega^*(s), v^i \right) \big\|_X ds$$

$$\leq \bar{M} L \int_0^t \delta_4(s) + M \left| \alpha_i^k(s) - \alpha_i^*(s) \right| ds. \tag{29}$$

Using again (22) and the Gronwall inequality we obtain that

$$\delta_4(t) \leq M C_J (1 + \eta + \xi) e^{\bar{M} L t} \varepsilon^k. \tag{30}$$

Thus, summing up the estimates for $\delta_1(t), \ldots, \delta_4(t)$ and rearranging terms we obtain from (23) that for all $t \in [0, t_f]$

$$\left\| z^k(t) - y^*(t) \right\|_X \leq C_1 \varepsilon^k + C_2 (N - 1) \Delta t^k \tag{31}$$

with $C_1 = (C_J (M+1)(1+\eta+\xi)e^{t_f \bar{M} L} + 1)$ and $C_2 = (M + t_f C)e^{t_f \bar{M} L}$. This proves (16).

By definition of the cost function (1) we get from the triangular inequality that

$$\left| J\left(\omega^*, \alpha^*, y^*\right) - J\left(u^k, v^k, z^k\right) \right|$$
$$\leq \left| \phi\left(y^*(t)\right) - \phi\left(z^k(t)\right) \right|$$
$$+ \int_0^{t_f} \left| \psi\left(y^*(t), \omega^*(t)\right) - \psi\left(z^k(t), \omega^k(t)\right) \right| dt \tag{32}$$

so that using hypothesis (H$_1$), (22) and (31) we obtain

$$\left| J\left(\omega^*, \alpha^*, y^*\right) - J\left(u^k, v^k, z^k\right) \right| \leq C_3 \varepsilon^k + C_4 (N - 1) \Delta t^k \tag{33}$$

with the constants $C_3 = (\eta + t_f \xi)C_1 + t_f \xi C_J$ and $C_4 = (\eta + t_f \xi)C_2(t_f)$. This proves (17).

Next, suppose that the main loop in Algorithm 1 terminates in step 4 or in step 7. In the first case, the termination criterion implies that

$$\left| J\left(u^*, v^*, z^*\right) - J\left(\omega^*, \alpha^*, y^*\right) \right| = \left| J\left(\omega^k, \alpha^k, y^k\right) - J\left(\omega^*, \alpha^*, y^*\right) \right| \leq \varepsilon^k \leq \varepsilon \tag{34}$$

for some k, because $J(\omega^k, \alpha^k, z^k)$ is an ε^k-accurate optimal solution of (11a)–(11f). Similarly, in the second case the termination criterion implies that

$$\left| J\left(u^*, v^*, z^*\right) - J\left(\omega^*, \alpha^*, y^*\right) \right|$$
$$\leq \left| J\left(u^*, v^*, z^*\right) - J\left(\omega^k, \alpha^k, y^k\right) \right| + \left| J\left(\omega^k, \alpha^k, y^k\right) - J\left(\omega^*, \alpha^*, y^*\right) \right|$$
$$\leq \left| J^k - J_{\text{rel}}^k \right| + \varepsilon^k \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \leq \varepsilon \tag{35}$$

for some k. This proves (18) under the assumption that Algorithm 1 terminates.

Finally suppose that Algorithm 1 loops infinitely many times, that is,

$$\left| J_{\text{rel}}^k - J^k \right| > \frac{\varepsilon}{2} \quad \text{or} \quad \varepsilon^k > \frac{\varepsilon}{2} \quad \text{for all } k = 0, 1, 2, \ldots \tag{36}$$

By adding $0 = -y^*(t) + y^*(t)$ and using that y^k is an ε^k-accurate optimal solution, we obtain from (31) that

$$\left\| z^k(t) - y^k(t) \right\|_X \le \left\| z^k(t) - y^*(t) \right\|_X + \left\| y^*(t) - y^k(t) \right\|_X$$

$$\le C_1 \varepsilon^k + C_2(N-1)\Delta t^k + \varepsilon^k. \tag{37}$$

Using that $\varepsilon^k \to 0$ and $\Delta t^k \to 0$ as $k \to \infty$ by assumption, we see from (37) that $\sup_{t \in [0,t_f]} \| z^k(t) - y^k(t) \|_X \to 0$ as $k \to \infty$. By definition of J_{rel}^k and J^k and using the triangular inequality we have

$$\left| J_{\mathrm{rel}}^k - J^k \right| = \left| J(\omega^k, \alpha^k, y^k) - J(\omega^k, \beta^k, z^k) \right| \le \left| \phi(y^k(t_f)) - \phi(z^k(t_f)) \right|$$

$$+ \int_0^{t_f} \left| \psi(y^k(t), \omega^k(t)) - \psi(z^k(t), \omega^k(t)) \right| dt, \tag{38}$$

so that $\sup_{t \in [0,t_f]} \| z^k(t) - y^k(t) \|_X \to 0$ as $k \to \infty$ implies by continuity of ϕ and ψ that also $|J_{\mathrm{rel}}^k - J^k| \to 0$ as $k \to \infty$. Together with the assumption that $\varepsilon^k \to 0$ as $k \to \infty$ this contradicts (36) and completes the proof. □

   Theorem 1 can be seen as a performance analysis of the relaxation method proposed in Algorithm 1. The estimates (16) and (17) prove a bilinear dependency of the mixed-integer control approximation error for the differential state and the optimal value in terms of the chosen maximal integer-control discretization mesh size Δt^k and accuracy ε^k. This relates to the convergence speed of Algorithm 1 in terms of the chosen refinements for Δt^k and ε^k. Note that the estimates (16) and (17) suggest to choose Δt^k and ε^k of the same order. On the other hand, Theorem 1 can be regarded as an existence result of suboptimal solutions for (MIOCP). To emphasize this, we formulate the precise statement explicitly.

Corollary 1 *Under the hypothesis* (H$_0$)–(H$_3$) *there exists for every $\varepsilon > 0$ a feasible solution* $(u^\varepsilon, v^\varepsilon, z^\varepsilon) \in U_{[0,t_f]} \times V_{[0,t_f]} \times X_{[0,t_f]}$ *of problem* (MIOCP) *satisfying*

$$J(u^\varepsilon, v^\varepsilon, z^\varepsilon) \le J(\omega^*, \alpha^*, y^*) + \varepsilon, \tag{39}$$

where $(\omega^*, \alpha^*, y^*)$ is the optimal solution of the relaxed problem (11a)–(11f).

Proof Apply Theorem 1 with $\varepsilon^k = 0$ for all k. Then, $(\omega^k, \alpha^k, y^k) = (\omega^*, \alpha^*, y^*)$ and it can be seen from proof of Theorem 1 that (H$_4$) is then not needed for the estimate (17). Moreover, Ξ_0 is contained in the feasible set of problem (11a)–(11f) and thus $J(\omega^*, \alpha^*, y^*) \le J(u, v, z)$ for all $(u, v, z) \in \Xi_0$. Thus, (39) follows from (18). □

   Hypothesis (H$_0$) can be checked by classical arguments, cf., e.g., [28]. Hypothesis (H$_1$)–(H$_3$) are needed to estimate the proximity of mixed-integer solutions of (MIOCP) to the optimal solution of the relaxed problem (11a)–(11f) while hypothesis (H$_4$) guarantees in a sense the proximity of the ε-accurate optimal solutions of (11a)–(11f) to the optimal ones. Hypothesis (H$_1$) and (H$_3$) are standard assumptions and can be weakened to appropriate 'local' versions using standard arguments. This

can also tighten the estimates in Theorem 1. The last conclusion of Theorem 1 concerning the termination of Algorithm 1 and Corollary 1 even hold for the functions ϕ and ψ just continuous as it can be seen from the respective proofs. Hypothesis (H$_4$) can be verified for problems that are well-posed in the Tikhonov sense, cf. the discussion in [15, Sect. 4]. For sufficiently regularized parabolic problems it could also be checked using methods as in [19]. Alternatively, instead of invoking Theorem 1, Corollary 1 or a weaker conclusion presented in Proposition 2 below may be used where (H$_4$) is not needed.

Hypothesis (H$_2$) of Theorem 1 clearly imposes certain regularity assumptions on the linear operator A generating the semigroup $\{T(t)\}_{t\geq 0}$, the function f, but also on the time regularities of the optimal control functions of the relaxed problem (11a)–(11f) in $U_{[0,t_f]}$ and $\tilde{V}_{[0,t_f]}$. The main difficulty with proving (H$_2$) is that y^* as a solution of (8) with A unbounded may only be continuous and not absolutely continuous in time, hence not necessarily differentiable almost everywhere as this is always true when $A = 0$ (with $T(\cdot) = \mathrm{Id}$) and X is a finite dimensional space. This can be delicate in particular for nonlinear systems. We will therefore exemplary discuss hypothesis (H$_2$) in Example 2 below for the case of a semilinear system where A is the generator of an analytic semigroup. A more general analysis is possible for linear systems

$$\dot{z}(t) = Az(t) + f\big(t, u(t), v(t)\big), \quad z(0) = z_0 \tag{40}$$

when f is sufficiently smooth. We formulate this as an auxiliary result.

Proposition 1 *Consider the problem (MIOCP) with equation (2) replaced by equation (40), let $\bar{M} = \sup_{t\in[0,t_f]} \|T(t)\|_{\mathcal{L}(X)}$ and suppose that the functions $g^i : [0, t_f] \to X$ defined by $g^i(t) = f(t, \omega^*(t), v^i)$, $i = 1, \ldots, N$, satisfy the following conditions.*

(i) *$g^i(t) \in D(A)$ for a.e. $t \in [0, t_f]$ and there exists constants $\bar{L}^i$ such that*

$$\mathrm{ess}\sup_{t\in[0,t_f]} \big\| Ag^i(t) \big\|_X \leq \bar{L}^i.$$

(ii) *g^i is differentiable for a.e. $t \in [0, t_f]$ and there exist constants $\bar{C}^i$ such that*

$$\mathrm{ess}\sup_{t\in[0,t_f]} \left\| \frac{d}{dt} g^i(t) \right\|_X \leq \bar{C}^i.$$

Then hypothesis (H$_2$) of Theorem 1 holds with $C_i := \bar{M}(\bar{C}^i + \bar{L}^i)$.

Proof From condition (i) we get from the chain rule that

$$\frac{d}{ds} T(t - s)g(t) = T(t - s)\frac{d}{ds} g(t) - T(t - s)Ag(t)$$

for all $i = 1, \ldots, N$ and thus, by taking the norm, applying the triangular inequality and using the definition of the constants in (i) and (ii) we obtain

$$\left\| \frac{d}{ds} T(t - s) f\big(s, \omega^*(s), v^i\big) \right\|_X \leq \bar{M}(\bar{C}^i + \bar{L}^i). \qquad \square$$

The conditions (i) and (ii) are a natural extension of the differentiability assumptions imposed in [23, Corollary 6] for the case when $A = 0$ and $X = \mathbb{R}^n$. In Sect. 4, we will use such arguments in order to verify hypothesis (H$_2$) in Example 1.

We note that the relaxation method works under much weaker assumptions with slightly weaker conclusions. Suppose that we replace the main hypothesis (H$_0$) by the following much weaker hypothesis.

(H$_0'$) Problem (11a)–(11f) has a feasible solution in $U_{[0,t_f]} \times \tilde{V}_{[0,t_f]} \times X_{[0,t_f]}$.

Then, we still get the following result, being useful in particular in many practical applications when the solutions $[\omega^k, \alpha^k, y^k]$ found in step 3 only satisfy, for example, necessary optimality conditions (up to an accuracy of ε^k). The following conclusion can then still be very useful in order to provide bounds for the mixed-integer problem (MIOCP) and we will take advantage of this when discussing the examples in Sect. 4. Nevertheless, an approximation of a globally optimal solution of problem (MIOCP) can of course only be obtained when the relaxed problem in step 3 is solved to ε^k-global optimality.

Proposition 2 *Under the hypothesis* (H$_0'$), *consider Algorithm* 1 *where we replace step* 3 *by*

3′ *Select some* $[\omega^k, \alpha^k, y^k]$ *such that* ω^k *and* α^k *is feasible for problem* (11a)–(11f) *and* y^k *is an* ε^k-*accurate solution of* (11b), (11c).

Assume hypothesis (H$_1$) *and that the hypothesis* (H$_2$) *and* (H$_3$) *hold with* $\omega^* = \omega^k$ *and* $y^* = y(\cdot; \omega^k, \alpha^k)$ *with constants* C^k *and* M^k *for all* $k = 0, 1, 2, \dots$ *and* $y(\cdot; \omega^k, \alpha^k)$ *being the mild solution of* (11b), (11c). *Further assume that, as in Theorem* 1, $\varepsilon^k \to 0$ *and that the sequence* $\{\mathcal{G}^k\}_k$ *is such that* $\Delta t^k \to 0$. *Define the constants* $C_1 = 2$, $C_2^k = ((M^k + t_f C^k) e^{t_f \bar{M} L})$, $C_3 = (\eta + t_f \xi) C_1$ *and* $C_4^k = (\eta + t_f \xi) C_2^k$, *where* $\bar{M} = \sup_{t \in [0,t_f]} \|T(t)\|_{\mathcal{L}(X)}$ *and the constants* η, ξ *and* L *are given by hypothesis* (H$_1$). *Then,* $[u^k, v^k, z^k]$ *defined by Algorithm* 1 *is in* Ξ_{ε^k} *for all* $k = 0, 1, 2, \dots$ *and satisfies the estimates*

$$\left\| y^k(t) - z^k(t) \right\|_X \leq C_1 \varepsilon^k + C_2^k (N-1) \Delta t^k, \quad t \in [0, t_f], \tag{41}$$

and

$$\left| J\left(\omega^k, \alpha^k, y^k \right) - J\left(u^k, v^k, z^k \right) \right| \leq C_3 \varepsilon^k + C_4^k (N-1) \Delta t^k. \tag{42}$$

In particular, if M^k *and* C^k *can be chosen independently of* k, *then Algorithm* 1 *terminates in a finite number of steps* κ *with an* ε-*feasible mixed-integer solution* $[u^*, v^*, z^*] \in \Xi_\varepsilon$ *of Problem* (MIOCP) *satisfying the estimate*

$$\left| J\left(\omega^\kappa, \alpha^\kappa, y^\kappa \right) - J\left(u^*, v^*, z^* \right) \right| \leq \varepsilon, \tag{43}$$

where $\varepsilon > 0$ *was chosen arbitrarily in step* 1.

Proof Let the assumptions of Proposition 2 hold true. First observe that the sequence $[u^k, v^k, z^k]$ obtained by Algorithm 1 with step 3 replaced by 3′ is ε^k-feasible for the problem (MIOCP) by the same arguments as in the proof of Theorem 1.

To show (41) and (42), let $y(\cdot; \omega^k, \alpha^k)$ denote the mild solution of (11b), (11c) and $z(\cdot; u^k, \beta^k)$ be the mild solution of (A$_3$) with the respective controls. Then,

$$\left\| y^k(t) - z^k(t) \right\|_X \le \left\| y(t; \omega^k, \alpha^k) - z(t; u^k, \beta^k) \right\|_X + 2\varepsilon^k, \quad t \in [0, t_f] \qquad (44)$$

because y^k and z^k are both ε^k-accurate solutions of (11b), (11c) and (A$_3$), respectively. From Lemma 2 with $\alpha = \alpha^k$, $\beta = \beta^k$, and $\Delta t = \Delta t^k$ with Δt^k from (15), we get that

$$\max_{i=1,\dots,N} \left| \int_0^t \alpha_i^k(\tau) - \beta_i^k(\tau) \, d\tau \right| \le (N-1)\Delta t^k, \quad t \in [0, t_f]. \qquad (45)$$

Moreover, under hypothesis (H$_1$) and the assumption that the hypothesis (H$_2$) and (H$_3$) hold with $\omega^* = \omega^k$ and $y^* = y^k$ with constants C^k and M^k for all $k = 0, 1, 2, \dots$, we may apply Lemma 1 with $y^* = y(\cdot; \omega^k, \alpha^k)$, $\alpha^* = \alpha^k$, $\beta = \beta^k$, $\omega^* = \omega^k = u^k$ and $\varepsilon = (N-1)\Delta t^k$ and obtain that

$$\left\| y(t; \omega^k, \alpha^k) - z(t; u^k, \beta^k) \right\|_X \le \left((M^k + C^k t) e^{\bar{M} L t} \right)(N-1)\Delta t^k, \quad t \in [0, t_f].$$

This proves (41) with $C_1 = 2$ and $C_2^k = ((M^k + t_f C^k)e^{t_f \bar{M} L})$. Using again the continuity assumptions of ϕ and ψ in (H$_1$), we obtain similarly as in the proof of Theorem 1 that

$$\left\| J(\omega^k, \alpha^k, y^k) - J(u^k, v^k, z^k) \right\|_X \le C_3 \varepsilon^k + C_4^k(t)(N-1)\Delta t^k \qquad (46)$$

$C_3 = C_1(\eta + t_f \xi)$ and $C_4^k = (\eta + t_f \xi)C_2^k$. This proves (42).

Finally, when the constants C^k and M^k can be chosen independently of k, $\varepsilon^k \to 0$ and $\Delta t^k \to 0$, then C_4^k is independent of k and we get by similar arguments as in the proof of Theorem 1 that the iteration terminates after a finite number of steps such that (43) holds. □

In order to solve the optimal control problem (11a)–(11f) numerically, the problem may have to be (adaptively) discretized. In particular, direct or indirect numerical methods may be used. For an introduction to the basic concepts see, e.g., [12]. Depending on the method of choice for the time discretization of the control functions $\omega \in U_{[0,t_f]}$ and $\alpha \in \tilde{V}_{[0,t_f]}$ it may in many cases be advantageous to discretize ω^k and α^k simultaneously using the grid $\mathcal{G}^k$. This is for example implemented in the software package MS MINTOC designed for solving mixed-integer optimal control problems with ordinary differential equations [22, 24].

We conclude this section with an interesting remark saying that, in the fashion of Theorem 1 and Proposition 2, the relaxation method can also deal with state constraints.

Remark 1 Suppose that we wish to include a constraint of the form

$$G(z(t), t) \ge 0, \quad t \in [0, t_f] \qquad (47)$$

in the mixed-integer optimal control problem (MIOCP). Including this constraint also in (11a)–(11f) with $z(t)$ replaced by $y(t)$ and assuming that there exists a function $\zeta \in L^\infty(0, t_f)$ such that

$$\big|G(y_1, t) - G(y_2, t)\big| \le \zeta(t)\|y_1 - y_2\|_X, \quad y_1, y_2 \in X \tag{48}$$

then (16) yields that

$$\big|G\big(y^*(t), t\big) - G\big(z^k(t), t\big)\big| \le \zeta(t)C_1\varepsilon^k + \zeta(t)C_2(N-1)\Delta t^k \tag{49}$$

with C_1 and C_2 as in Theorem 1. This shows also a bilinear dependency of the integer-control approximation error for the state constraint violation on Δt^k and ε^k. The conclusion of Proposition 2 can be adapted accordingly.

3 Combinatorial constraints

Suppose we wish to include combinatorial constraints of the form

$$\#_{v^i \frown v^j}(v) \le K^{i,j}, \quad i \in I, \quad j \in J \tag{50}$$

into the mixed-integer optimal control problem (MIOCP) given by (1)–(3), where $\#_{v^i \frown v^j}(v)$ denotes the number of switches of the control function $v : [0, t_f] \to V_{\mathrm{ad}}$ from value v^i to value v^j, $K^{i,j}$ are given, non-negative constants and $I, J \subset \{1, \ldots, N\}$.

Note that the relaxation method considered in Sect. 2 typically satisfies

$$\#_{v^i \frown v^j}(v) \to +\infty$$

for some $i, j \in \{1, \ldots, N\}$ as we let $\varepsilon \to 0$, so eventually violating (50) for small ε. Therefore, along the lines of [25], we propose in Algorithm 2 a modification of Algorithm 1. The min-max problem (A_5) can be written as a standard mixed-integer linear problem (MILP) using slack variables and can be computed efficiently [25]. We then have the following result.

Theorem 2 *Suppose that the hypotheses of Theorem 1 hold true and let C_1, C_2, C_3 and C_4 be as in Theorem 1. Then Algorithm 2 terminates for every $\varepsilon > 0$ and $k_{\max} \ge 0$ after a finite number of steps $\kappa \le k_{\max}$ with an ε^κ-feasible solution $[u^*, v^*, z^*] \in \Xi_{\varepsilon^\kappa}$ of problem (MIOCP) satisfying the combinatorial constraints (50) and the estimates*

$$\big\|y^*(t) - z^*(t)\big\|_X \le C_1\varepsilon^\kappa + C_2\big(J_{\mathrm{sub}}\big(p^{\kappa,*}\big) + \delta\big), \quad t \in [0, t_f], \tag{51}$$

and

$$\big|J\big(\omega^*, \alpha^*, y^*\big) - J\big(u^*, v^*, z^*\big)\big| \le C_3\varepsilon^\kappa + C_4\big(J_{\mathrm{sub}}\big(p^{\kappa,*}\big) + \delta\big) \tag{52}$$

for $J_{\mathrm{sub}}(p^{\kappa,*})$ given by Algorithm 2 and some $0 \le \delta \le \max_{l=1,\ldots,n^\kappa} \Delta t_l^\kappa$.

Algorithm 2 Relaxation method with switching constraints

Consider Algorithm 1 where we replace steps 4, 5 and 7 by

$4'$ If $\alpha^k \in PC(0, t_f; \{0, 1\}^N)$, (50) holds with $v = \sum_{i=1}^{N} \alpha_i^k v^i$ and $\varepsilon^k \leq \varepsilon$, then set $\beta^k(t) = \alpha^k(t)$ and $z^k(t) = y^k(t)$ for $t \in [0, t_f]$ and STOP.

$5'$ Using $\mathcal{G}^k$, define a piecewise constant function $\beta^k = (\beta_1^k, \ldots, \beta_N^k) : [0, t_f] \to \{0, 1\}^N$ by

$$\beta_i^k(t) = p_{i,j}^{k,*}, \quad t \in [t_j^k, t_{j+1}^k) \tag{A4}$$

where $p_{i,j}^{k,*}$ is given by the solution of the min-max problem

$$\begin{cases} \displaystyle\min_{p^k} J_{\mathrm{sub}}(p^k) = \max_{i=1,\ldots,N} \max_{r=1,\ldots,n^k} \left| \sum_{l=1}^{r} (q_{i,l}^k - p_{i,l}^k) \Delta t_l^k \right| \\[2ex] \text{subject to} \\[1ex] \displaystyle\sum_{r=1}^{n^k} \left| p_{i,r}^k - p_{j,r+1}^k \right| \leq K^{i,j}, \quad i \in I, \quad j \in J \\[2ex] \displaystyle\sum_{i=1}^{N} p_{i,r}^k = 1, \quad r = 1, \ldots, n^k \\[2ex] p_{i,r}^k \in \{0, 1\}, \quad i = 1, \ldots, N, \quad r = 1, \ldots, n^k \end{cases} \tag{A5}$$

with $\Delta t_l^k = t_{l+1}^k - t_l^k$, $l = 1, \ldots, n^k$, and

$$q_{i,l}^k = \frac{1}{\Delta t_l^k} \int_{t_l}^{t_{l+1}} \alpha_i^k(t) \, dt, \quad i = 1, \ldots, N, \quad l = 1, \ldots, n^k. \tag{A6}$$

$7'$ If $|J_{\mathrm{sub}}(p^{k,*}) - J_{\mathrm{sub}}(p^{k-1,*})| < \varepsilon$ or $k \geq k_{\max}$ then STOP.

Proof Algorithm 2 terminates by the criterion in step $7'$ after κ steps, $\kappa \leq k_{\max}$, with a solution $[u^*, v^*, z^*]$. We first show that this solution is ε^κ-feasible for the problem (MIOCP). We have $u^* = \omega^\kappa \in U_{\mathrm{ad}}$ and by definition of ω^κ in step 3, $v^\kappa(t) = \sum_{i=1}^{N} \beta_i^\kappa(t) v^i$, $t \in [0, t_f]$, by definition in step 6, so $v^\kappa(t) \in V_{\mathrm{ad}}$, $t \in [0, t_f]$, because $\beta_i^k(t) \in \{0, 1\}$ for all i, k and $t \in [0, t_f]$ as seen from (A4) and the constraints in (A5).By construction in step 6, z^κ is an ε^κ-accurate solution of (A3). Thus, recalling (9), $[u^*, v^*, z^*] \in \Xi_{\varepsilon^\kappa}$. The constraints in (A5) also ensure that the combinatorial constraints (50) are satisfied.

Next we show (51). For all $k = 0, 1, \ldots$ the cost function in (50) is defined as

$$J_{\mathrm{sub}}(p^{k,*}) = \max_{i=1,\ldots,N} \max_{r=1,\ldots,n^k} \left| \sum_{l=1}^{r} (q_{i,l}^k - p_{i,l}^k) \Delta t_l^k \right|. \tag{53}$$

By definition of $q_{i,l}^k$ in (A$_6$) and $\beta_i^k(t)$ in (A$_4$) and rearranging terms, we get

$$J_{\text{sub}}\big(p^{k,*}\big) = \max_{i=1,\ldots,N} \max_{r=1,\ldots,n^k} \left| \int_0^{t_{r+1}} \alpha_i^k(t) - \beta_i^k(t)\,dt \right|. \tag{54}$$

Using that $\alpha_i^k(t) \in [0,1]$ and $\beta_i^k(t) \in \{0,1\}$ for all $t \in [0,t_f]$, this yields

$$J_{\text{sub}}\big(p^{k,*}\big) = \max_{i=1,\ldots,N} \sup_{t \in [0,t_f]} \left| \int_0^t \alpha_i^k(\tau) - \beta_i^k(\tau)\,d\tau \right| - \delta \tag{55}$$

for some $0 \le \delta \le \max_{l=1,\ldots,n^k} \Delta t_l^k$. Fixing some $t \in [0,t_f]$, we have as in the proof of Theorem 1

$$\big\| z^k(t) - y^*(t) \big\|_X \le \delta_1(t) + \delta_2(t) + \delta_3(t) + \delta_4(t), \tag{56}$$

with $\delta_i(t)$, $i = 1,\ldots,4$, as in (15). Moreover, as in the proof of Theorem 1, we see that $\delta_1(t) \le \varepsilon^k$, $\delta_2(t) \le C_J(1 + \eta + \xi)e^{\bar{M}Lt}\varepsilon^k$ and $\delta_4(t) \le MC_J(1 + \eta + \xi)e^{\bar{M}Lt}\varepsilon^k$. Using (H$_1$)–(H$_3$), we can apply Lemma 1 with $y^* = y(\cdot;\omega^*,\alpha^k)$, $\alpha^* = \alpha^k$, $\beta = \beta^k$ and $\varepsilon = J_{\text{sub}}(p^{k,*}) + \delta$ and get that

$$\delta_3(t) = \big\| y\big(t;\omega^*,\beta^k\big) - y\big(t;\omega^*,\alpha^k\big) \big\|_X \le (M + Ct)e^{\bar{M}Lt}\big(J_{\text{sub}}\big(p^{k,*}\big) + \delta\big). \tag{57}$$

Summing up the estimates for $\delta_1(t), \ldots, \delta_4(t)$ and rearranging terms, this proves (51) with the definition of C_1 and C_2 as in Theorem 1. The estimate (52) then follows from (51) and the definition of the constants C_3 and C_4 as in Theorem 1 using the definition of the cost function J in (1) and the Lipschitz constants η and ξ from (H$_1$). This completes the proof of Theorem 2. $\qquad\square$

Remark 2 As already remarked in the case without combinatorial constraints, the method can also deal with state constraints such as (47). Assuming again existence of a function $\zeta \in L^\infty(0,t_f)$ such that (48) holds true, (51) yields a bounded deviation of the feasible reference trajectory

$$\big| G\big(y^*(t),t\big) - G\big(z^k(t),t\big) \big| \le \zeta(t)C_1\varepsilon^k + \zeta(t)C_2\big(J_{\text{sub}}\big(p^{k,*}\big) + \delta\big), \tag{58}$$

and hence a bound on the worst case constraint violation. Also, the conclusion of Proposition 2 can be adapted analogously.

4 Examples

In this section we discuss the hypothesis (H$_1$)–(H$_3$) of Theorem 1 exemplary for a linear and a semilinear control problem where A is the generator of an analytic semigroup in view of Proposition 2 and present numerical results for a test problem in each case using the conclusions.

4.1 A linear parabolic equation with lumped controls

Let Ω be a domain in $\mathbb{R}^n$ and $f_i : \Omega \to \mathbb{R}$, $i = 1, \ldots, N$, be fixed control profiles. Consider the internally controlled heat equation

$$
\begin{cases}
\dfrac{\partial z}{\partial t}(x, t) - \rho \sum_{j=1}^{n} \dfrac{\partial^2 z}{\partial x_j^2}(x, t) = f_{\sigma(t)}(x) u(t), & \text{in } Q \\[2mm]
z(x, t) = 0, & \text{on } \Sigma \\[2mm]
z(x, 0) = z_0(x), & \text{in } \Omega
\end{cases}
\tag{59}
$$

where $Q = \Omega \times (0, t_f)$, $\Sigma = \partial \Omega \times (0, t_f)$ and ρ is a positive constant.

Suppose that for some $\lambda_1 \geq 0$ and $\lambda_2 > 0$ the control task is to minimize the cost function

$$
J = \int_{\Omega} \left| z(t_f, x) \right|^2 dx + \lambda_1 \int_0^{t_f} \int_{\Omega} \left| z(t, x) \right|^2 dx\, dt + \lambda_2 \int_0^{t_f} \left| u(t) \right|^2 dt
\tag{60}
$$

where z is the weak solution of (59) by selecting $u : [0, t_f] \to \mathbb{R}$ and a switching signal $\sigma(\cdot) : [0, t_f] \to \{1, \ldots, N\}$ determining the control profile f_i applied at time $t \in [0, t_f]$.

In order to write the above problem in abstract form (2), we let $X = L^2(\Omega)$, set $V = U = U_{\text{ad}} = \mathbb{R}$, $V_{\text{ad}} = \{1, \ldots, N\}$ and define $f : [0, t_f] \times U \times V \to X$ by $f(t, u, v)(x) := f_v(x) u$, $\phi(z) = \|z\|_X^2$, $\psi(z, u) = \lambda_1 \|z\|_X^2 + \lambda_2 |u|^2$ and define $(A, D(A))$ as

$$
D(A) = H^2(\Omega) \cap H_0^1(\Omega)
$$

$$
(Az)(x) = \sum_{j=1}^{n} \frac{\partial^2 z}{\partial x_j^2}(x), \quad z \in D(A).
\tag{61}
$$

It is well-known that $(A, D(A))$ is the generator of a strongly continuous (analytic) semigroup of contractions $\{T(t)\}_{t \geq 0}$ on X, see, e.g., [20]. We choose $X_{[0, t_f]} = C([0, t_f]; X)$, $U_{[0, t_f]} = PC(0, t_f; \mathbb{R})$ and $V_{[0, t_f]} = L^\infty(0, t_f; \mathbb{R})$.

Let $[\omega^k, \alpha^k, y^k]$ be a sequence of feasible solutions of the corresponding relaxed problem (11a)–(11f) so that $[\omega^k, \alpha^k, y^k] \in \mathcal{S}$ for $k = 0, 1, 2, \ldots$, where $\mathcal{S}$ is a bounded subset of $X_{[0, t_f]} \times U_{[0, t_f]} \times \tilde{V}_{[0, t_f]}$ and assume that the fixed control profiles f_i satisfy

$$
f_i \in D(A) \quad \text{for all } i = 1, \ldots, N.
\tag{62}
$$

We now want to check if the assumptions of Proposition 2 are satisfied. For this, we can restrict our analysis without loss of generality to $\mathcal{S}$ and, using that the functions ϕ, ψ and f are locally Lipschitz continuous and $\mathcal{S}$ is bounded, we can see that hypothesis (H1) holds. Moreover, due to (62) we have that $f(t, \omega^k(t), v^i) = f_v(x) \omega^k(t) \in D(A)$ and

$$
\left\| A f\left(t, \omega^k(t), v^i\right) \right\|_X \leq \|A f_i\|_X \left| \omega^k(t) \right| \leq \bar{C}^i
\tag{63}
$$

for some constant $\bar{C}^i$ and due to the choice of $U_{[0,t_f]}$ we have that $f(s, \omega^k(s), v^i)$ is differentiable for a.e. $s \in [0, t_f]$ and

$$\left\| \frac{d}{ds} f\left(s, \omega^k(s), v^i\right) \right\|_X = \left\| \frac{d}{ds} f_i \omega^k(s) \right\|_X \leq \|f_i\|_X \left| \frac{d}{ds} \omega^k(s) \right| \leq \bar{L}^i \qquad (64)$$

for some constant $\bar{L}^i$. Noting that (59) and thus also the abstract system is linear, we may apply Proposition 1 to see that hypothesis (H$_2$) holds for every $k = 0, 1, 2, \ldots$. Also, the bound in (H$_3$) holds for all k, observing that

$$\sup_{t \in (0,t_f)} \left\| f\left(t, \omega^k(t), v^i\right) \right\|_X \leq \|f_i\|_X \left\| \omega^k(t) \right\|_\infty \leq M^i, \qquad (65)$$

for some constant M^i. Hence, for any such choices $[\omega^k, \alpha^k, y^k]$, any sequence $\varepsilon^k \to 0$, $\Delta t^k \to 0$ and any $\varepsilon > 0$, the relaxation method terminates after finitely many steps κ with an ε-feasible solution $[u^*, v^*, z^*]$ of problem (59) satisfying the estimate

$$\left| J\left(\omega^\kappa, \alpha^\kappa, y^\kappa\right) - J\left(u^*, v^*, z^*\right) \right| \leq \varepsilon, \qquad (66)$$

by Proposition 2. The desired switching structure $\sigma : [0, t_f] \to \{1, \ldots, N\}$ is finally given by $\sigma(t) = v^*(t)$.

Example 1 To demonstrate the applicability of the approach, we implemented the relaxation method for a test problem of the form (59)–(60) with a two-dimensional rectangular domain Ω and the following parameters.

Let $\Omega = [0, L_\xi] \times [0, L_\zeta]$, $\rho = 0.01$, $L_\xi = 1$, $L_\zeta = 2$ and $t_f = 15$ and suppose that there are given 9 actuator locations x_i with the positions given by $(\xi_j, \zeta_k) \in \Omega$, where

$$\xi_j = \frac{j + 0.005 L_\xi}{4}, \quad \zeta_k = \frac{k + 0.005 L_\zeta}{4}, \quad j, k = 1, 2, 3. \qquad (67)$$

Further suppose that there is a point actuator for each of these locations x_i which we model here by setting $f_i = B_i$ with

$$B_i(x) = \frac{1}{\sqrt{2\pi\epsilon}} e^{\frac{-(x_i - x)^2}{2\epsilon}} \qquad (68)$$

for some small, but fixed $\epsilon > 0$. Note that $\int_\Omega B_i(x)\,dx = 1$ and that $B_i(x)$ converges to the Dirac delta function $\delta(x - x_i)$ as $\epsilon \to 0$.

As initial data we take

$$z_0(\xi, \zeta) = 10 \sin(\pi\xi) 10 \sin(\pi\zeta) \qquad (69)$$

and as parameters in the cost function we take $\lambda_1 = 2$ and $\lambda_2 = \frac{1}{500}$.

We have chosen these numerical values to match as closely as possible the two-dimensional example in [13] motivated by thermal manufacturing. The only difference is that the pointwise actuators $\delta(x - x_i)$ were approximated in [13] by indicator functions of an epsilon environment while we choose here a smoother approximation

Table 1 Performance of the relaxation method for Example 1

k	$\Delta t^k_{\max}$	$J^k_{\text{rel}} = J(\omega^k, \tilde{\alpha}^k, y^k)$	$J^k = J(u^k, v^k, z^k)$	Error $(J^2_{\text{rel}})^{-1} \lvert J^2_{\text{rel}} - J^k \rvert$
0	1.8750	5.634024E+04	1.283813E+05	2.9809
1	0.9375	4.190360E+04	7.080185E+04	1.1955
2	0.4688	3.224914E+04	6.175488E+04	0.9149

in view of (62). Regarding a direct treatment of $\delta(x - x_i)$ as an unbounded control operator instead of using the bounded approximation (68), see the comments in Sect. 5.

The solution of the relaxed optimal control problem (11a)–(11f) has been computed numerically. We discretized the state equation (59) in space using a standard Galerkin approach with triangular elements and linear Ansatz-functions. We eliminated one control by setting $\tilde{\alpha}_i(t) = \alpha_i(t)$, $i = 1, \ldots, N - 1$, where we then get $\alpha_N(t) = 1 - \sum_{i=1}^{N-1} \tilde{\alpha}_i(t)$ using the constraint $\sum_{i=1}^{N} \alpha_i(t) = 1$, $t \in [0, t_f]$. This constraint is then always fulfilled and the condition $\alpha_N(t) \in [0, 1]$, $t \geq t_0$, is equivalent to imposing that $\tilde{\alpha}_i \in [0, 1]$ and $\sum_{i=1}^{N-1} \tilde{\alpha}_i - 1 \leq 0$. The resulting semi-discretized control problem was solved with Bock's direct multiple shooting method [3, 17] implemented in the software-package MUSCOD-II. The control functions ω and $\tilde{\alpha}$ were chosen as piecewise constant and initialized with $\omega(t) = 0$ and $\tilde{\alpha}_i(t) = \frac{1}{9}$, $t \in [0, t_f]$, $i = 1, \ldots, 8$.

The computations were made for an unstructured grid with 162 triangular elements and 8, 16 and 32 equidistant shooting intervals. Time integration was carried out by a BDF-method and sensitivities were computed using internal numerical differentiation. Error estimates for these methods provide an accuracy of some $\varepsilon_1 \geq 0$ for the so obtained approximations of the mild solutions.

We implemented Algorithm 1 where we compute in step 3 solutions satisfying first order necessary conditions with an accuracy of ε_2. The above discussion of the abstract example thus applies. In particular, using that $\lambda_2 > 0$, we obtain the existence of a bounded set $\mathcal{S}$ containing the iterates $[\omega^k, \alpha^k, y^k]$. We adaptively solved the relaxed problem on a common control discretization grid $\mathcal{G}^k$ for u and $\tilde{\alpha}$. For the computations, we have chosen $\varepsilon_1 = 1.0\text{E}{-}04$ and used bisection for refinements of the control grids $\mathcal{G}^k$ in step 7. Thus, ε^k is given implicitly as a function of ε_1, ε_2 and Δt^k. This construction ensures that $\varepsilon^k \to 0$ as $k \to \infty$.

The performance of the relaxation method is summarized in Table 1. We see that the relative error of the mixed-integer solution compared with the best found relaxed solution decreases with the grid refinements in accordance with Proposition 2. The best found controls and the evolution of the state norm of the corresponding solution are displayed in Fig. 1. We see the rounding error in form of an overshooting behavior when comparing the evolution of the $L^2(\Omega)$-norm of the relaxed and the mixed-integer solution. This effect decreases with the size of the time discretization step size. The cost corresponding to the best found solution is 6175. Unfortunately, [13] does not report the cost of the best found solution, but a cumulative $L^2(0, 15; L^2(\Omega))$-norm of 90.27. The cumulative $L^2(0, 15; L^2(\Omega))$-norm of our best found solution is 78.58.
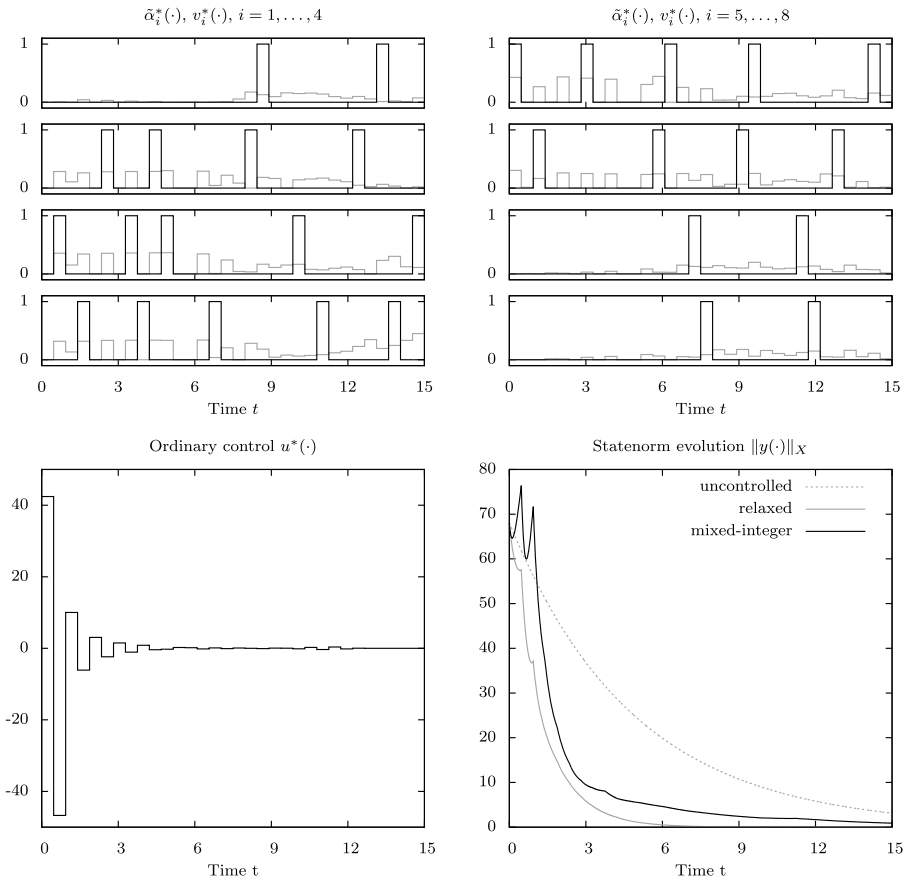
Fig. 1 Numerical results for Example 1. The *upper figures* show the best found integer controls $v_i^*(\cdot)$ and their relaxation $\tilde{\alpha}_i^*(\cdot)$, from *bottom* to *top*, $i = 1, \ldots, 4$ (*left*) and $i = 5, \ldots, 8$ (*right*). Control $v_9^*(\cdot)$ is defined by $v_9^*(t) = 1 - \sum_{i=1}^{8} v_i^*(t)$, $t \in [0, 15]$. The lower figures show the corresponding ordinary control $u^*(\cdot)$ (*left*) and the evolution of the state norm (*right*)

4.2 A semilinear reaction-diffusion system

Let Ω be a bounded domain in $\mathbb{R}^n$ with a smooth boundary Γ and consider the classical Lotka-Volterra system with diffusion

$$
\begin{cases}
\dfrac{\partial z_1}{\partial t}(x, t) - d_1 \displaystyle\sum_{j=1}^{n} \dfrac{\partial^2 z_1}{\partial x_j^2}(x, t) = z_1(x, t)\big(a_1 - b_1 v(t) - c_1 z_2(x, t)\big) & \text{in } Q \\[4mm]
\dfrac{\partial z_2}{\partial t}(x, t) - d_2 \displaystyle\sum_{j=1}^{n} \dfrac{\partial^2 z_2}{\partial x_j^2}(x, t) = z_2(x, t)\big(a_2 - b_2 v(t) - c_2 z_1(x, t)\big) & \text{in } Q \\[4mm]
\dfrac{\partial z_1}{\partial \nu}(x, t) = \dfrac{\partial z_2}{\partial \nu}(x, t) = 0 & \text{on } \Sigma \\[2mm]
z_1(x, 0) = z_{1,0}(x), \quad z_2(x, 0) = z_{2,0}(x) & \text{in } \Omega
\end{cases}
\tag{70}
$$

with constants $a_i, b_i, c_i, d_i > 0$, $i = 1, 2$, domains $Q = \Omega \times [0, t_f]$, $\Sigma = \Gamma \times [0, t_f]$ and control $0 \leq v(t) \leq 1$. System (70) describes the interaction of two populations z_1 and z_2, both spatially distributed and diffusing in Ω. The initial distribution $z_{1,0}, z_{2,0}$ at $t = 0$ is assumed to be non-negative. The boundary conditions then imply that the populations z_1 and z_2 are confined in Ω for all $t \geq 0$. The function v models a control of the system and we shall investigate to approximate optimal controls $v^*(t)$ taking values in $\{0, 1\}$ as to minimize the distance of the population (z_1, z_2) to its uncontrolled ($v = 0$) steady state distribution $(\bar{z}_1, \bar{z}_2)$ given by the constant functions

$$\bar{z}_1(x) = \frac{a_2}{c_2}, \quad \bar{z}_2(x) = \frac{a_1}{c_1}, \quad x \in \Omega.$$

In order to bring the system into abstract form (2), set $X = L^2(\Omega) \times L^2(\Omega)$, $U = U_{\mathrm{ad}} = \{\}$, $V = \mathbb{R}$, $V_{\mathrm{ad}} = \{0, 1\}$, define the operator $A : D(A) \to X$ by

$$D(A) = \left\{ (z_1, z_2) \in H^2(\Omega) \times H^2(\Omega) : \frac{\partial z_1}{\partial \nu}(x, t) = \frac{\partial z_2}{\partial \nu}(x, t) = 0, \text{ on } \Gamma \right\}$$

$$A(z_1, z_2)(x) = \left(d_1 \sum_{j=1}^{n} \frac{\partial^2 z_1}{\partial x_j^2}(x), d_2 \sum_{j=1}^{n} \frac{\partial^2 z_2}{\partial x_j^2}(x) \right), \quad (z_1, z_2) \in D(A),$$

define the non-linear function $f : X \times U \times V = X \times V \to X$ by

$$f\big((z_1, z_2), v\big)(x) = \big(z_1(x)\big(a_1 - b_1 v - c_1 z_2(x)\big), z_2(x)\big(a_2 - b_2 v - c_2 z_2(x)\big)\big)$$

and define the cost functions ϕ and ψ by

$$\phi\big((z_1, z_2)\big) = 0, \quad \psi\big((z_1, z_2)\big) = \int_{\Omega} \big\|z_1(x) - \bar{z}_1(x)\big\|^2 + \big\|z_2(x) - \bar{z}_2(x)\big\|^2 \, dx. \quad (71)$$

We choose $X_{[0,t_f]} = C([0, t_f]; X)$ and $V_{[0,t_f]} = PC(0, t_f; \mathbb{R})$.

It is well-known, that $(A, D(A))$ is the generator of an analytic semigroup on X and that for any non-negative initial data $z_{1,0}, z_{1,0} \in X$, the system (70) has a non-negative unique mild solution $(z_1, z_2) \in C([0, t_f], X \times X)$ for every $v \in L^\infty(0, t_f; [0, 1])$. Moreover, for initial data $z_{1,0}, z_{1,0} \in D(A)$ and $v \in C_{\mathrm{pw}}^{0,\vartheta}(0, t_f; [0, 1])$, $\vartheta > 0$, this solution is classical and satisfies

$$(z_1, z_2) \in C\big([0, t_f]; D(A) \times D(A)\big) \cap H^1\big([0, t_f]; X \times X\big). \quad (72)$$

Existence (local in time) and uniqueness follows from classical theory for semilinear parabolic equations, see, e.g., [20, Chap. 6]. Global existence results for (70) are obtained by a-priori bounds on the solution using contracting rectangles [4].

Assume that the initial data satisfies $z_{1,0}, z_{1,0} \in D(A)$ and that $[\alpha^k, y^k]$ is a sequence of feasible solutions to the corresponding relaxed problem (11a)–(11f) in a bounded set $\mathcal{S} \subset \tilde{V}_{[0,t_f]} \times X_{[0,t_f]}$. We want to discuss again the assumptions of Proposition 2.

Hypothesis (H$_1$) holds by the same arguments as in the previous example. Moreover, we claim that hypothesis (H$_2$) holds. Let $\bar{M}$ be the growth bound of $\{T(t)\}_{t \geq 0}$

on $[0, t_f]$. By analyticity of $\{T(t)\}_{t\geq 0}$, we have for every feasible $y \in X_{[0, t_f]}$ that s almost everywhere in $(0, t_f)$,

$$\frac{d}{ds} T(t-s) f(y(s), v^i)$$
$$= -AT(t-s) f(y(s), v^i) + T(t-s) f_y(y(s), v^i) y_s(s), \quad (73)$$

where $f_y = \frac{d}{dy} f$ and $y_s = \frac{d}{ds} y$. Using that $y^k(s) \in D(A) \times D(A)$ for all $s \in [0, t_f]$ and $f : D(A) \to D(A)$, we see that

$$\left\| -AT(t-s) f(y^k(s), v^i) \right\|_X \leq \left\| T(t-s) \right\|_{\mathcal{L}(X)} \left\| Af(y^k(s), v^i) \right\|_X \leq C_1^{k,i} \quad (74)$$

for some constants $C_1^{k,i}$. Using that f is a smooth function, we see that

$$\left\| T(t-s) f_y(y^k(s), v^i) y_s^k(s) \right\|_X \leq \bar{M} \left\| f_y(y^k(s), v^i) \right\|_X \left\| y_s^k(s) \right\|_X \leq C_2^{k,i}. \quad (75)$$

Thus (73) yields the estimate

$$\left\| \frac{d}{ds} T(t-s) f(y(s), v^i) \right\|_X \leq C_1^{k,i} + C_2^{k,i} \quad (76)$$

for $s \in [0, t_f]$ a.e. By well-posedness of the problem (70) for every $v \in V_{[0, t_f]}$ and using the boundedness of $\mathcal{S}$, we get an estimate

$$\sup_{t \in [0, t_f]} \left\| f(t, y^k(t), v^i) \right\|_X \leq M^i, \quad (77)$$

for some constants M^i verifying hypothesis (H$_3$). Further, using (H$_1$) and the boundedness of $\mathcal{S}$, $\| f_y(y^k(s), v^i) \|_X$ and $\| Af(y^k(s), v^i) \|_X$ can be bounded independently of k. Using the state equation (11b), we get

$$\left\| y_s^k(s) \right\|_X \leq \left\| Ay^k(s) \right\|_X + \left\| f(y^k(s), v^i) \right\|_X \leq \left\| y^k(s) \right\|_{D(A)} + M^i \quad (78)$$

for a.e. $s \in [0, t_f]$. So, boundedness of S and the regularity of y in (72) implies that $\| y_s^k(s) \|_X$ can be bounded independently of k for a.e. $s \in [0, t_f]$. Hence, the constants $C_1^{k,i}$ and $C_2^{k,i}$ in (76) can be chosen independently of k and we can conclude from Proposition 2 that the relaxation methods terminates after κ steps with an integer solution $[v^*, z^*]$ satisfying the estimate

$$\left| J(\alpha^\kappa, y^\kappa) - J(v^*, z^*) \right| \leq \varepsilon, \quad (79)$$

for any sequences $\varepsilon^k \to 0$, $\Delta t^k \to 0$ and every $\varepsilon > 0$.

Example 2 We applied the relaxation method to a semilinear test problem of the form (70) again for a two dimensional domain Ω with the following parameters. Let Ω being a circle with radius 1 centered at $(1, 1)$ and choose $a_1 = a_2 = c_1 = c_2 = 1$,

Table 2 Performance of the relaxation method for Example 2

k	$\Delta t_{\max}$	Rel. Cost $J^k_{\text{rel}} = J(\alpha^*, y^*)$	$J^k = J(v^*, z^*)$	Error $(J^2_{\text{rel}})^{-1}|J^2_{\text{rel}} - J^k|$
0	2.0000	7.066392E+01	8.287875E+01	0.4065
1	1.0000	5.978818E+01	6.958250E+01	0.1809
2	0.5000	5.892414E+01	5.875641E+01	0.0028
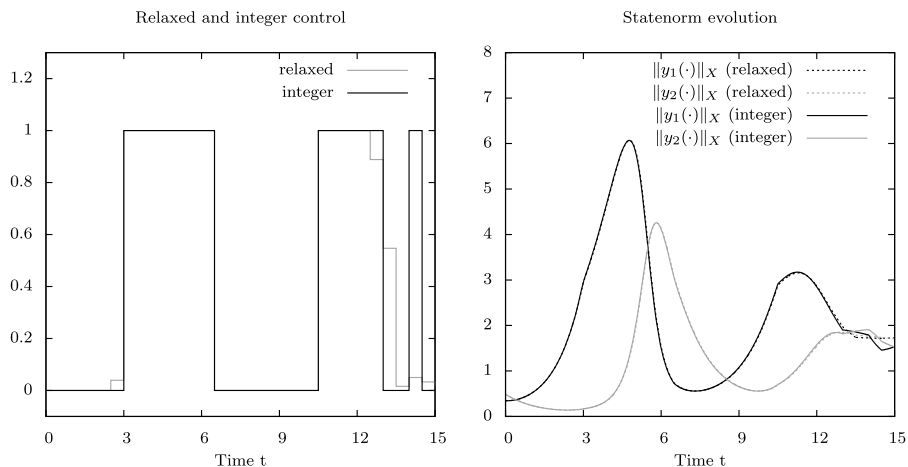
Fig. 2 Numerical results for Example 2. *The left figure* shows the best found relaxed and integer control $\alpha^*(\cdot)$, $v^*(\cdot)$ and *the right figure* shows the corresponding evolutions of the populations $y_1(\cdot)$, $y_2(\cdot)$

$b_1 = \frac{7}{10}$, $b_2 = \frac{1}{2}$, $d_1 = 0.05$, $d_2 = 0.01$, initial data $z_{1,0}$, $z_{2,0} \in D(A)$ approximated by $\tilde{z}_{1,0}(x) = \frac{1}{2}d_{\frac{1}{2}}(x-1)$, $\tilde{z}_{2,0}(x) = \frac{7}{10}d_{\frac{1}{2}}(x-1)$, where $d_\epsilon(x)$ given by

$$d_\epsilon(x) = \frac{1}{\sqrt{2\pi\epsilon}}e^{\frac{-x^2}{2\epsilon}} \tag{80}$$

models a population concentrated at the origin. For $v(t) = 0$, $t \geq 0$, the solution $z_1(t, x)$, $z_2(t, x)$ converges asymptotically to a spatially constant and temporarily non-constant, periodic solution.

The computations for the optimal control are made by the same numerical method as in the previous example, but using a grid with 258 finite elements. For the performance of the relaxation method see Table 2. The best found controls and the evolution of the state norm of the corresponding solutions are displayed in Fig. 2. Again we see a decrease of the integer-approximation error in accordance with Proposition 2. The best found integer control yields a cost of 58.76.

5 Conclusions and open problems

We considered mixed-integer optimal control problems for abstract semilinear evolution equations and obtained conditions guaranteeing that the value function and the

state of a relaxed optimal control problem can be approximated with arbitrary precision using a control that satisfies integer restrictions. In particular, our approach is constructive and gives rise to a numerical method for mixed-integer optimal control problems with certain partial differential equations. Moreover, we showed how these conditions imply a-priori estimates on the quality of the solution when combinatorial constraints are enforced.

We note that we did not discuss convergence of the constructed sequence of integer controls approximating the optimal value of the relaxed problem. It is not even clear in which topology such a convergence would be meaningful. Several issues related to this questions, in particular in a PDE-context, is discussed in [10, 11] and [8].

Compared to the previously available results on mixed-integer optimal control problems with ordinary differential equations in [23, 24], the setting treated in this paper involves a differential operator A, taken to be a generator of a strongly continuous semigroup. This requires careful regularity considerations. When A is a Laplace operator, we showed on a linear and a semilinear example how such regularity assumptions can be met and provided numerical examples demonstrating the practicability of the approach.

It is clear that the methodology considered in this paper generalizes to the case when the generator A of a strongly continuous semigroup is replaced by a family $\{A(t)\}_{t \in [0, t_f]}$ of unbounded linear operators generating an evolution operator in the sense of [16]. On the other hand it is not so clear how to extend the results in case of unbounded control action, for example, Neumann or Dirichlet boundary control for the heat equation. Recalling the density of solutions to (5) in the set of solution to (6) which motivated our approach, we note that the case of unbounded control is not covered by the available results on operator differential inclusions. While in principle semigroup techniques can deal with unbounded control operators, see for example the exposition in [1, Chap. 3], this extension is non-trivial and requires additional work.

References

1. Bensoussan, A., Da Prato, G., Delfour, M.C., Mitter, S.K.: Representation and Control of Infinite-Dimensional Systems. Volume 1. Systems & Control: Foundations & Applications. Birkhäuser Boston, Boston (1992)
2. de Blasi, F.S., Pianigiani, G.: Evolution inclusions in non-separable Banach spaces. Comment. Math. Univ. Carol. **40**(2), 227–250 (1999)
3. Bock, H., Plitt, K.: A multiple shooting algorithm for direct solution of optimal control problems. In: Proceedings of the 9th IFAC World Congress, pp. 242–247. Pergamon, Budapest (1984)
4. Brown, P.N.: Decay to uniform states in ecological interactions. SIAM J. Appl. Math. **38**(1), 22–37 (1980)
5. Engell, S., Toumi, A.: Optimisation and control of chromatography. Comput. Chem. Eng. **29**, 1243–1252 (2005)

6. Frankowska, H.: A priori estimates for operational differential inclusions. J. Differ. Equ. **84**(1), 100–128 (1990)
7. Gugat, M.: Optimal switching boundary control of a string to rest in finite time. Z. Angew. Math. Mech. **88**(4), 283–305 (2008)
8. Hante, F.M.: Hybrid dynamics comprising modes governed by partial differential equations: modeling analysis and control for semilinear hyperbolic systems in one space dimension. Dissertation, University Erlangen-Nuremberg, Erlangen, Germany (2010). [Online]
9. Hante, F.M., Leugering, G.: Optimal boundary control of convention-reaction transport systems with binary control functions. In: Hybrid Systems: Computation and Control. Lecture Notes in Comput. Sci., vol. 5469, pp. 209–222. Springer, Berlin (2009)
10. Hante, F.M., Leugering, G., Seidman, T.I.: Modeling and analysis of modal switching in networked transport systems. Appl. Math. Optim. **59**(2), 275–292 (2009)
11. Hante, F.M., Leugering, G., Seidman, T.I.: An augmented BV setting for feedback switching control. J. Syst. Sci. Complex. **23**(3), 456–466 (2010)
12. Hinze, M., Pinnau, R., Ulbrich, M., Ulbrich, S.: Optimization with PDE Constraints. Mathematical Modelling: Theory and Applications, vol. 23. Springer, New York (2009)
13. Iftime, O.V., Demetriou, M.A.: Optimal control of switched distributed parameter systems with spatially scheduled actuators. Automatica J. IFAC **45**(2), 312–323 (2009)
14. Kawajiri, Y., Biegler, L.: A nonlinear programming superstructure for optimal dynamic operations of simulated moving bed processes. Ind. Eng. Chem. Res. **45**(25), 8503–8513 (2006)
15. Kogut, P.I., Leugering, G.: Optimal Control Problems for Partial Differential Equations on Reticulated Domains. Systems and Control: Foundations and Applications. Springer, Berlin (2011)
16. Kreĭn, S.G.: Linear Differential Equations in Banach Space. Am. Math. Soc., Providence (1971). Translated from the Russian by J.M. Danskin, Translations of Mathematical Monographs, vol. 29
17. Leineweber, D., Bauer, I., Schäfer, A., Bock, H., Schlöder, J.: An efficient multiple shooting based reduced SQP strategy for large-scale dynamic process optimization (Parts I and II). Comput. Chem. Eng. **27**, 157–174 (2003)
18. Li, X.J., Yong, J.M.: Optimal Control Theory for Infinite-Dimensional Systems. Systems & Control: Foundations & Applications. Birkhäuser Boston, Boston (1995)
19. Malanowski, K., Tröltzsch, F.: Lipschitz stability of solutions to parametric optimal control problems for parabolic equations. Z. Anal. Anwend. **18**(2), 469–489 (1999)
20. Pazy, A.: Semigroups of Linear Operators and Applications to Partial Differential Equations. Applied Mathematical Sciences Series. Springer, New York (1983)
21. Sager, S.: Numerical methods for mixed–integer optimal control problems. Der andere Verlag, Tönning (2005). http://mathopt.de/PUBLICATIONS/Sager2005.pdf. ISBN 3-89959-416-9
22. Sager, S.: Reformulations and algorithms for the optimization of switching decisions in nonlinear optimal control. J. Process Control **19**(8), 1238–1247 (2009)
23. Sager, S., Bock, H., Diehl, M.: The integer approximation error in mixed-integer optimal control. Math. Program. (2010). doi:10.1007/s10107-010-0405-3
24. Sager, S., Bock, H.G., Reinelt, G.: Direct methods with maximal lower bound for mixed-integer optimal control problems. Math. Program., Ser. A **118**(1), 109–149 (2009)
25. Sager, S., Jung, M., Kirches, C.: Combinatorial integral approximation. Math. Methods Oper. Res. **73**(3), 363–380 (2011)
26. Sakata, T., Jackson, D.K., Mao, S., Marriott, G.: Optically switchable chelates: Optical control and sensing of metal ions. J. Org. Chem. **73**(1), 227–233 (2008)
27. van Sint Annaland, M., Kuipers, J., van Swaaij, W.: Safety analysis of switching between reductive and oxidative conditions in a reaction coupling reverse flow reactor. Chem. Eng. Sci. **56**(4), 1517–1524 (2001)
28. Tröltzsch, F.: Optimal control of partial differential equations Graduate Studies in Mathematics, vol. 112. Am. Math. Soc., Providence (2010). Theory, Methods and Applications. Translated from the 2005 German original by Jürgen Sprekels
29. Zuazua, E.: Switching control. J. Eur. Math. Soc. **13**, 85–117 (2011)