OPTIMAL CONTROL APPLICATIONS AND METHODS
Optim. Control Appl. Meth. (2009)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/oca.880

Incorporating a class of constraints into the dynamics of optimal control problems

K. Graichen[1,*,†] and N. Petit[2]

[1]*Automation and Control Institute, Vienna University of Technology, Gusshausstrasse 27–29/E376, A-1040 Vienna, Austria*
[2]*Centre Automatique et Systèmes, Unité Mathématiques et Systèmes, MINES ParisTech, 60, Boulevard Saint–Michel, 75272 Paris, France*

SUMMARY

A method is proposed to systematically transform a constrained optimal control problem (OCP) into an unconstrained OCP, which can be treated in the standard calculus of variations. The considered class of constraints comprises up to m input constraints and m state constraints with well-defined relative degree, where m denotes the number of inputs of the given nonlinear system. Starting from an equivalent normal form representation, the constraints are incorporated into a new system dynamics by means of saturation functions and differentiation along the normal form cascade. This procedure leads to a new unconstrained OCP, where an additional penalty term is introduced to avoid the unboundedness of the saturation function arguments if the original constraints are touched. The penalty parameter has to be successively reduced to converge to the original optimal solution. The approach is independent of the method used to solve the new unconstrained OCP. In particular, the constraints cannot be violated during the numerical solution and a successive reduction of the constraints is possible, e.g. to start from an unconstrained solution. Two examples in the single and multiple input case illustrate the potential of the approach. For these examples, a collocation method is used to solve the boundary value problems stemming from the optimality conditions.

Received 15 December 2007; Revised 24 September 2008; Accepted 16 January 2009

KEY WORDS: optimal control problem; state and input constraints; normal form; saturation functions; calculus of variations; boundary value problem

1. INTRODUCTION

This paper proposes a new method to solve constrained optimal control problems (OCPs). In general, numerical methods to solve OCPs can roughly be divided in two different classes. In *direct methods*, the OCP is discretized to obtain a finite-dimensional parameter optimization problem, see e.g. [1–6]. Well-known advantages of the direct approach are the good domain of convergence as well as the efficient handling of constraints. On the other hand, *indirect approaches* are based on the calculus of variations and require the solution of a two-point boundary value problem (BVP), see e.g. [7]. Indirect methods are known to show a fast numerical convergence in the neighborhood of the optimal solution and to deliver highly accurate solutions, which makes them particularly attractive in aerospace industries [8–11]. However, the handling of constraints via Pontryagin's maximum principle [12] is in general non-trivial, since the overall structure of the BVP depends on the sequence between

*Correspondence to: K. Graichen, Automation and Control Institute, Vienna University of Technology, Gusshausstrasse 27–29/E376, A-1040 Vienna, Austria.
†E-mail: graichen@acin.tuwien.ac.at

singular/nonsingular and unconstrained/constrained arcs and requires *a priori* knowledge concerning the structure of the optimal solution.

In order to avoid these problems in handling constraints, this paper exposes a method to incorporate a set of constraints of a given OCP (called OCP_x) into a new unconstrained one. For a nonlinear system with m inputs, the method can handle up to m state constraints and m (state-dependent) input constraints. The state constraints are required to have a well-defined relative degree in the sense of nonlinear geometric control. The technique is systematic and allows a straightforward numerical treatment of the new unconstrained OCP in the calculus of variations. For sake of simplicity, the main principles of the approach are presented for the single input case with one state constraint and one input constraint, and are then extended to multiple input systems.

In the first step, the system dynamics of OCP_x are transformed into a normal form consisting of an internal dynamics and a cascade of integrators with the state constraint function corresponding to the first variable. In these preliminary coordinates, an equivalent OCP_y is defined, where the constraints enter precisely at the top and at the bottom of the normal form cascade. In a next step, the constraints are represented by means of saturation functions and successive differentiation along the cascade. These substitutions propagate through the internal dynamics and eventually define a new unconstrained dynamics. Its trajectories have inverse images in the original coordinates, which intrinsically satisfy the constraints.

Using the saturation functions, a new $\mathrm{OCP}_\xi^\varepsilon$ is derived, which includes an additional penalty term with parameter ε to avoid unboundedness of the new states or input, if one of the original constraints is touched. The penalty parameter ε has to be successively reduced during the numerical solution of $\mathrm{OCP}_\xi^\varepsilon$ to eventually approach the constrained optimal solution of OCP_y. The systematic incorporation of the constraints into the formulation of $\mathrm{OCP}_\xi^\varepsilon$ has the advantage that the constraints cannot be violated during the numerical solution of $\mathrm{OCP}_\xi^\varepsilon$ and that the constraints can be successively reduced, e.g. to start from an unconstrained solution.

The paper is organized as follows. In Section 2, the considered OCP_x in the single input case is exposed and transformed to the equivalent OCP_y in the normal form coordinates. Section 3 describes the saturation function approach to incorporate the constraints in a new system representation and to derive a new unconstrained $\mathrm{OCP}_\xi^\varepsilon$ with an additional penalty term. In Section 4, the convergence properties of $\mathrm{OCP}_\xi^\varepsilon$ are studied for $\varepsilon \to 0$. Section 5 is devoted to the solution of $\mathrm{OCP}_\xi^\varepsilon$ by deriving the optimality conditions from the calculus of variations. A modified version of a standard Matlab BVP solver is shortly introduced to numerically solve the BVP stemming from the optimality conditions. An example system with state and input constraints illustrates the method and the numerical solution. Section 6 extends the results to multiple input systems and applies the concept to an example application with one state constraint and two input constraints. Finally, conclusions are given in Section 7, where some of the advantages of the proposed method are discussed.

2. PROBLEM FORMULATION IN THE SINGLE INPUT CASE

The considered OCP is initially introduced for nonlinear systems with one control and a set of one state constraint and one input constraint (Section 6 addresses the multiple input case). A normal form representation is derived by using the state constraint as linearizing output. In this way, the constraints appear at the top and the bottom of the normal form cascade, which is the basis for the saturation function approach presented in Section 3.

2.1. Optimal control problem

We consider a nonlinear control-affine single input system of the form

$$\dot{x} = f(x) + g(x)u, \quad x \in \mathbb{R}^n, \quad u \in \mathbb{R} \tag{1}$$

DOI: 10.1002/oca

with $f, g: \mathbb{R}^n \to \mathbb{R}^n$ being sufficiently smooth. The initial and (desired) final conditions are given by

$$x(0) = x_0, \quad \chi(x(T)) = 0 \tag{2}$$

with $\chi: \mathbb{R}^n \to \mathbb{R}^q$. It is assumed that for each input u, the system (1) and initial conditions in (2) yield a unique state x. The cost functional to be minimized is of the form

$$J(u) = \varphi(x(T)) + \int_0^T L(x, u, t)\,\mathrm{d}t \tag{3}$$

where the functions $\varphi: \mathbb{R}^n \to \mathbb{R}$ and $L: \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ are sufficiently smooth. The end time T is fixed for the sake of simplicity. The following two constraints are considered:

$$c(x) \in [c^-, c^+], \quad u \in [u^-(x), u^+(x)] \tag{4}$$

The function $c(x)$ of the state constraint is assumed to have a well-defined relative degree (in the sense of geometric nonlinear control) with respect to the dynamics (1). The second state-dependent input constraint corresponds to a mixed input-state constraint $d(x, u) \in [d^-, d^+]$, which is well defined with respect to u, i.e. $\partial d/\partial u \neq 0$, such that d can be inverted with respect to u.

In summary, we consider the following OCP, noted OCP_x, and postulate the existence of an (at least local) optimal solution in Assumption 1.

Problem OCP_x:

$$\begin{aligned} \text{minimize} \quad & J(u) = \varphi(x(T)) + \int_0^T L(x, u, t)\,\mathrm{d}t \\ \text{subject to} \quad & \dot{x} = f(x) + g(x)u \\ & x(0) = x_0, \quad \chi(x(T)) = 0 \\ & c(x) \in [c^-, c^+], \quad u \in [u^-(x), u^+(x)] \end{aligned}$$

Assumption 1
OCP_x has an optimal solution (u^*, x^*) with the optimal cost $J(u^*) = J^*$.

2.2. Normal form representation for state constraint

Following [13], the relative degree $r \leqslant n$ of the constraint function $c(x)$ at a point $x = x^0$ is defined by

$$\begin{aligned} & L_g L_f^i c(x) = 0, \quad i = 1, \ldots, r-2 \\ & L_g L_f^{r-1} c(x) \neq 0 \end{aligned} \tag{5}$$

where L_f and L_g denote the Lie derivatives along the vector fields $f(x)$ and $g(x)$. Literally, r reveals how many times the constraint function $c(x)$ has to be differentiated until the input u appears (see again [13]).

The constraint function $c(x)$ can be used as (partially) linearizing output to derive a change of coordinates

$$\begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} \theta_y(x) \\ \theta_z(x) \end{pmatrix} = \theta(x) \tag{6}$$

with $y = (y_1, \ldots, y_r)^\mathsf{T}$ and $\theta_y = (\theta_1, \ldots, \theta_r)^\mathsf{T}$ defined by

$$\begin{aligned} & y_1 = c(x) = \theta_1(x), \quad y_i = L_f^{i-1} c(x) = \theta_i(x) \\ & i = 2, \ldots, r \end{aligned} \tag{7}$$

The additional coordinates $z = \theta_z(x) \in \mathbb{R}^{n-r}$ are necessary to complete the transformation (6) if $r < n$. Since the relative degree of $c(x)$ is assumed to be well defined in a sufficiently large neighborhood of the point x^0, the Jacobian $\partial\theta/\partial x$ is non-singular such that the inverse transformation

$$x = \theta^{-1}(y, z) \tag{8}$$

exists. In these coordinates, we can transform OCP_x into an equivalent OCP_y as follows:

Problem OCP_y:

$$\begin{aligned} \text{minimize} \quad & \bar{J}(u) = \bar{\varphi}(y(T), z(T)) \\ & \qquad + \int_0^T \bar{L}(y, z, u, t)\,\mathrm{d}t && \text{(9a)} \\ \text{subject to} \quad & \dot{y}_i = y_{i+1}, \quad i = 1, \ldots, r-1 && \text{(9b)} \\ & \dot{y}_r = a_0(y, z) + a_1(y, z)u && \text{(9c)} \\ & \dot{z} = b_0(y, z) + b_1(y, z)u && \text{(9d)} \end{aligned}$$

DOI: 10.1002/oca

$$y(0)=\theta_y(x_0),\quad z(0)=\theta_z(x_0)$$

$$\bar{\chi}(y(T),z(T))=0 \tag{9e}$$

$$y_1\in[c^-,c^+],\quad u\in[\bar{u}^-(y,z),\bar{u}^+(y,z)] \tag{9f}$$

where $\bar{\varphi}=\varphi\circ\theta^{-1}$, $\bar{L}=L\circ\theta^{-1}$, $\bar{\chi}=\chi\circ\theta^{-1}$, $\bar{u}^{\pm}=u^{\pm}\circ\theta^{-1}$ follow from OCP_x with the change of coordinates (6).

The notation '$\circ$' is consistently used throughout the text as substitution rule to replace a specific argument of a function $p(\cdot,v,\cdot)$ by a given transformation $v=q(w)$, i.e. $p(\cdot,q(w),\cdot)=p(\cdot,v,\cdot)\circ q$.

In nonlinear control, the dynamics (9b)–(9d) are often called input–output normal form, see e.g. [13]. The chain of integrators (9b)–(9c) with the functions $a_0=L_f^r c(x)\circ\theta^{-1}$ and $a_1=L_g L_f^{r-1}c(x)\circ\theta^{-1}$ are the input–output dynamics, where the transformed constraints (9f) appear at the top and bottom of the cascade. The second part (9d) of the dynamics with $b_0=L_f\theta_z(x)\circ\theta^{-1}$ and $b_1=L_g\theta_z(x)\circ\theta^{-1}$ represents the internal dynamics of the normal form. It is always possible in the single input case to choose $z=\theta_z(x)$ such that the internal dynamics (9d) are independent of the input u [13]. Nevertheless, (9d) is the more general form.

The last Equation (9c) of the input–output dynamics can be inverted to determine the input u in dependence of the states y,z, and $\dot{y}_r$:

$$u=\frac{\dot{y}_r-a_0(y,z)}{a_1(y,z)} \tag{10}$$

where $a_1(y,z)\neq 0$ due to the well-defined relative degree of $c(x)$, see (5).

Note that OCP_x and OCP_y are two different forms of the same OCP due to the one-to-one correspondence of the coordinates x and (y,z) via the transformations (6) and (8). This property leads to the following proposition:

Proposition 1
Under Assumption 1 and due to the bijective state transformation (6), OCP_y has an optimal solution $(u^*, y^*=\theta_y(x^*), z^*=\theta_z(x^*))$ with optimal cost $\bar{J}(u^*)=J^*$.

3. USING SATURATION FUNCTIONS TO REPRESENT THE CONSTRAINTS

This section presents an approach to transform the constrained OCP_y to a new unconstrained OCP and utilizes ideas from [14] originally developed in the context of feedforward control design. The proposed method takes advantage of the normal form cascade (9b)–(9c) and systematically incorporates the constraints on y_1 and u within a new system representation by means of saturation functions and successive differentiation of y_1. The derived unconstrained system defines a new $\text{OCP}_\xi^\varepsilon$ that contains an additional penalty term with parameter ε in order to avoid unboundedness of the saturation function arguments if the constraints are touched.

3.1. Derivation of new system representation

The idea of the approach is to replace the coordinates y and the corresponding input–output dynamics (9b)–(9c) by a new system in unconstrained coordinates that automatically satisfies the constraints (9f). In the first step, the state constraint $y_1\in[c^-,c^+]$ is replaced by a saturation function

$$y_1=\psi(\xi_1,c^{\pm})\in(c^-,c^+) \tag{11}$$

with the new unconstrained variable $\xi_1\in\mathbb{R}$. The saturation function $\psi(\xi_1,c^{\pm})$ is assumed to be smooth and strictly monotonically increasing, i.e. $\partial\psi/\partial\xi_1>0$ for all $\xi_1\in\mathbb{R}$, which means that the limits $c^{\pm}$ are only reached asymptotically for $\xi_1\to\pm\infty$, as shown in Figure 1.

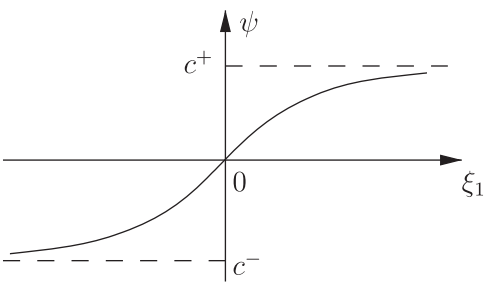

Figure 1. Asymptotic saturation function (11) with saturation limits $c^{\pm}$ and coordinate ξ_1.

DOI: 10.1002/oca

Hence, the constraint $y_1 \in [c^- c^+]$ is actually satisfied on the open intervals (c^-, c^+).

In order to substitute the next coordinate y_2, Equation (11) is differentiated and a new coordinate ξ_2 is introduced:

$$\dot{y}_1 = y_2 = \psi' \dot{\xi}_1 \quad \text{with } \dot{\xi}_1 = \xi_2 \tag{12}$$

with the notation $\psi' = (\partial\psi/\partial\xi_1)(\xi_1, c^\pm)$. By introducing ξ_2, a first differential equation $\dot{\xi}_1 = \xi_2$ is derived for the previous coordinate ξ_1. Further differentiation of (12) leads to (if $r > 2$)

$$\dot{y}_2 = y_3 = \psi'' \xi_2^2 + \psi' \dot{\xi}_2 \quad \text{with } \dot{\xi}_2 = \xi_3 \tag{13}$$

The Equations (12) and (13) show the concept behind the successive differentiation $y_1^{(i)} = y_{i+1}$ and the introduction of new coordinates $\dot{\xi}_i = \xi_{i+1}$ until y_r is reached. Hence, the following relations are obtained:

$$y_1 = h_1(\xi_1) = \psi(\xi_1, c^\pm) \tag{14a}$$

$$\begin{aligned} y_i &= h_i(\xi_1, \ldots, \xi_i) \\ &= \gamma_i(\xi_1, \ldots, \xi_{i+1}) + \psi' \xi_i, \quad i = 2, \ldots, r \end{aligned} \tag{14b}$$

where the nonlinear terms γ_i are determined with respect to the previous equation for y_{i-1}, i.e. $\gamma_2(\xi_1) = 0$ and

$$\gamma_i(\xi_1, \ldots, \xi_{i-1}) = \sum_{j=1}^{i-2} \frac{\partial h_{i-1}}{\partial \xi_j} \xi_{j+1}, \quad i = 3, \ldots, r$$

As a result, the successive differentiation of y_1 leads to a new set of coordinates $\xi = (\xi_1, \ldots, \xi_r)^\mathsf{T}$ that replaces the normal form coordinates y by the relation

$$y = h(\xi) = (h_1(\xi_1), \ldots, h_r(\xi))^\mathsf{T} \tag{15}$$

The final differentiation of $y_r = h_r(\xi)$ eventually gives

$$\dot{y}_r = \gamma_{r+1}(\xi) + \psi' \dot{\xi}_r \tag{16}$$

Since $\dot{y}_r$ is affected by the input u via the final equation (9c) of the input–output dynamics, the input constraint in (9f) can be interpreted as the constraint

$$\dot{y}_r \in [a^-(y, z), a^+(y, z)] \tag{17a}$$

where $a^\pm(y, z)$ is defined with respect to the (constant) sign of $a_1(y, z) \neq 0$:

$$a^\pm(y, z) = \begin{cases} a_0(y, z) + a_1(y, z)\bar{u}^\pm(y, z) & \text{if } a_1(y, z) > 0 \\ a_0(y, z) + a_1(y, z)\bar{u}^\mp(y, z) & \text{if } a_1(y, z) < 0 \end{cases} \tag{17b}$$

In order to incorporate the constraint (17a) in Equation (16), a second saturation function

$$\dot{\xi}_r = \phi(\tilde{u}, \phi^\pm) \in (\phi^-, \phi^+) \tag{18a}$$

with a new input $\tilde{u}$ is introduced, which eventually will substitute the original input u in the equations. Similar to $\psi(\xi_1, c^\pm)$ shown in Figure 1, the function $\phi(\tilde{u}, \phi^\pm)$ is smooth and reaches the limits $\phi^\pm$ only for $\tilde{u} \to \pm\infty$.[‡] The saturation limits $\phi^\pm$ have to be chosen such that the constraint (17a) is satisfied. Using (16) and (18a), the inequalities $a^-(y, z) \leqslant \dot{y}_r \leqslant a^-(y, z)$ can be written as

$$\frac{\tilde{a}^-(\xi, z) - \gamma_{r+1}(\xi)}{\psi'(\xi_1, c^\pm)} \leqslant \phi(\tilde{u}, \phi^\pm) \leqslant \frac{\tilde{a}^+(\xi, z) - \gamma_{r+1}(\xi)}{\psi'(\xi_1, c^\pm)}$$

with $\psi' > 0$ and $\tilde{a}^\pm = a^\pm \circ h$ being expressed in the new coordinates (15). Hence, the saturation limits $\phi^\pm$ directly follow to

$$\phi^\pm := \phi^\pm(\xi, z) = \frac{\tilde{a}^\pm(\xi, z) - \gamma_{r+1}(\xi)}{\psi'(\xi_1, c^\pm)} \tag{18b}$$

and thus depend on the states ξ and z in order to satisfy (17a). Inserting (18) in (17a) leads to

$$\begin{aligned} \dot{y}_r &= h_{r+1}(\xi, z, \tilde{u}) \\ &= \gamma_{r+1}(\xi) + \psi' \phi(\tilde{u}, \phi^\pm(\xi, z)) \end{aligned} \tag{19}$$

in addition to the relations (15) for the coordinates y. Finally, the original input u can be expressed in terms

[‡]The explicit formulas for the saturation functions $\psi(\xi_1, c^\pm)$ and $\phi(\tilde{u}, \phi^\pm)$ used in this paper are stated in (A1) and (A2) in Appendix A.1.

DOI: 10.1002/oca

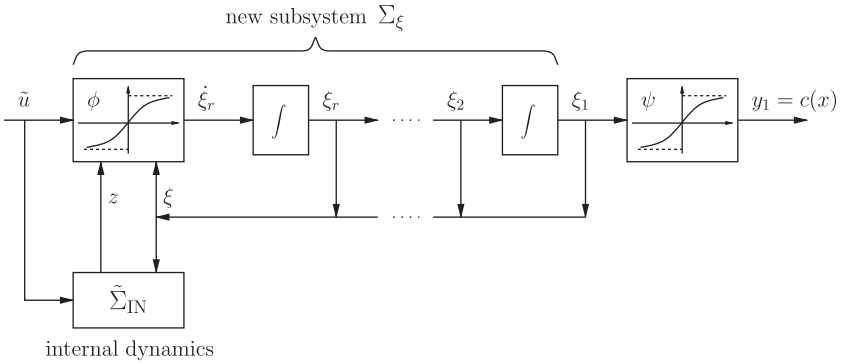

Figure 2. New normal form (21) with subsystem Σ_ξ and transformed internal dynamics $\tilde{\Sigma}_{\text{IN}}$.

of the states (ξ, z) and the new input $\tilde{u}$ by using the inverse dynamics (10) and (15), (19):

$$u = h_u(\xi, z, \tilde{u}) = \frac{h_{r+1}(\xi, z, \tilde{u}) - \tilde{a}_0(\xi, z)}{\tilde{a}_1(\xi, z)} \tag{20}$$

with $\tilde{a}_0 = a_0 \circ h$ and $\tilde{a}_1 = a_1 \circ h$.

Owing to the introduced saturation functions $\psi(\xi_1, c^\pm)$ and $\phi(\tilde{u}, \phi^\pm(\xi, z))$ and the successive differentiation of y_1, the coordinates y and input u are replaced by ξ and the new input $\tilde{u}$, which leads to the new representation of the normal form (9b)–(9d)

$$\Sigma_\xi\colon\ \dot{\xi}_i = \xi_{i+1}, \quad i = 1, \ldots, r-1 \tag{21a}$$

$$\dot{\xi}_r = \tilde{\phi}(\xi, z, \tilde{u}) = \phi(\tilde{u}, \phi^\pm(\xi, z)) \tag{21b}$$

$$\tilde{\Sigma}_{\text{IN}}\colon\ \dot{z} = \tilde{b}(\xi, z, \tilde{u}) = \tilde{b}_0(\xi, z) + \tilde{b}_1(\xi, z) h_u(\xi, z, \tilde{u}) \tag{21c}$$

with $\tilde{b}_0 = b_0 \circ h$ and $\tilde{b}_1 = b_1 \circ h$. The block diagram in Figure 2 illustrates the structure of the new system (21). The two saturation functions $\psi(\xi_1, c^\pm)$ and $\phi(\tilde{u}, \phi^\pm(\xi, z))$ are arranged at the top and bottom of the chain of integrators, whereby the states ξ feed back into the modified internal dynamics $\tilde{\Sigma}_{\text{IN}}$ and (together with z) into $\phi(\tilde{u}, \phi^\pm(\xi, z))$ to determine the saturation limits $\phi^\pm(\xi, z)$.

Remark 1

The structure of the saturation limits $\phi^\pm(\xi, z)$ in (18b) has a particular advantage if the state constraint $y_1 \in [c^-, c^+]$ is approached, which implies $\psi'(\xi_1, c^\pm) \to 0$. For certain properties of the constraint functions (17b), it can be shown that the limits $\phi^\pm(\xi, z)$ approach $\pm\infty$ for $\psi'(\xi_1, c^\pm) \to 0$, thus 'opening' the (normalized) saturation function in (A2) (see Appendix A.1), i.e. $\phi(\tilde{u}, \phi^\pm(\xi, z)) \approx \tilde{u}$. More details can be found in Appendix A.2 and Section 5.3.

3.2. Inverse relations

An important point is that the strict monotonicity of the saturation functions (11) and (18) ensures the one-to-one correspondence between the original (constrained) normal form coordinates $y_1 \in (c^-, c^+)$ with $\dot{y}_i = y_{i+1}$ and the new (unconstrained) coordinates $\xi = (\xi_1 < \infty, \ldots, \xi_r)^\mathsf{T}$ related by $\dot{\xi}_i = \xi_{i+1}$, $i = 1, \ldots, r-1$. The unique correspondence is defined from the inverse relations of (15)

$$\xi_1 = \psi^{-1}(y_1, c^\pm) = h_1^{-1}(y_1) \tag{22a}$$

$$\xi_i = \frac{y_i - \gamma_i(\xi_1, \ldots, \xi_{i-1})}{\psi'(\xi_1, c^\pm)} = h_i^{-1}(\xi_1, \ldots, \xi_{i-1}, y_i), \quad i = 2, \ldots, r \tag{22b}$$

which successively determine the coordinates ξ. In summary, the overall inverse relation to (15) is denoted by

$$\xi = h^{-1}(y) \tag{23}$$

with $h^{-1}\colon (c^-, c^+) \times \mathbb{R}^{r-1} \to \mathbb{R}^r$. Owing to the asymptotic nature of the saturation function $\psi(\xi_1, c^\pm)$, the

DOI: 10.1002/oca

coordinate $\xi_1 = \psi^{-1}(y_1, c^{\pm})$ becomes unbounded if y_1 touches one of the constraints $c^{\pm}$. For the particular choice (A1) of $\psi(\xi_1, \psi^{\pm})$ in Appendix A.1, the inverse function can be explicitly stated as

$$\begin{aligned}\xi_1 &= \psi^{-1}(y_1, c^{\pm}) \\ &= \tfrac{1}{4}(c^+ - c^-)[\log(y_1 - c^-) - \log(c^+ - y_1)] \quad (24)\end{aligned}$$

where the two log-terms lead to unboundedness of ξ_1 if y_1 touches either c^- or c^+, see Figure 1.

Similarly, the second saturation function $\phi(\tilde{u}, \phi^{\pm})$ can be solved for the new input $\tilde{u}$ on the open intervals of the constraints (9f). Using (19), the new input $\tilde{u}$ can be formally written with the inverse relation

$$\tilde{u} = \phi^{-1}\left(\frac{\dot{y}_r - \gamma_{r+1}(\xi)}{\psi'(\xi_1, c^{\pm})}, \phi^{\pm}(\xi, z)\right) \circ h^{-1} \quad (25a)$$

In addition, $\dot{y}_r$ is replaced by the right-hand side of (9c), which formally leads to

$$\tilde{u} = h_{\tilde{u}}(y, z, u) \quad (25b)$$

where the function $h_{\tilde{u}} : (c^-, c^+) \times \mathbb{R}^{r-1} \times \mathbb{R}^{n-r} \times (\bar{u}^-, \bar{u}^+) \to \mathbb{R}$ with $\bar{u}^{\pm} := \bar{u}^{\pm}(y, z)$ has bounded values on the open intervals of the constraints (9f) (see also (17a)). This equation can be simplified, in the case when (A2) in Appendix A.1 is used, by inserting (17b) and (18b):

$$\begin{aligned}\tilde{u} = \frac{a_1(y,z)(\bar{u}^{\pm} - \bar{u}^{\mp})}{4\psi'(\xi_1, c^{\pm}) \circ \psi^{-1}} [\log|u - \bar{u}^{\mp}| - \log|\bar{u}^{\pm} - u|] \\ \text{for} \quad a_1(y,z) \gtrless 0 \quad (26)\end{aligned}$$

Similar to the previous considerations for ξ_1, the two log terms show that $\tilde{u}$ becomes unbounded if u reaches one of its constraints $\bar{u}^{\pm}(y, z)$.

3.3. *New penalized* $\mathrm{OCP}^{\varepsilon}_{\xi}$

The derived system (21) is used to define a new unconstrained OCP with respect to the new input $\tilde{u}$. The cost functional $\bar{J}(u)$ of the previous OCP_y can be expressed in the new coordinates (ξ, z) and $\tilde{u}$ by

$$\tilde{J}(\tilde{u}) = \tilde{\varphi}(\xi(T), z(T)) + \int_0^T \tilde{L}(\xi, z, \tilde{u}, t)\, \mathrm{d}t \quad (27a)$$

with the substituted cost terms $\tilde{\varphi} = \bar{\varphi} \circ h$ and $\tilde{L} = L \circ h \circ h_u$. However, as it was discussed in the last section, the state ξ_1 and input $\tilde{u}$ become unbounded if one of the constraints (9f) is touched. This problem is taken into account by adding an additional penalty term

$$p(\tilde{u}) = \int_0^T \xi_1^2 + \tilde{u}^2\, \mathrm{d}t \quad (27b)$$

to the cost $\tilde{J}(\tilde{u})$. This yields the following penalized $\mathrm{OCP}^{\varepsilon}_{\xi}$ with penalty parameter ε and repeating the dynamics (21) for the sake of completeness, we have

Problem $\mathrm{OCP}^{\varepsilon}_{\xi}$:

$$\begin{aligned}&\text{minimize} && P(\tilde{u}, \varepsilon) = \tilde{J}(\tilde{u}) + \varepsilon p(\tilde{u}) && (28a)\\ &\text{subject to} && \dot{\xi}_i = \xi_{i+1}, \quad i = 1, \ldots, r-1 && (28b)\\ & && \dot{\xi}_r = \tilde{\phi}(\xi, z, \tilde{u}) = \phi(\tilde{u}, \phi^{\pm}(\xi, z)) && (28c)\\ & && \dot{z} = \tilde{b}(\xi, z, \tilde{u}) \\ & && \quad = \tilde{b}_0(\xi, z) + \tilde{b}_1(\xi, z) h_u(\xi, z, \tilde{u}) && (28d)\\ & && \xi(0) = h^{-1}(\theta(x_0)), \quad z(0) = \theta_z(x_0) \\ & && \tilde{\chi}(\xi(T), z(T)) = 0 && (28e)\end{aligned}$$

where $\tilde{b}_0 = b_0 \circ h, \tilde{b}_1 = b_1 \circ h$ and $\tilde{\chi} = \bar{\chi} \circ h$ follow from OCP_y. The constraints (9f) are incorporated in the dynamics by the asymptotic saturation functions $\psi(\xi_1, c^{\pm})$ and $\phi(\tilde{u}, \phi^{\pm}(\xi, z))$ with the variable limits $\phi^{\pm}(\xi, z)$ defined in (18b). Their successive derivatives uniquely define $y = h(\xi)$, $\dot{y}_r = h_{r+1}(\xi, z, \tilde{u})$, and $u = h_u(\xi, z, \tilde{u})$ stated in (15), (19), and (20).

Note that the penalized $\mathrm{OCP}^{\varepsilon}_{\xi}$ is truly unconstrained because the constraints (9f) are incorporated in the normal form dynamics (28b)–(28d). In practice, the new $\mathrm{OCP}^{\varepsilon}_{\xi}$ will be successively solved with decreasing values of the penalty parameter $\varepsilon \to 0$. Before we discuss convergence, we state the following assumption:

Assumption 2
For each penalty parameter $\varepsilon > 0$, $\mathrm{OCP}^{\varepsilon}_{\xi}$ has an optimal solution $(\tilde{u}^{\varepsilon}, \xi^{\varepsilon}, z^{\varepsilon})$. Moreover, this solution has bounded components $\tilde{u}^{\varepsilon}$ and ξ_1^{ε}.

DOI: 10.1002/oca

Assumption 2 is reasonable from a practical point of view to ensure solvability of $\mathrm{OCP}_\xi^\varepsilon$. Moreover, the assumption of boundedness guarantees that the inverse transformations (23) and (25) are well defined, which implies that y_1 and u strictly remain inside the constraints (9f).

By successively decreasing $\varepsilon \to 0$, one intuitively expects that the penalized cost $P(\tilde{u}^\varepsilon, \varepsilon)$ converges to the optimal value J^* and that $y^\varepsilon = h(\xi^\varepsilon, z^\varepsilon), z^\varepsilon$, and $u^\varepsilon = h_u(\xi^\varepsilon, z^\varepsilon, \tilde{u}^\varepsilon)$ converge to the optimal solution (u^*, y^*, z^*) of OCP_y (see Proposition 1). This point is addressed in the next section.

Remark 2
Under certain assumptions, the penalization of ξ_1 used to avoid its unboundedness is not required, since $\tilde{u}$ also becomes unbounded if y_1 touches one of the constraints $c^\pm$. Moreover, in these cases, $\tilde{u}$ gives an infinite penalty value $\int_0^T \tilde{u}^2 \,\mathrm{d}t$ (locally not square-summable), which automatically implies that y_1 strictly stays inside the constraints (c^-, c^+) and therefore ξ_1 and $\tilde{u}$ remain bounded. These points are developed in Appendix A.2.

4. INVESTIGATION OF CONVERGENCE

This section investigates the convergence of the cost and states of $\mathrm{OCP}_\xi^\varepsilon$ for the penalty parameter $\varepsilon \to 0$. Although the variables ξ^ε and $\tilde{u}^\varepsilon$ as part of the solution of $\mathrm{OCP}_\xi^\varepsilon$ may become unbounded in the limit $\varepsilon \to 0$, the convergence of the trajectories in the (y, z)-coordinates can additionally be concluded under assumption of strong convexity on OCP_y.

4.1. Some definitions and further assumptions

Several norms are used in the following. The Euclidian norm for a vector $p \in \mathbb{R}^q$ is denoted by $\|p\|$. For time (vector) functions $p(t) \in \mathbb{R}^q$ defined on $t \in [0, T]$, the standard norms $L^i(0, T; \mathbb{R}^q), i = 1, 2, \infty$ are used and denoted by $\|p\|_1, \|p\|_2$, and $\|p\|_\infty$.

Moreover, some definitions and assumptions are necessary, which are directly stated for OCP_y and not for the original OCP_x for the sake of simplicity. Define the set

$$\begin{aligned} S = \{u \in L^\infty(0, T; \mathbb{R}) :\ & y(0) = \theta_y(x_0), \\ & z(0) = \theta_z(x_0), \\ & \bar{\chi}(y(T), z(T)) = 0, \\ & y_1 \in [c^-, c^+], \\ & u \in [\bar{u}^-(y, z), \bar{u}^+(y, z)] \\ & \forall t \in [0, T]\} \end{aligned} \tag{29}$$

denoting the set of admissible inputs u, which—together with their associated unique states (y, z) following from the dynamics (9b)–(9d)—satisfy the boundary conditions (9e) and constraints (9f). With the definition of S, OCP_y can alternatively be stated as

$$\mathrm{OCP}_y : \min_{u \in S} J(u)$$

with the optimal solution $\bar{J}(u^*) = J^*$ (see Proposition 1). In order to allow statements concerning the convergence of $\mathrm{OCP}_\xi^\varepsilon$, define the following subset of admissible inputs u for which the constraints (9f) are strictly satisfied on the open intervals:

$$\begin{aligned} S^0 = \{u \in S :\ & y_1 \in (c^-, c^+), \\ & u \in (\bar{u}^-(y, z), \bar{u}^+(y, z))\ \forall t \in [0, T]\} \end{aligned} \tag{30}$$

For each admissible input $u \in S^0$, the inverse relations (23) and (25) exist and yield bounded variables ξ and $\tilde{u}$. This defines the image $\tilde{S}^0$ of set S^0 as

$$\tilde{S}^0 = \{\tilde{u} = h_{\tilde{u}}(y, z, u) : u \in S^0\} \tag{31}$$

with respect to all $u \in S^0$ and the corresponding states (y, z). Hence, each new input $\tilde{u} \in \tilde{S}^0$ is admissible in the sense that its associated states (ξ, z) are bounded and unique and satisfy the boundary conditions (28e). This allows to reformulate $\mathrm{OCP}_\xi^\varepsilon$ as

$$\mathrm{OCP}_\xi^\varepsilon :\ \min_{\tilde{u} \in \tilde{S}^0} P(\tilde{u}, \varepsilon)$$

DOI: 10.1002/oca

Note that $\tilde{S}^0$ is non-empty due to Assumption 2, which in turn implies the non-emptiness of S^0.

Finally, we impose the following additional assumptions on OCP_y and not on $\mathrm{OCP}_\xi^\varepsilon$ for the sake of generality:

Assumption 3

(a) The functional $\bar{J}$ is continuous in u for all $u \in S$.
(b) The optimal control u^* lies in the closure of S^0.
(c) The functions a_0, a_1, b_0, and b_1 satisfy the linear growth and boundedness properties

$$\|a_0(\bar{x})\| \leqslant \bar{a}_0(1+\|\bar{x}\|), \quad \|a_1(\bar{x})\| \leqslant \bar{a}_1$$

$$\|b_0(\bar{x})\| \leqslant \bar{b}_0(1+\|\bar{x}\|), \quad \|b_1(\bar{x})\| \leqslant \bar{b}_1 \quad \forall \bar{x} = \begin{pmatrix} y \\ z \end{pmatrix} \in \mathbb{R}^n \tag{32}$$

$\forall \bar{x}^\mathsf{T} = (y^\mathsf{T}, z^\mathsf{T}) \in \mathbb{R}^n$ and for some constants $\bar{a}_0$, $\bar{a}_1$, $\bar{b}_0$, and $\bar{b}_1$.

Although we already assumed uniqueness of the states $\bar{x}$, the properties (32) of the (sufficiently smooth) functions a_0, a_1, b_1, b_2 of the normal form (9b)–(9d) state more precisely that there exists a unique and bounded solution $\bar{x}$ to each input $u \in S$. Moreover, an important consequence is that two solutions $\bar{x}_u$ and $\bar{x}_v$ for associated inputs u and v satisfy

$$\|\bar{x}_u - \bar{x}_v\|_\infty \leqslant C\|u-v\|_1 \quad \forall u, v \in S \tag{33}$$

for some constant $C>0$. The proof can be found in [15] using (32) and Gronwall's lemma. The additional assumption that the optimal control $u^* \in S$ also lies in the closure of S^0 is necessary to ensure that u^* can be approached from within S^0, see e.g. [16] for a similar assumption in the context of interior point methods.

4.2. Convergence results

The proof of convergence is adapted from the results of Lasdon *et al.* [17] for OCPs with interior penalty functions. Since $\mathrm{OCP}_\xi^\varepsilon$ has to be successively solved for decreasing penalty parameters $\varepsilon^{k+1} < \varepsilon^k$, the following lemma is of importance concerning the non-increase of the cost (28a):

Lemma 1
Let $\tilde{u}^{k+1}$ and $\tilde{u}^k$ be the optimal controls of $\mathrm{OCP}_\xi^\varepsilon$ for $0 < \varepsilon^{k+1} < \varepsilon^k$. Then, the following inequalities hold for the cost functional (28a):

$$\tilde{J}(\tilde{u}^{k+1}) \leqslant \tilde{J}(\tilde{u}^k), \quad p(\tilde{u}^{k+1}) \geqslant p(\tilde{u}^k)$$

$$P(\tilde{u}^{k+1}, \varepsilon^{k+1}) \leqslant P(\tilde{u}^k, \varepsilon^k) \tag{34}$$

Proof
The proof directly follows from [16]. Since the optimal controls $\tilde{u}^k$ and $\tilde{u}^{k+1}$ minimize the cost (27a) for ε^{k+1} and ε^k, the following inequalities are true:

$$\tilde{J}(\tilde{u}^k) + \varepsilon^k p(\tilde{u}^k) \leqslant \tilde{J}(\tilde{u}^{k+1}) + \varepsilon^k p(\tilde{u}^{k+1}) \tag{35a}$$

$$\tilde{J}(\tilde{u}^{k+1}) + \varepsilon^{k+1} p(\tilde{u}^{k+1}) \leqslant \tilde{J}(\tilde{u}^k) + \varepsilon^{k+1} p(\tilde{u}^k) \tag{35b}$$

Multiplying the first inequality with $\varepsilon^{k+1}/\varepsilon^k$ (which satisfies $0 < \varepsilon^{k+1}/\varepsilon^k < 1$) and adding the resulting inequality to the second one gives

$$\tilde{J}(\tilde{u}^{k+1})\left(1 - \frac{\varepsilon^{k+1}}{\varepsilon^k}\right) \leqslant \tilde{J}(\tilde{u}^k)\left(1 - \frac{\varepsilon^{k+1}}{\varepsilon^k}\right) \tag{35c}$$

Since $0 < \varepsilon^{k+1} < \varepsilon^k$, it follows that $\tilde{J}(\tilde{u}^{k+1}) \leqslant \tilde{J}(\tilde{u}^k)$. Using this result in (35a) and dividing by $\varepsilon^k > 0$ leads to $p(\tilde{u}^k) \leqslant p(\tilde{u}^{k+1})$. The last inequality to be proven follows from the nested relations $P(\tilde{u}^{k+1}, \varepsilon^{k+1}) \leqslant P(\tilde{u}^k, \varepsilon^{k+1}) \leqslant P(\tilde{u}^k, \varepsilon^k)$. □

The following theorem concerns the convergence of the cost $P(\tilde{u}^k, \varepsilon^k)$ using the results of Lemma 1:

Theorem 1
Let $\{\varepsilon^k\}$ be a decreasing sequence of positive penalty parameters with $\lim_{k\to\infty} \varepsilon^k = 0$. Then, $P(\tilde{u}^k, \varepsilon^k)$ converges to the optimal cost

$$\lim_{k\to\infty} P(\tilde{u}^k, \varepsilon^k) = J^* \tag{36a}$$

with

$$\lim_{k\to\infty} \tilde{J}(\tilde{u}^k) = J^*, \quad \lim_{k\to\infty} \varepsilon^k p(\tilde{u}^k) = 0 \tag{36b}$$

DOI: 10.1002/oca

Proof
The proof of the theorem is adapted from [17]. Since $\bar{J}(u)$ is continuous over S and $u^* \in S$ also lies in the closure of S^0 (see Assumption 3), it follows that for any parameter $\delta J > 0$, one can always find an admissible input $u^\delta \in S^0$ with associated states (y^δ, z^δ) such that $\bar{J}(u^\delta) < J^* + \delta J/2$. For this $u^\delta \in S^0$, there exists a corresponding new (bounded) input $\tilde{u}^\delta = h_{\tilde{u}}(y^\delta, z^\delta, u^\delta) \in \tilde{S}^0$ with $\tilde{J}(\tilde{u}^\delta) = \bar{J}(u^\delta)$, which allows to rewrite the previous inequality as

$$\tilde{J}(\tilde{u}^\delta) < J^* + \delta J/2 \tag{37a}$$

Select ε^l such that

$$\varepsilon^l \, p(\tilde{u}^\delta) < \delta J/2 \tag{37b}$$

Then, for any $k > l$ with $\varepsilon^k < \varepsilon^l$ and using Lemma 1, it follows that

$$P(\tilde{u}^k, \varepsilon^k) \leqslant P(\tilde{u}^l, \varepsilon^l) \leqslant P(\tilde{u}^\delta, \varepsilon^l) \tag{37c}$$

where $\tilde{u}^k$ and $\tilde{u}^l$ are the optimal solutions for ε^k and ε^l, respectively. With (37a) and (37b), there exists an upper estimate on $P(\tilde{u}^\delta, \varepsilon^l)$ with

$$P(\tilde{u}^\delta, \varepsilon^l) < J^* + \delta J/2 + \delta J/2 = J^* + \delta J \tag{37d}$$

Finally, using $P(\tilde{u}^k, \varepsilon^k) > \tilde{J}(\tilde{u}^k) > J^* > J^* - \delta J$, Equations (37c) and (37d) lead to the conclusion that $\forall \delta J > 0, \exists l$ such that $\forall k > l, |P(\tilde{u}^k, \varepsilon^k) - J^*| < \delta J$. This proves (36a) and additionally (36b) by remembering that $p(\tilde{u}^k) > 0$. □

Note that until this point no convexity assumption was necessary to prove the convergence of the cost (36). In order to prove convergence of the states, we require the following strong convexity assumption:

Assumption 4
The cost functional $\bar{J}(u)$ of OCP_y satisfies the strong convexity property

$$\begin{aligned} &D\|u - v\|_2^2 \\ &\quad \leqslant \bar{J}(u) + \bar{J}(v) - 2\bar{J}\left(\tfrac{1}{2}u + \tfrac{1}{2}v\right) \quad \forall u, v \in S \end{aligned} \tag{38}$$

for some $D > 0$.

The strong convexity property (38) e.g. holds for linear systems with quadratic cost functional.

The strong convexity assumption (38) would imply uniqueness of the optimal control u^* if the set S was convex. However, since this is not known (in particular due to the presence of state constraints), we assume uniqueness in the next theorem to prove convergence of the trajectories.

Theorem 2
Assume that the optimal control u^* of OCP_y is unique and that Assumption 4 holds. Then, the input $u^k = h_u(\xi^k, z^k, \tilde{u}^k)$ as well as $y^k = h(\xi^k, z^k)$ and z^k following from the solution of $\mathrm{OCP}_\xi^\varepsilon$ with $\varepsilon^{k+1} < \varepsilon^k$ converge to the optimal trajectories (u^*, y^*, z^*) according to

$$\begin{aligned} &\lim_{k\to\infty} \|u^k - u^*\|_2 = 0, \quad \lim_{k\to\infty} \|y^k - y^*\|_\infty = 0 \\ &\lim_{k\to\infty} \|z^k - z^*\|_\infty = 0 \end{aligned} \tag{39}$$

Proof
The uniqueness of u^* ensures that $\bar{J}(u) \geqslant \bar{J}(u^*) \forall u \in S$. Hence, the strong convexity property (38) can be used to conclude

$$\begin{aligned} &D\|u - u^*\|_2^2 \\ &\quad \leqslant \bar{J}(u) + \bar{J}(u^*) - 2\bar{J}\left(\tfrac{1}{2}u + \tfrac{1}{2}u^*\right) \\ &\quad \leqslant \bar{J}(u) + \bar{J}(u^*) - 2\bar{J}(u^*) = \bar{J}(u) - \bar{J}(u^*) \end{aligned} \tag{40}$$

The uniqueness of u^* also ensures that $J^* = \bar{J}(u^*)$ holds. Hence, with the results of Theorem 1 and the equivalent cost values $\bar{J}(u^k) = \tilde{J}(\tilde{u}^k)$, it follows from (40) that $u^k = h_u(\xi^k, z^k, \tilde{u}^k)$ converges to u^* in L^2. The convergence of $y^k = h(\xi^k, z^k)$ and z^k relies on the linear growth property in Assumption 3. The L^1-norm in (33) can be related to L^2 used in (38) by means of Hölder's inequality, which leads to $\|u - u^*\|_1 \leqslant \sqrt{T}\|u - u^*\|_2$. Hence, (33) can be formulated as $\|\bar{x} - \bar{x}^*\|_\infty \leqslant C\|u - u^*\|_1 \leqslant C\sqrt{T}\|u - u^*\|_2$ with $\bar{x}^\mathsf{T} = (y^\mathsf{T}, z^\mathsf{T})$, which shows the convergence of y^k and z^k in L^∞. □

Note that no statement is made concerning the limits of ξ^k and $\tilde{u}^k$ for $k \to \infty$, since they become

DOI: 10.1002/oca

unbounded if the optimal trajectories y^* and u^* touch the constraints (9f). However, the corresponding mappings $h_u(\xi^k, z^k, \tilde{u}^k)$ and $h(\xi^k)$ converge to the optimal trajectories u^* and y^* as stated in Theorem 2.

5. NUMERICAL SOLUTION OF $\mathrm{OCP}_\xi^\varepsilon$

This section now focuses on the numerical solution of $\mathrm{OCP}_\xi^\varepsilon$. In the first step, the optimality conditions are derived, which lead to a two-point BVP for $\mathrm{OCP}_\xi^\varepsilon$. This BVP can be solved numerically by a collocation method, which is a modified version of a standard Matlab BVP solver. A simple example from the literature illustrates the derivation of $\mathrm{OCP}_\xi^\varepsilon$ and its numerical solution.

5.1. Necessary optimality conditions

The necessary optimality conditions for $\mathrm{OCP}_\xi^\varepsilon$ follow from the classical calculus of variations. For the sake of compactness, one can comprise the states in $\tilde{x}^\mathsf{T} = (\xi^\mathsf{T}, z^\mathsf{T})$ and rewrite the dynamics (28b)–(28d) and boundary conditions (28e) of $\mathrm{OCP}_\xi^\varepsilon$ under the form

$$\begin{aligned} \dot{\tilde{x}} &= \tilde{f}(\tilde{x}, \tilde{u}) \\ \tilde{x}(0) &= \begin{pmatrix} h^{-1}(\theta(x_0)) \\ \theta_z(x_0) \end{pmatrix}, \quad \tilde{\chi}(\tilde{x}(T)) = 0 \end{aligned} \tag{41}$$

where $\tilde{f}$ follows from the right-hand side of (28b)–(28d). Define the Hamiltonian

$$\begin{aligned} H(\tilde{x}, \lambda, \tilde{u}, t) &= \tilde{L}(\tilde{x}, \tilde{u}, t) + \varepsilon p(\tilde{u}) + \lambda^\mathsf{T} \tilde{f}(\tilde{x}, \tilde{u}) \\ &= \tilde{L}(\tilde{x}, \tilde{u}, t) + \varepsilon(\xi_1^2 + \tilde{u}^2) \\ &\quad + \sum_{i=1}^{r-1} \lambda_{\xi,i} \xi_{i+1} + \lambda_{\xi,r} \tilde{\phi}(\xi, z, \tilde{u}) \\ &\quad + \lambda_z^\mathsf{T} \tilde{b}(\xi, z, \tilde{u}) \end{aligned}$$

with the adjoint states $\lambda_\xi = (\lambda_{\xi_1}, \ldots, \lambda_{\xi,r})^\mathsf{T}$ and $\lambda^\mathsf{T} = (\lambda_\xi^\mathsf{T}, \lambda_z^\mathsf{T})$. Then, the minimization of H with respect to the new input $\tilde{u}$ is given by[§]

$$\frac{\partial H}{\partial \tilde{u}} = \frac{\partial \tilde{L}}{\partial \tilde{u}} + 2\varepsilon\tilde{u} + \lambda_{\xi,r} \frac{\partial \tilde{\phi}}{\partial \tilde{u}} + \lambda_z^\mathsf{T} \frac{\partial \tilde{b}}{\partial \tilde{u}} = 0 \tag{42a}$$

Using the transformations (15), (20), and the internal dynamics (28d), the partial derivatives of $\tilde{L}$ and $\tilde{b}$ become

$$\begin{aligned} \frac{\partial \tilde{L}}{\partial \tilde{u}} &= \left[\frac{\partial L}{\partial u} \circ h \circ h_u \right] \frac{\partial h_u}{\partial \tilde{u}} \\ \frac{\partial \tilde{b}}{\partial \tilde{u}} &= \tilde{b}_1(\xi, z) \frac{\partial h_u}{\partial \tilde{u}} \end{aligned} \tag{42b}$$

The term $\partial h_u / \partial \tilde{u}$ can be further detailed with the help of (20)[¶]

$$\frac{\partial h_u}{\partial \tilde{u}} = \frac{1}{\tilde{a}_1(\xi, z)} \frac{\partial h_{r+1}}{\partial \tilde{u}} = \frac{\psi'(\xi_1, c^\pm)}{\tilde{a}_1(\xi, z)} \frac{\partial \tilde{\phi}}{\partial \tilde{u}} \tag{42c}$$

The adjoint system for λ is defined by $\dot{\lambda}^\mathsf{T} = -\partial H / \partial \tilde{x}$, which can be written in more detail as

$$\dot{\lambda}_{\xi,1} = -\frac{\partial \tilde{L}}{\partial \xi_1} - 2\varepsilon\xi_1 - \lambda_{\xi,r} \frac{\partial \tilde{\phi}}{\partial \xi_1} - \lambda_z^\mathsf{T} \frac{\partial \tilde{b}}{\partial \xi_1} \tag{43a}$$

$$\begin{aligned} \dot{\lambda}_{\xi,i} &= -\frac{\partial \tilde{L}}{\partial \xi_i} - \lambda_{\xi,i-1} - \lambda_{\xi,r} \frac{\partial \tilde{\phi}}{\partial \xi_i} - \lambda_z^\mathsf{T} \frac{\partial \tilde{b}}{\partial \xi_i} \\ & \qquad i = 2, \ldots, r \end{aligned} \tag{43b}$$

$$\dot{\lambda}_z^\mathsf{T} = -\frac{\partial \tilde{L}}{\partial z} - \lambda_{\xi,r} \frac{\partial \tilde{\phi}}{\partial z} - \lambda_z^\mathsf{T} \frac{\partial \tilde{b}}{\partial z} \tag{43c}$$

[§] For the numerical solution of the optimality conditions (see Section 5.2), we assume $\partial^2 H / \partial \tilde{u}^2 > 0$ (strengthened Legendre–Clebsch condition). Note that the positive definiteness of $\partial^2 H / \partial \tilde{u}^2$ represents a sufficient (second-order) optimality condition.

[¶] The detailed expressions in (42) allow some further statements concerning the penalty term $\varepsilon(\xi_1^2 + \tilde{u}^2)$. If one of the input constraint in (9f), or alternatively (17a), is approached, the second saturation function ϕ similarly goes to ϕ^- or ϕ^+ with $\partial \tilde{\phi} / \partial \tilde{u} = \partial \phi / \partial \tilde{u} \to 0$. Hence, a side-effect of the remaining penalty term $2\varepsilon\tilde{u}$ in (42) is that it helps to avoid singularity of $\partial H / \partial \tilde{u}$ in the case of saturation.

DOI: 10.1002/oca

with the final condition

$$\lambda^{\mathsf{T}}(T) = \frac{\partial \tilde{\varphi}}{\partial \tilde{x}} + \nu^{\mathsf{T}} \frac{\partial \tilde{\chi}}{\partial \tilde{x}} \tag{43d}$$

and the additional multipliers $\nu \in \mathbb{R}^q$. The differential equations and boundary conditions (41), (43) together with the algebraic equation (42) for $\tilde{u}$ defines a two-point boundary value problem (BVP), which (in general) must be solved numerically to obtain the input u^ε, the states $(\tilde{x}^\varepsilon, \lambda^\varepsilon)$, and the multipliers ν^ε. The normal form coordinates y^ε and finally the original input u^ε and state x^ε follow from the relations (15), (20), and (8):

$$\begin{aligned} y^\varepsilon &= h(\xi^\varepsilon), \quad u^\varepsilon = h_u(\xi^\varepsilon, z^\varepsilon, \tilde{u}^\varepsilon) \\ x^\varepsilon &= \theta^{-1}(y^\varepsilon, z^\varepsilon) \end{aligned} \tag{44}$$

In order to approach the optimal solution (u^*, y^*, z^*) and optimal cost $\bar{J}(u^*) = J^*$, the BVP (41)–(43) has to be solved successively for a sequence $\{\varepsilon^k\}$ of decreasing penalty parameters $\varepsilon^{k+1} < \varepsilon^k$ and using the previous solutions for ε^k within a continuation scheme. If the optimal solution (u^*, y^*, z^*) touches one of the constraints (9f), the reduction of $\varepsilon_k \to 0$ is not possible, since, in this case, the internal variables ξ_1^k and $\tilde{u}^k$ of $\mathrm{OCP}_\xi^\varepsilon$ would become unbounded in the limit. In practice, the sequence $\{\varepsilon^k\}$ is stopped at a certain step k when the corresponding solution is sufficiently close to the optimal one.

5.2. Numerical solution with collocation

An efficient method to numerically solve two-point BVPs is collocation, see e.g. [18]. A convenient collocation code is the solver bvp4c [19] implemented under MATLAB, which can be used to solve nonlinear two-point BVPs. However, to be applicable to OCPs, we extended the bvp4c-code to additionally account for algebraic equations like (42) as they arise from the optimality conditions. This leads to the general BVP formulation of (index 1) differential–algebraic equations (DAE)

$$\dot{x}_d = f_d(x_d, x_a, t, p) \tag{45a}$$

$$0 = f_a(x_d, x_a, t, p) \tag{45b}$$

$$0 = f_{bc}(x_d(t_0), x_d(t_f), x_a(t_0), x_a(t_f), p) \tag{45c}$$

with the differential and algebraic equations (45a) and (45b) for the dynamic and algebraic states $x_d(t)$ and $x_a(t)$ on the time interval $t \in [t_0, t_f]$ and the boundary conditions (45c). Unknown parameters p can additionally be considered in the DAE formulation (45).

The general collocation method and its implementation in bvp4c has been left unchanged as it was designed to be applicable and numerically robust for a wide range of BVPs. The function bvp4c discretizes the differential equations (45a) along a time mesh $t_i \in [t_0, t_f], i = 1, \ldots, N$. In addition, the bvp4c-code has been extended to additionally evaluate the algebraic equation (45b) at the mesh points t_i. The resulting discretized equations (45a) and (45b) together with the boundary conditions (45c) results in a set of nonlinear algebraic equations for the variables $x_d(t_i)$ and $x_a(t_i), i = 1, \ldots, N$, which is solved with a Newton iteration scheme. In addition, bvp4c employs a mesh refinement strategy to adapt the time mesh $t_i \in [t_0, t_f], i = 1, \ldots, N$, and the number of grid points N in each Newton step based on the residual along the discretized ODEs (45a).

In order to use the collocation method for solving $\mathrm{OCP}_\xi^\varepsilon$, the BVP (41)–(43) has to be adapted to the DAE form (45). The ODEs (45a) are given by the system and adjoint equations in (41) and (43) for the dynamic state $x_d^{\mathsf{T}} = (\tilde{x}^{\mathsf{T}}, \lambda^{\mathsf{T}})$. The input $\tilde{u}$ denotes the algebraic variable $z = \tilde{u}$ with (42) corresponding to (45b). The boundary conditions for $\tilde{x}$ and λ in (41) and (43) are comprised in (45c). The multipliers ν in the final condition (43d) can be treated as unknown parameters $p = \nu$.

Remark 4

The normal form dynamics (28b)–(28d), which is comprised in the compactly written dynamics (41), can be written in a higher-order representation of ξ_1. The same structure is reflected in the adjoint system (43). By successively differentiating (42), the adjoint equations (43) can be expressed in terms of $\lambda_{\xi,r}$ and its derivatives [20]. These higher-order representations in ξ_1 and $\lambda_{\xi,r}$ result in fewer unknowns in the collocation scheme and lead to a higher accuracy of the numerical solution [18, 20].

DOI: 10.1002/oca

5.3. Example system

Consider the following modified version of the classical double integrator problem in [7]:

$$\text{minimize} \quad J(u)=\frac{1}{2}\int_0^1 u^2\,\mathrm{d}t \tag{46a}$$

$$\text{subject to} \quad \dot{y}_1=y_2, \quad \dot{y}_2=u \tag{46b}$$

$$y(0)=(0,1)^\mathsf{T}, \quad y(1)=(0,-1)^\mathsf{T} \tag{46c}$$

$$y_1\in[c^-,c^+], \quad u\in[u^-,u^+] \tag{46d}$$

The system (46b) is already written in the normal form (9b)–(9c) with the states $y=(y_1,y_2)^\mathsf{T}$ and no internal dynamics (9d). The state constraint in (46d) has relative degree $r=2$ and, together with the input constraint $u\in[u^-,u^+]$, directly correspond to (9f).

Following Section 3, the constraints (46d) are represented by two saturation functions ψ and ϕ, which yields the relations (15) and (20)

$$\begin{aligned} y_1&=\psi(\xi_1,c^\pm), \quad y_2=\psi'\xi_2\\ u&=h_u(\xi,\tilde{u})=\psi''\xi_2^2+\psi'\phi(\tilde{u},\phi^\pm(\xi)) \end{aligned} \tag{47a}$$

with the saturation limits (18b)

$$\phi^\pm(\xi)=\frac{u^\pm-\psi''\xi_2^2}{\psi'(\xi_1,c^\pm)} \tag{47b}$$

In the new coordinates $\xi=(\xi_1,\xi_2)^\mathsf{T}$ with input $\tilde{u}$, the dynamics (46b) and boundary conditions (46c) are replaced by

$$\begin{aligned} \dot{\xi}_1&=\xi_2, \quad \dot{\xi}_2=\tilde{\phi}(\xi,\tilde{u})\\ \xi(0)&=(0,1)^\mathsf{T}, \quad \xi(1)=(0,-1)^\mathsf{T} \end{aligned} \tag{48}$$

with $\tilde{\phi}=\phi(\tilde{u},\phi^\pm(\xi))$. The boundary conditions in (48) correspond to (46c), since symmetry is assumed for the state constraints $c^+=-c^-$, which yields $0=\psi(0,c^\pm)$ and $\psi'(0,c^\pm)=1$ for the saturations functions (A1) and (A2) in Appendix A.1.

As described in Section 3.3, the cost (46a) is transformed and penalized in the new coordinates

$$P(\tilde{u},\varepsilon)=\int_0^T \frac{1}{2}h_u(\xi,\tilde{u})^2+\varepsilon(\xi_1^2+\tilde{u}^2)\,\mathrm{d}t$$

to account for the unboundedness of ξ_1 or $\tilde{u}$ if one of the constraints (46d) is touched. Finally, the optimality conditions (42), (43) read as

$$\frac{\partial H}{\partial\tilde{u}}=h_u(\xi,\tilde{u})\psi'\frac{\partial\tilde{\phi}}{\partial\tilde{u}}+2\varepsilon\tilde{u}+\lambda_{\xi,2}\frac{\partial\tilde{\phi}}{\partial u}=0 \tag{49a}$$

$$\dot{\lambda}_{\xi,1}=-h_u(\xi,\tilde{u})\frac{\partial h_u}{\partial\xi_1}-2\varepsilon\xi_1-\lambda_{\xi,2}\frac{\partial\tilde{\phi}}{\partial\xi_1} \tag{49b}$$

$$\dot{\lambda}_{\xi,2}=-h_u(\xi,\tilde{u})\frac{\partial h_u}{\partial\xi_2}-\lambda_{\xi,2}\frac{\partial\tilde{\phi}}{\partial\xi_2} \tag{49c}$$

The final conditions $\lambda_{\xi,i}(T)=v_i, i=1,2$ following from (43d) can be omitted, since the multipliers v_1, v_2 do not appear elsewhere.

The transformations (47a) and the optimality conditions (49) are analytically calculated with the software package MATHEMATICA using the explicit formulas (A1)–(A2) in Appendix A.1 for ψ and ϕ. The equations of the two-point BVP (46b) and (46c), (48) and (49) are adapted to the form (45) and are provided as MATLAB Cmex-functions to the collocation solver, as described in Section 5.2.

The initial guess for the state $\xi(t_i)$ is a linear interpolation between the boundary conditions (46c) on a uniform time mesh $t_i\in[0,1], i=1,\ldots,N$, with $N=30$ mesh points. The initial guess for $\lambda(t_i)$ and $\tilde{u}(t_i)$ is zero. The BVP is successively solved for the penalty terms $\varepsilon\in\{10^0,10^{-1},\ldots,10^{-12}\}$ using the previous solution as initial guess for the next run. For the simulation studies, we considered the constraints $c^\pm=\pm0.15$ and $u^\pm=\pm3.5$.

Figure 3 shows the optimal solutions $(\tilde{u}^\varepsilon,\xi^\varepsilon)$ and the corresponding original variables $(u^\varepsilon,y^\varepsilon)$ following from (47a) for several penalty parameters ε. Remarkable is that the first solution for $\varepsilon=10^0$ already closely approaches the constraints and has been readily obtained by starting from a trivial initial guess. This illustrates the particular advantage of the approach that the constraints (46d) cannot be violated during the numerical solution due to their systematic incorporation in (47) and (48).

Clearly visible is the non-violation of the constraints (46d) and the convergence to the optimal solution

DOI: 10.1002/oca

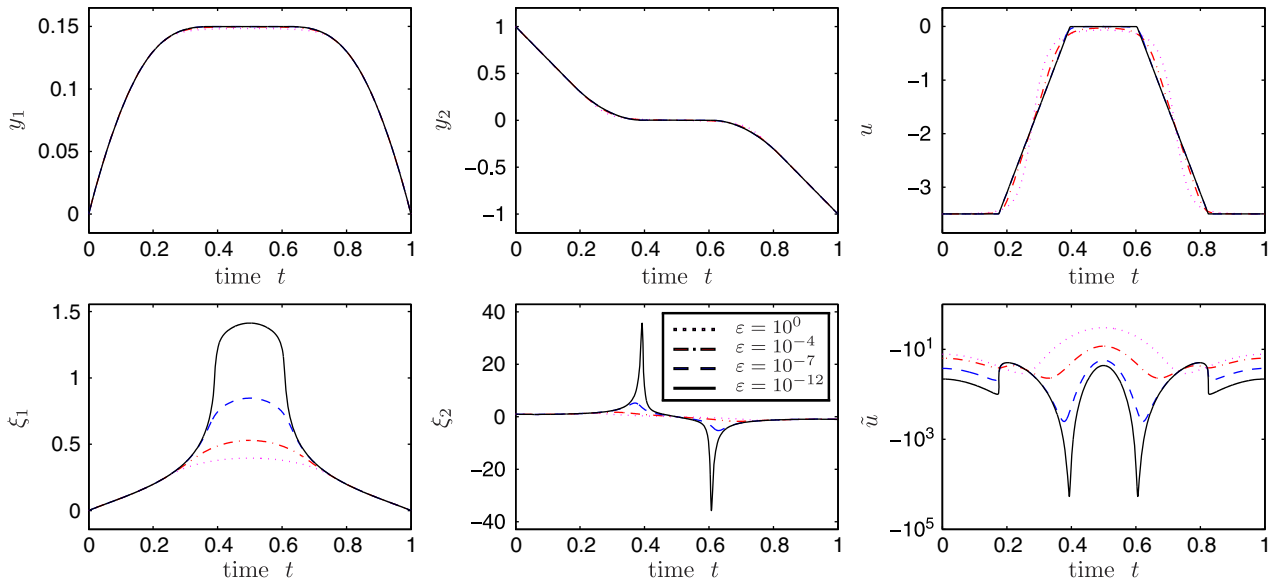

Figure 3. Optimal trajectories for the example (46) with decreasing penalty parameter ε.

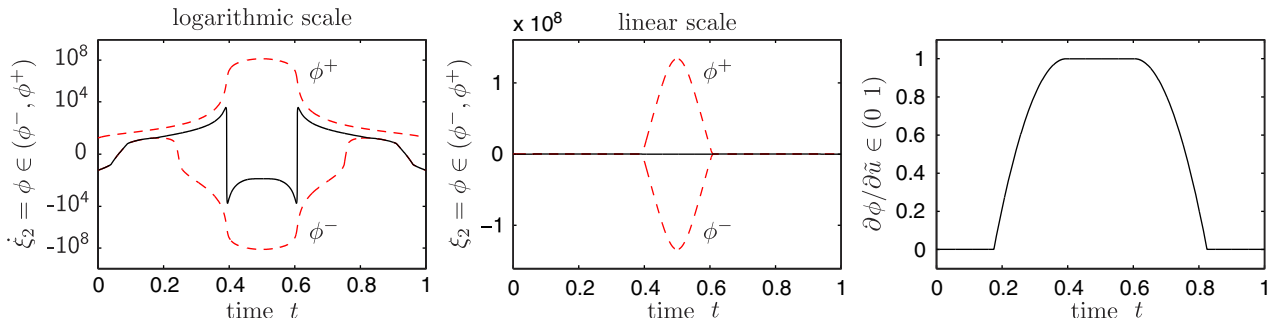

Figure 4. Behavior of second saturation function $\dot{\xi}_2 = \phi(\tilde{u}, \phi^{\pm}(\xi))$ for $\varepsilon = 10^{-12}$.

(u^*, y^*) for decreasing ε.[‖] As discussed in Section 3.2, the internal states (ξ_1, ξ_2) and the new input $\tilde{u}$ (plotted logarithmically in Figure 3) tend to become unbounded when the constraints (46d) are approached and ε is successively reduced. For the final run with $\varepsilon = 10^{-12}$, the minimal distance of y_1^{ε} and u^{ε} to the constraints $c^+ = 0.15$ and $u^- = -3.5$ is of order 10^{-9} and 10^{-11}, respectively. The distance to the optimal cost J^* is of order 10^{-7}.

Figure 4 additionally shows the trajectories of $\dot{\xi}_2 = \phi(\tilde{u}, \phi^{\pm}(\xi))$ to illustrate the behavior of ϕ. At the beginning and end of the time interval $[0, 1]$, ϕ almost reaches (with negligible distance) the lower limit $\phi^-(\xi)$ corresponding to the input constraint in Figure 3. In the middle of the interval, the bounds $\phi^{\pm}(\xi)$ behave in a symmetric manner and significantly increase in magnitude due to the gradient $\psi'(\xi_1, c^{\pm})$ appearing in the

[‖] Note that (u^*, y^*) can be analytically computed, which is omitted here due to the lack of space. The optimal value of the cost is $J^* = u^+(1 - \sqrt{2c^+u^+ - 1}/\sqrt{3}) \approx 3.04815$ for the constraint values $c^{\pm} = \pm 0.15$ and $u^{\pm} = \pm 3.5$.

DOI: 10.1002/oca

denominator in (47b). Hence, ϕ 'opens' in the neighborhood of the state constraint, which leads to $\dot{\xi}_2 = \phi(\tilde{u}, \phi^{\pm}(\xi)) \approx \tilde{u}$ due to the normalization $(\partial\phi/\partial\tilde{u}) \in (0, 1)$ of the saturation function (A1) in Appendix A.1. This effect is shown in the two right plots of Figure 4 and is explained in more details in Appendix A.2 (also see Remark 1).

6. EXTENSION TO THE MULTIPLE INPUT CASE

This section extends the results from the previous sections to the multiple input case in a compact manner. The main extension concerns the incorporation of multiple input constraints, which—as will appear—is more convoluted than in the single input case.

6.1. *Optimal control problem* OCP_x

Considered is the following nonlinear control-affine multiple input system:

$$\dot{x} = f(x) + \sum_{i=1}^{m} g_i(x) u_i \tag{50}$$

with the state $x \in \mathbb{R}^n$, the input vector $u = (u_1, \ldots, u_m)^{\mathsf{T}} \in \mathbb{R}^m$, and the sufficiently smooth vector fields $f, g_i : \mathbb{R}^n \to \mathbb{R}^n, i = 1, \ldots, m$. The boundary conditions (2) and cost function (3) with $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ are in essence the same as in the single input case.

In consistency with the constraints (4), the following state and input constraints are assumed:

$$c_i(x) \in [c_i^-, c_i^+], \quad u_i \in [u_i^-(x), u_i^+(x)], \quad i = 1, \ldots, m \tag{51}$$

The vector relative degree $\{r_1, \ldots, r_m\}$ of the m functions $c_i(x)$ at a point x^0 is defined by [13]

$$L_{g_j} L_f^k c_i(x) = 0 \tag{52a}$$

for all $1 \leqslant j \leqslant m, k < r_i - 1, 1 \leqslant i \leqslant m$, and for all x in a neighborhood of x^0. Moreover, the $m \times m$ matrix

$$A(x) = \begin{pmatrix} L_{g_1} L_f^{r_1 - 1} c_1(x) & \cdots & L_{g_m} L_f^{r_1 - 1} c_m(x) \\ \vdots & & \vdots \\ L_{g_1} L_f^{r_m - 1} c_1(x) & \cdots & L_{g_m} L_f^{r_m - 1} c_m(x) \end{pmatrix} \tag{52b}$$

has to be non-singular at $x = x^0$. In the following, we assume that the m state constraints in (51) have a well-defined relative degree $\{r_1, \ldots, r_m\}$, which means that the conditions (52a) as well as the non-singularity of the decoupling matrix (52b) are satisfied in a sufficiently large neighborhood of x^0. The OCP OCP_x is summarized as follows.

Problem OCP_x (multiple input case):

$$\begin{aligned} \text{minimize} \quad & J(u) = \varphi(x(T)) + \int_0^T L(x, u, t)\, \mathrm{d}t \\ \text{subject to} \quad & \dot{x} = f(x) + \sum_{i=1}^{m} g_i(x) u_i \\ & x(0) = x_0, \quad \chi(x(T)) = 0 \\ & c_i(x) \in [c_i^-, c_i^+] \\ & u_i \in [u_i^-(x), u_i^+(x)], \quad i = 1, \ldots, m \end{aligned}$$

Note that the consideration of m state constraints and m input constraints is the most general case considered here. If an input u_i is unconstrained, the respective limits can be set to $u_i^{\pm} \to \pm\infty$. If the number of state constraints is less than the number m of inputs, the remaining functions $c_i(x)$ have to be chosen to achieve a well-defined relative degree $\{r_1, \ldots, r_m\}$. This case is addressed by the example application in Section 6.5.

6.2. *Normal form representation*

Owing to the well-defined relative degree $\{r_1, \ldots, r_m\}$ of the constraint functions $c_i(x), i = 1, \ldots, m$, there exists a change of coordinates [13]

$$\begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} \theta_y(x) \\ \theta_z(x) \end{pmatrix} = \theta(x) \tag{53a}$$

with $y^{\mathsf{T}} = (y_1^{\mathsf{T}}, \ldots, y_m^{\mathsf{T}})$ and $y_i = (y_{i,1}, \ldots, y_{i,r_i})^{\mathsf{T}}$ defined by

$$\begin{aligned} & y_{i,1} = c_i(x) = \theta_{i,1}(x), \quad y_{i,j} = L_f^j c_i(x) = \theta_{i,j}(x) \\ & \quad j = 2, \ldots, r_i, \quad i = 1, \ldots, m \end{aligned} \tag{53b}$$

The single functions $\theta_{i,j}$ are comprised in $\theta = (\theta_{1,1}, \ldots, \theta_{m,r_m})^{\mathsf{T}}$. The additional coordinates $z = \theta_z(x) \in \mathbb{R}^{n-r}$ with $r = \sum_{i=1}^{m} r_i$ are necessary to complete the

DOI: 10.1002/oca

transformation (53) if $r<n$. In these coordinates, the original OCP_x can be stated under the following form:

Problem OCP_y (multiple input case).

$$\text{minimize} \quad \bar{J}(u)=\bar{\varphi}(y(T),z(T)) + \int_0^T \bar{L}(y,z,u,t)\,\mathrm{d}t \tag{54a}$$

$$\text{subject to} \quad \dot{y}_{i,j}=y_{i,j+1}, \quad j=1,\ldots,r_i-1 \tag{54b}$$

$$\dot{y}_{i,r_i}=a_{i,0}(y,z) + \sum_{j=1}^{m} a_{i,j}(y,z)u_j, \quad i=1,\ldots,m \tag{54c}$$

$$\dot{z}=b_0(y,z)+B(y,z)u \tag{54d}$$

$$y(0)=\theta_y(x_0), \quad \bar{\chi}(y(T),z(T))=0 \tag{54e}$$

$$y_{i,1}\in[c_i^-,c_i^+], \quad u_i\in[\bar{u}_i^-(y,z),\bar{u}_i^+(y,z)] \quad i=1,\ldots,m \tag{54f}$$

where $a_{i,0}=L_f^{r_i}c_i(x)\circ\theta^{-1}$, $a_{i,j}=L_{g_j}L_f^{r_i-1}c_i(x)\circ\theta^{-1}$, and $u=(u_1,\ldots,u_m)^\mathsf{T}$. The functions of the cost $\bar{\varphi}=\varphi\circ\theta^{-1}$, $\bar{L}=L\circ\theta^{-1}$, and constraints $\bar{u}_i^{\pm}=u_i^{\pm}\circ\theta^{-1}$ correspond to OCP_x.

As in [13], the normal form dynamics of OCP_y comprises the input–output dynamics (54b)–(54c) and the internal dynamics (54d) with the matrix function $B:\mathbb{R}^r\times\mathbb{R}^{n-r}\to\mathbb{R}^{n-r\times m}$. The equations for $\dot{y}_{i,r_i}$ can be written in vector notation

$$\dot{y}_r=a_0(y,z)+\bar{A}(y,z)u \tag{55a}$$

with $\dot{y}_r=(\dot{y}_{1,r_1},\ldots,\dot{y}_{m,r_m})^\mathsf{T}$ and $a_0=(a_{0,1},\ldots,a_{0,m})^\mathsf{T}$ to determine the input vector u:

$$u=\bar{A}^{-1}(y,z)(\dot{y}_r-a_0(y,z)) \tag{55b}$$

The inverse of the decoupling matrix $\bar{A}(y,z)=\{a_{i,j}(y,z)\}=A(x)\circ\theta^{-1}$ is well-defined due to the full rank condition (52b).

6.3. Using saturation functions to represent the constraints

In a straightforward extension of the single input case in Section 3, the state constraints in (54f) can be represented by m saturation functions $y_{i,1}=\psi_i(\xi_{i,1},c_i^{\pm})$ and using successive differentiation of $y_{i,1}$. This defines the mappings

$$y_{i,1}=h_{i,1}(\xi_{i,1})=\psi_i(\xi_{i,1},c_i^{\pm}), \quad i=1,\ldots,m \tag{56a}$$

$$\begin{aligned} y_{i,j}&=h_{i,j}(\xi_{i,1},\ldots,\xi_{i,j})\\ &=\gamma_{i,j}(\xi_{i,1},\ldots,\xi_{i,j-1})+\psi_i'\xi_{i,j}, \quad j=2,\ldots,r_i \end{aligned} \tag{56b}$$

comprised in

$$y=h(\xi)=(h_{1,1}(\xi_{1,1}),\ldots,h_{m,r_m}(\xi_m))^\mathsf{T} \tag{57}$$

The vector notation $\xi_i=(\xi_{i,1},\ldots,\xi_{i,r_i})^\mathsf{T}$ is used when it is beneficial. The nonlinear terms $\gamma_{i,j}$ are determined with respect to the previous equation for $y_{i,j-1}$, i.e.

$$\gamma_{i,2}(\xi_{i,1})=0$$

$$\gamma_{i,j}(\xi_{i,1},\ldots,\xi_{i,j-1})=\sum_{k=1}^{j-2}\frac{\partial h_{i,j-1}}{\partial\xi_{i,k}}\xi_{i,k+1}$$

$$j=3,\ldots,r_i, \quad i=1,\ldots,m$$

Similar to the single input case, the successive differentiations of $y_{i,1}$ along the multiple cascades lead to a new set of coordinates $\xi^\mathsf{T}=(\xi_1^\mathsf{T},\ldots,\xi_m^\mathsf{T})\in\mathbb{R}^r$ that replaces y. The inverse mapping is denoted by $y=h^{-1}(\xi)$ and is addressed in more detail in Section 3.2. The final differentiations to reach y_{i,r_i} yield

$$\dot{y}_{i,r_i}=\gamma_{i,r_i+1}(\xi_i)+\psi_i'\dot{\xi}_{i,r_i}, \quad i=1,\ldots,m \tag{58}$$

In contrast to the straightforward derivation of (56)–(57), the incorporation of the input constraints $u_i\in[\bar{u}_i(y,z),\bar{u}_i(y,z)]$ via the highest derivatives $\dot{y}_{i,r_i}$ is more complicated than in the single input case due to the influence of the decoupling matrix $\bar{A}(y,z)$ in (55a). Only in the exceptional case when $\bar{A}(y,z)$ is a diagonal matrix, i.e. $\dot{y}_{i,r_i}=a_{0,i}(y,z)+a_{i,i}(y,z)u_i$, the constraints on u_i can directly be mapped to $\dot{y}_{i,r_i}$ as in (19), in order to use further saturation functions for $\dot{\xi}_{i,r_i}=\phi_i(\tilde{u}_i,\phi_i^{\pm})$. In the general case, the structure of the decoupling matrix $\bar{A}(y,z)$ has to be taken into

DOI: 10.1002/oca

account by using $\phi_i(\tilde{u}_i, \phi_i^{\pm}), i=1,\ldots,m$ in a linear combination

$$\begin{pmatrix} \dot{\xi}_{1,r_1} \\ \vdots \\ \dot{\xi}_{m,r_m} \end{pmatrix} = D \begin{pmatrix} \phi_1(\tilde{u}_1, \phi_1^{\pm}) \\ \vdots \\ \phi_m(\tilde{u}_m, \phi_m^{\pm}) \end{pmatrix} \tag{59}$$

where the elements of the $m \times m$ matrix $D=\{d_{i,j}\}$ and the saturation limits $\phi_i^{\pm}$ still have to be determined. Combining (55a), (58), and (59) in vector notation with $\gamma_{r+1}=(\gamma_{1,r_1+1},\ldots,\gamma_{m,r_m+1})^{\mathsf{T}}$ (and omitting arguments where it is beneficial) leads to

$$\begin{aligned} &\begin{pmatrix} \psi_1' & & 0 \\ & \ddots & \\ 0 & & \psi_m' \end{pmatrix} D \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_m \end{pmatrix} \\ &\quad = \tilde{a}_0(\xi,z) + \tilde{A}(\xi,z)u - \gamma_{r+1}(\xi) \\ &\quad = |\tilde{A}|^{-1}\tilde{A}(|\tilde{A}|u + \delta(\xi,z)) \end{aligned} \tag{60a}$$

where $\tilde{a}_0 = a_0 \circ h$, $\tilde{A} = A \circ h$, and

$$\delta(\xi,z) = |\tilde{A}|\tilde{A}^{-1}(\tilde{a}_0(\xi,z) - \gamma_{r+1}(\xi)) \tag{60b}$$

The non-singularity of the decoupling matrix $\tilde{A}(\xi,z)$ follows from the well-defined relative degree, which ensures that the inverse $\tilde{A}^{-1}$ with the determinant $|\tilde{A}| \neq 0$ exists in a sufficiently large neighborhood of $\theta(x^0)$.** The vector equation (60a) can be reformulated and expanded with respect to the partial derivatives $(\psi_1', \ldots, \psi_m')$:

$$D \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_m \end{pmatrix} = \underbrace{|\tilde{A}|^{-1} \begin{pmatrix} \prod_{k\in\mathscr{K}_1} \psi_k' & & 0 \\ & \ddots & \\ 0 & & \prod_{k\in\mathscr{K}_m} \psi_k' \end{pmatrix} \tilde{A}}_{m\times m \text{ matrix}}$$

$$\times \underbrace{\begin{pmatrix} |\tilde{A}|u_1 + \delta_1 \\ \vdots \\ |\tilde{A}|u_m + \delta_m \end{pmatrix} \frac{1}{\prod_{k=1}^{m} \psi_k'}}_{m \text{ vector}} \tag{61}$$

The sets $\mathscr{K}_i$ are defined by $\mathscr{K}_i = \{k=1,\ldots,m : k \neq i\}$. Owing to the reformulation, the first part of the right-hand side of (61) is a $m \times m$ matrix independent of u, whereas the second part is a vector of dimension m that depends on u. By comparison with the left-hand side of (61), the elements of the matrix $D=\{d_{i,j}\}$ are set to

$$\begin{aligned} d_{i,j} := d_{i,j}(\xi,z) &= \frac{\tilde{a}_{i,j}(\xi,z)}{|\tilde{A}(\xi,z)|} \prod_{k\in\mathscr{K}_i} \psi_k'(\xi_{k,1}, c_k^{\pm}) \\ &\quad i=1,\ldots,m \end{aligned} \tag{62}$$

and thus depend on the states ξ and z. The functions $\tilde{a}_{i,j}(\xi,z)$ belong to the decoupling matrix $\tilde{A}(\xi,z) = \{\tilde{a}_{i,j}(\xi,z)\}$. Further comparison of the saturation functions $(\phi_1,\ldots,\phi_m)$ with the vector on the right-hand side of (61) shows that each ϕ_i is related to the ith input u_i via the expression $(|\tilde{A}|u_i + \delta_i)/\prod_{k=1}^{m}\psi_k'$. Hence, in order to satisfy the input constraints in (54f), the limits of the saturation functions $\phi_i(\tilde{u}, \phi_i^{\pm}), i=1,\ldots,m$ have to be chosen to

$$\begin{aligned} \phi_i^{\pm} &:= \phi_i^{\pm}(\xi,z) \\ &= \begin{cases} \dfrac{|\tilde{A}(\xi,z)|\tilde{u}_i^{\pm}(\xi,z) + \delta_i(\xi,z)}{\prod_{k=1}^{m}\psi_k'(\xi_{k,1}, c_k^{\pm})} & \text{if } |\tilde{A}(\xi,z)| > 0 \\ \dfrac{|\tilde{A}(\xi,z)|\tilde{u}_i^{\mp}(\xi,z) + \delta_i(\xi,z)}{\prod_{k=1}^{m}\psi_k'(\xi_{k,1}, c_k^{\pm})} & \text{if } |\tilde{A}(\xi,z)| < 0 \end{cases} \end{aligned} \tag{63}$$

depending on the sign of the determinant $|\tilde{A}(\xi,z)| \neq 0$.

The highest derivatives $\dot{y}_r = (\dot{y}_{1,r_1},\ldots,\dot{y}_{m,r_m})^{\mathsf{T}}$ in (58) can now be expressed as

$$\begin{aligned} \dot{y}_{i,r_i} &= \gamma_{i,r_i+1}(\xi_i) + \psi_i' \sum_{j=1}^{m} d_{i,j}(\xi,z)\tilde{\phi}_j(\xi,z,\tilde{u}_j) \\ &= h_{i,r_i+1}(\xi,z,\tilde{u}), \quad i=1,\ldots,m \end{aligned} \tag{64a}$$

**The normalization in (60) with respect to the determinant $|\tilde{A}(\xi,z)|$ is used to achieve consistency with the single input case, also see Remark 5.

DOI: 10.1002/oca

with $\tilde{\phi}_i = \phi_i(\tilde{u}_i, \phi_i^{\pm}(\xi, z))$ and the new input vector $\tilde{u} = (\tilde{u}_1, \ldots, \tilde{u}_m)^{\mathsf{T}}$. Summarizing these relations in $\dot{y}_r = h_{r+1}(\xi, z, \tilde{u})$, we can rewrite the input vector (55b) in the new coordinates ξ and inputs $\tilde{u}$ as

$$\begin{aligned} u &= h_u(\xi, z, \tilde{u}) \\ &= \tilde{A}^{-1}(\xi, z)(h_{r+1}(\xi, z, \tilde{u}) - \tilde{a}_0(\xi, z) \end{aligned} \tag{64b}$$

Remark 5

The expansion with respect to the partial derivatives ψ_k' in (61) has been undertaken in order to collect all ψ_k'–terms in the denominator of the saturation limits (63). The benefit of this formulation is mentioned in Remark 1 for the single input case. Moreover, it can easily be verified that the expressions (59) with (62)–(63) exactly reduce to (18) for $m = 1$.

6.4. New penalized OCP $\mathrm{OCP}_{\xi}^{\varepsilon}$

As stated in detail in Section 3.2, the coordinates $\xi_{i,1}$ and new inputs $\tilde{u}_i$ as the arguments of the saturation functions (56a) and (59) become unbounded if one of the corresponding state or input constraints (54f) is touched. This problem is addressed (as in the single input case) by penalizing $\xi_{i,1}$ and the new inputs $\tilde{u}_i, i = 1, \ldots, m$, in the cost function, which leads to the new penalized $\mathrm{OCP}_{\xi}^{\varepsilon}$.

Problem $\mathrm{OCP}_{\xi}^{\varepsilon}$ (multiple input case):

$$\text{minimize} \quad P(\tilde{u}, \varepsilon) = \tilde{J}(\tilde{u}) + \varepsilon \sum_{i=1}^{m} \int_0^T (\xi_{i,1}^2 + \tilde{u}_i^2)\, \mathrm{d}t \tag{65a}$$

$$\text{subject to} \quad \dot{\xi}_{i,j} = \xi_{i,j+1}, \quad j = 1, \ldots, r_i - 1$$

$$\dot{\xi}_{i,r_i} = \sum_{j=1}^{m} d_{i,j}(\xi, z)\tilde{\phi}_j(\xi, z, \tilde{u}), \quad i = 1, \ldots, m \tag{65b}$$

$$\begin{aligned} \dot{z} &= \tilde{b}(\xi, z, \tilde{u}) \\ &= \tilde{b}_0(\xi, z) + \tilde{B}(\xi, z) h_u(\xi, z, \tilde{u}) \end{aligned} \tag{65c}$$

$$\xi(0) = h^{-1}(\theta(x_0)), \quad z(0) = \theta_z(x_0)$$

$$\tilde{\chi}(\xi(T), z(T)) = 0 \tag{65d}$$

where $\tilde{b}_0 = b_0 \circ h$, $\tilde{B} = B \circ h$, and $\tilde{\chi} = \bar{\chi} \circ h$ follow from OCP_y. The constraints (54f) are incorporated in the dynamics by the asymptotic saturation functions $\psi_i(\xi_{i,1}, c_i^{\pm})$ and the linearly combined $\tilde{\phi}_i = \phi_i(\tilde{u}_i, \phi_i^{\pm}(\xi, z))$ with (62), (63). Their successive derivatives uniquely define $y = h(\xi)$, $\dot{y}_r = h_{r+1}(\xi, z, \tilde{u})$, and $u = h_u(\xi, z, \tilde{u})$ stated in (57) and (64).

Similar to the single input case, the penalty parameter ε has to be successively reduced during the numerical solution of $\mathrm{OCP}_{\xi}^{\varepsilon}$ in order to approach the optimal solution (u^*, y^*, z^*) of OCP_y via the mappings (57) and (64). For details, we refer back to Sections 4 and 5 of the single input case.

6.5. Example: Ducted fan

The incorporation of the constraints in the multiple input case is illustrated for the planar ducted fan [21], as shown in Figure 5. The system consists of a rigid body described by the position (x_1, x_2) in the center of gravity and the angle α to the vertical. The thrust of the ducted fan is given by the body-fixed forces u_1 and u_2, which can be adjusted by moving the flaps at the end of the duct. We consider the following constrained OCP with constraints on both inputs (u_1, u_2) and angle α:

$$\text{minimize} \quad J(u) = \int_0^T \mu + 2u_1^2 + u_2^2\, \mathrm{d}t \tag{66a}$$

$$\text{subject to} \quad m\ddot{x}_1 = u_1 \cos\alpha - u_2 \sin\alpha \tag{66b}$$

$$m\ddot{x}_2 = -mg + u_1 \sin\alpha + u_2 \cos\alpha \tag{66c}$$

$$J\ddot{\alpha} = r u_1 \tag{66d}$$

$$x(0) = (0, 0, 0, 0, 0, 0)^{\mathsf{T}}$$

$$x(T) = (0, 0, 1, 0, 0, 0)^{\mathsf{T}} \tag{66e}$$

$$\alpha \in [\alpha^-, \alpha^+]$$

$$u_1 \in [u_1^-, u_1^+], \quad u_2 \in [u_2^-, u_2^+] \tag{66f}$$

The simplified model equations (66b)–(66d) are taken from [21] as well as the model parameters[††]

[††] The experimental setup of the ducted fan used in [21] is attached to a vertical stand with a counter weight, which leads to the reduced gravity force.

DOI: 10.1002/oca

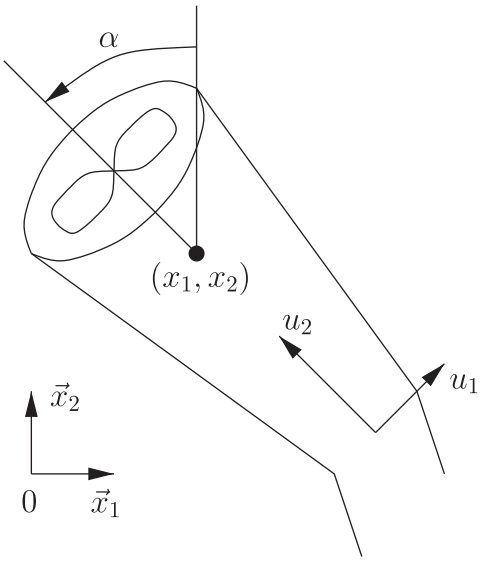

Figure 5. Ducted fan with position (x_1, x_2), angle α to the vertical, and thrusts u_1 and u_2.

$r = 0.2\,\mathrm{m}$, $J = 0.05\,\mathrm{kg\,m^2}$, $m = 2.2\,\mathrm{kg}$, $m\,g = 4\,\mathrm{N}$ and the input constraints $u_1 \in [-5, 5]\,\mathrm{N}$ and $u_2 \in [0, 17]\,\mathrm{N}$. An additional constraint is imposed on the angle $\alpha \in [-30, 30]^\circ$ to arbitrarily restrict the movement of the fan. The transition problem for the fan is to cover a horizontal distance of 1 m in the free end time T, which leads to the boundary conditions (66e) for the state vector $x = (x_1, \dot{x}_1, x_2, \dot{x}_2, \alpha, \dot{\alpha})^\mathsf{T}$. The cost (66a) can be interpreted as a trade-off between time and energy optimality with respect to the parameter μ, see again [21]. In the following, we choose $\mu = 1000$ to put a strong emphasis on the minimization of T. The following derivation of the penalized $\mathrm{OCP}_\xi^\varepsilon$ proceeds along the lines of the previous subsections. As will appear, the extra feature that T is a free parameter does not interfere.

Remark 6

The ducted fan belongs to the class of flat systems [22], i.e. there exists a so-called flat output $z = (z_1, z_2)$ with

$$z_1 = x_1 - \frac{J}{mr}\sin\alpha, \quad z_2 = x_2 + \frac{J}{mr}\cos\alpha \tag{67}$$

which allows to parameterize the states and inputs $u = (u_1, u_2)$

$$\begin{aligned}(x, y, \alpha) &= f_x(z_1, \ddot{z}_1, z_2, \ddot{z}_2)\\ u &= f_u(\ddot{z}_1, z_1^{(3)}, z_1^{(4)}, \ddot{z}_2, z_2^{(3)}, z_2^{(4)})\end{aligned} \tag{68}$$

in terms of z and its time derivatives. By planning an appropriate flat time trajectory $z(t)$, $t \in [0, T]$, the corresponding input trajectory $u(t)$ can be algebraically calculated, which steers the system (66b)–(66d) between the boundary conditions (66e) and satisfies the constraints (66f) for a large enough transition time T. Hence, the non-emptiness of the set S^0 of admissible controls, see (30), can explicitly be concluded.

Since only one state constraint $c_1(x) = \alpha$ is given, a second function $c_2(x)$ can be freely chosen in order to derive the normal form coordinates (53):

$$\begin{aligned}y_{1,1} &= \alpha, \quad y_{1,2} = \dot{\alpha}, \quad y_{2,1} = x_2\\ y_{2,2} &= \dot{x}_2, \quad z_1 = x_1, \quad z_2 = \dot{x}_1\end{aligned}$$

Note that x_2 and not x_1 is chosen as coordinate $y_{2,1}$ to achieve a well-defined relative degree $\{r_1, r_2\} = \{2, 2\}$ around the vertical position $\alpha = 0$. The normal form (54b)–(54d) directly follows from reordering the model equations (66b)–(66d):

$$\dot{y}_{1,1} = y_{1,2}, \quad \dot{y}_{1,2} = \frac{r}{J}u_1 \tag{69a}$$

$$\dot{y}_{2,1} = y_{2,2}, \quad \dot{y}_{2,2} = -g + \frac{\sin y_{1,1}}{m}u_1 + \frac{\cos y_{1,1}}{m}u_2 \tag{69b}$$

$$\dot{z}_1 = z_2, \quad \dot{z}_2 = \frac{\cos y_{1,1}}{m}u_1 - \frac{\sin y_{1,1}}{m}u_2 \tag{69c}$$

In addition, the inputs u_1 and u_2 can be obtained by solving (69a)–(69b):

$$u_1 = \frac{J}{r}\dot{y}_{1,2}, \quad u_2 = m\frac{g + \dot{y}_{2,2}}{\cos y_{1,1}} - \frac{J}{r}\dot{y}_{1,2}\tan y_{1,1} \tag{70}$$

In the normal form coordinates $y = (y_{1,1}, y_{1,2}, y_{2,1}, y_{2,2})^\mathsf{T}$ and $z = (z_1, z_2)^\mathsf{T}$, the state constraint $y_{1,1} = \alpha \in [\alpha^-, \alpha^+]$ can be incorporated by a saturation

DOI: 10.1002/oca

function $\psi_1(\xi_{1,1}, \alpha^{\pm})$, whereas the second coordinate $y_{2,1}=x_1$ is unconstrained. This leads to the relations (57)

$$\begin{aligned} y_{1,1} &= \psi_1(\xi_{1,1}, \alpha^{\pm}), \quad y_{1,2}=\psi_1'\xi_{1,2} \\ \dot{y}_{1,2} &= \psi_1''\xi_{1,2}^2+\psi_1'\dot{\xi}_{1,2}, \quad y_{2,1}=\xi_{2,1} \\ y_{2,2} &= \xi_{2,2}, \quad \dot{y}_{2,2}=\dot{\xi}_{2,2} \end{aligned} \tag{71}$$

between y and the new coordinates $\xi=(\xi_{1,1}, \xi_{1,2}, \xi_{2,1}, \xi_{2,2})^{\mathsf{T}}$. Note that the choice $y_{2,1}=\xi_{2,1}$ corresponds to a second saturation function $y_{2,1}=\psi_2(\xi_{2,1}, x_2^{\pm})$, if (arbitrary) constraints for $y_{2,1}=x_2\in[x_2^-, x_2^+]$ are set to $x_2^{\pm}\to\infty$. Then, the normalized saturation function (A1) in Appendix A.1 reduces to $\psi_2(\xi_{2,1}, x_2^{\pm})\to\xi_{2,1}$.

In order to incorporate the input constraints (66f), two new saturation functions $\phi_i(\tilde{u}_i, \phi_i^{\pm})$, $i=1,2$, are used to parameterize the highest derivatives $\dot{\xi}_{1,2}$ and $\dot{\xi}_{2,2}$ according to (59). With the decoupling matrix $\tilde{A}(\xi)$ following from (69a)–(69b) and $y_{1,1}=\psi_1(\xi_{1,1}, \alpha^{\pm})$,

$$\tilde{A}(\xi)=\begin{pmatrix} \dfrac{r}{J} & 0 \\ \dfrac{\sin\psi_1}{m} & \dfrac{\cos\psi_1}{m} \end{pmatrix}, \quad |\tilde{A}(\xi)|=\frac{r}{m\,J}\cos\psi_1 \tag{72}$$

the single elements $d_{i,j}$ of the 2×2 matrix $D=\{d_{i,j}\}$ are derived from (62):

$$D(\xi)=\begin{pmatrix} \dfrac{m}{\cos\psi_1} & 0 \\ \dfrac{J}{r}\psi_1'\tan\psi_1 & \dfrac{J}{r}\psi_1' \end{pmatrix} \tag{73}$$

In addition, the expression (60b) evaluates to

$$\begin{aligned} \delta(\xi) &= \begin{pmatrix} \dfrac{\cos\psi_1}{m} & 0 \\ -\dfrac{\sin\psi_1}{m} & \dfrac{r}{J} \end{pmatrix} \begin{pmatrix} -\psi_1''\xi_{1,2}^2 \\ -g \end{pmatrix} \\ &= \begin{pmatrix} -\dfrac{\cos\psi_1}{m}\psi_1''\xi_{1,2}^2 \\ \dfrac{\sin\psi_1}{m}\psi_1''\xi_{1,2}^2-\dfrac{r\,g}{J} \end{pmatrix} \end{aligned}$$

and is used to determine the saturation limits (63)

$$\begin{aligned} \phi_1^{\pm}(\xi) &= \frac{\cos\psi_1}{m\,\psi_1'}\left(\frac{r}{J}u_1^{\pm}+\psi_1''\xi_{1,2}^2\right) \\ \phi_2^{\pm}(\xi) &= \frac{1}{\psi_1'}\left(\frac{r\cos\psi_1}{Jm}u_2^{\pm}+\frac{\sin\psi_1}{m}\psi_1''\xi_{1,2}^2-\frac{rg}{J}\right) \end{aligned} \tag{74}$$

This finally leads to the unconstrained system (66b)–(66d) with the states (ξ, z) and the new inputs $(\tilde{u}_1, \tilde{u}_2)$:

$$\dot{\xi}_{1,1}=\xi_{1,2}, \quad \dot{\xi}_{1,2}=\frac{m}{\cos\psi_1}\tilde{\phi}_1(\xi, \tilde{u}_1) \tag{75a}$$

$$\begin{aligned} \dot{\xi}_{2,1} &= \xi_{2,2}, \quad \dot{\xi}_{2,2}=\frac{J}{r}\psi_1'\tan\psi_1\tilde{\phi}_1(\xi, \tilde{u}_1) \\ &\quad +\frac{J}{r}\psi_1'\tilde{\phi}_2(\xi, \tilde{u}_2) \end{aligned} \tag{75b}$$

$$\begin{aligned} \dot{z}_1 &= z_2, \quad \dot{z}_2=\frac{J\psi_1'}{r}(\tilde{\phi}_1-\tan\psi_1\tilde{\phi}_2) \\ &\quad +\frac{J\psi_1''\xi_{1,2}^2}{mr\cos\psi_1}-g\tan\psi_1 \end{aligned} \tag{75c}$$

where $\tilde{\phi}_i=\phi_i(\tilde{u}_i, \phi_i^{\pm}(\xi))$, $i=1,2$. The internal dynamics (75c) is obtained by inserting (70)–(71) and (75a)–(75b) into (69c). With the new system dynamics (75), the OCP (66) of the ducted fan can be transformed to $\mathrm{OCP}_{\xi}^{\varepsilon}$, whereby an additional penalty term $\varepsilon\int_0^T \xi_{1,1}^2+\tilde{u}_1^2+\tilde{u}_2^2\,\mathrm{d}t$ with parameter ε is added to the cost (66a) to avoid the unboundedness of the saturation function arguments if one of the constraints (66f) is touched.

The optimality conditions for $\mathrm{OCP}_{\xi}^{\varepsilon}$ are derived according to Section 5.1 and are omitted here due to the lack of space. Note that the final conditions (43d) for the adjoint state $\lambda\in\mathbb{R}^6$ can be omitted since the final conditions (66e) for the ducted fan encompass the whole state vector x. In addition to (42) and (43), the transversality condition $H(\tilde{x}, \lambda, \tilde{u})|_T=0$ forms a further final condition to account for the free end time T.

The new system (75) as well as the optimality conditions (42) and (43) are analytically calculated with

DOI: 10.1002/oca

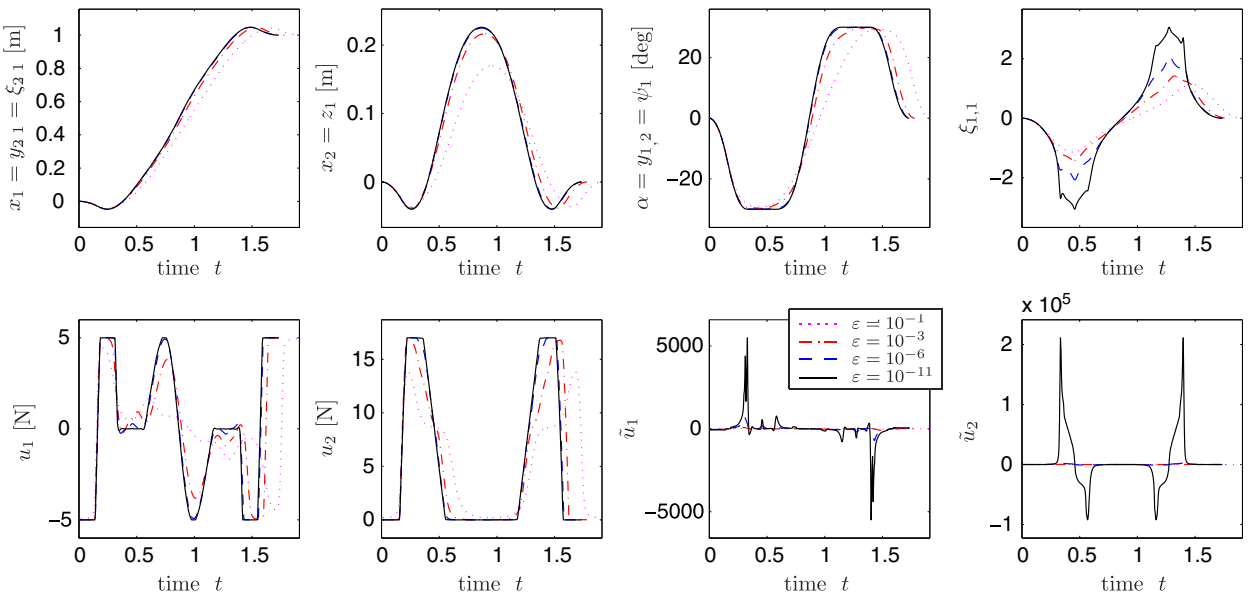

Figure 6. Optimal trajectories for the ducted fan with decreasing penalty parameter ε.

MATHEMATICA using the explicit formulas (A1)–(A2) in Appendix A.1 and are provided as Cmex-functions to the MATLAB collocation solver described in Section 5.2. A time transformation $\tau = \delta t$ with the normalized time coordinate $\tau \in [0, 1]$ and the scaling factor δ is used to address the free end time $T = \delta$. In the DAE representation (45) of the BVP solver, δ is used as free parameter p.

The initial guess for the states (ξ, z) is a linear interpolation between the respective boundary conditions on a uniform mesh with 200 points for the new time coordinate $\tau \in [0, 1]$. The initial guess for λ and $(\tilde{u}_1, \tilde{u}_2)$ is simply zero, whereas the free parameter $p = \delta$ is initialized with $p = 1$. The BVP is successively solved for the penalty terms $\varepsilon \in \{10^0, 10^{-1}, \ldots, 10^{-11}\}$ using the previous solution as initial guess for the next run. The mesh refinement of the BVP solver is turned off during the successive solutions and the single trajectories are computed on the fixed uniform mesh with 200 points. The reason for using a fixed mesh is that the mesh refinement leads to an increase of mesh points during the successive solutions, while the complex shape of the trajectories strongly changes.

Figure 6 shows the simulation results for several penalty parameters ε. The trajectories for the final run with $\varepsilon = 10^{-11}$ show an aggressive behavior, where all the constraints (66f) are clearly exploited. This aggressive maneuver of the ducted fan corresponds to the strong emphasis on time optimality in the cost (66a) with $\mu = 1000$.

7. CONCLUSIONS

The presented approach describes a systematic way to transform a constrained optimal control problem OCP_x into a new unconstrained one. For a given nonlinear system with m inputs, the treated class of constraints comprises up to m input constraints and m state constraints with well-defined relative degree. Starting from an equivalent OCP_y in normal form coordinates, a new system representation is derived by means of saturation functions and successive differentiation along the normal form cascade. This system in new coordinates is used to define a new unconstrained $\text{OCP}_\xi^\varepsilon$ with an

DOI: 10.1002/oca

additional penalty term in order to avoid unboundedness of the saturation function arguments, which occurs if the original constraints are touched. After deriving the optimality conditions for $\text{OCP}_\xi^\varepsilon$, the resulting BVP can be solved numerically (e.g. with the collocation method as used in this paper), whereby the penalty parameter ε has to be successively reduced within a continuation scheme. The single analytical steps in the derivation of $\text{OCP}_\xi^\varepsilon$ can be automated by using computer algebra systems such as MATHEMATICA or MAPLE.

The systematic incorporation of the constraints shows the philosophy behind the approach, which stresses the importance of analytical preprocessing to derive a truly unconstrained OCP. Compared, for instance, with interior penalty methods, where barrier functions are added to the cost to account for constraints, the saturation function approach directly includes the constraints in the new system dynamics. A particular benefit of this procedure is that the constraints cannot be violated in the new coordinates, which is of advantage for finding an initial numerical solution or to successively reduce the magnitude of the constraints, e.g. to start from an unconstrained solution.

Moreover, the proposed method is independent of the numerical method that is finally used to solve the derived unconstrained OCP. Besides the collocation method used in this paper, first investigations with indirect gradient and shooting methods have shown that the intrinsic incorporation of the constraints has further advantages over the classical constraint penalization concerning speed of convergence and non-violation of the constraints.

The performance of the proposed methodology has been tested for several challenging benchmark problems, including the Goddard problem with thrust and dynamic pressure constraints [23] and the space shuttle reentry problem with input and heating constraints [24]. Current research concerns the application of the proposed methodology to real-time trajectory optimization, in particular in the context of real-time iteration schemes [25]. Further research is done on the extension of the approach to a more general class of constraints.

APPENDIX A

A.1. Choice of saturation functions

An appropriate choice to construct the saturation function $\psi(\xi_1, c^\pm)$ is for instance

$$\psi(\xi_1, c^\pm) = c^+ - \frac{c^+ - c^-}{1+\exp(s\xi_1)} \quad \text{with } s = \frac{4}{c^+ - c^-} \tag{A1}$$

The parameter s adjusts the slope at the position $\xi_1 = 0$ and is chosen in (A1) to normalize the slope at $\xi_1 = 0$ to $\partial\psi/\partial\xi_1 = 1$. Note that the saturation limits $c^\pm$ are only reached asymptotically for $\xi_1 \to \pm\infty$.

The second saturation function in (16) with the new input $\tilde{u}$ and the saturation limits $\phi^\pm := \phi^\pm(\xi, z)$ is constructed accordingly by

$$\phi(\tilde{u}, \phi^\pm) = \phi^+ - \frac{\phi^+ - \phi^-}{1+\exp(s\tilde{u})} \quad \text{with } s = \frac{4}{\phi^+ - \phi^-} \tag{A2}$$

Other choices for asymptotic saturation functions can e.g. be obtained for tanh-functions.

A.2. Limit behavior of saturation functions

This appendix explores the behavior of the saturation functions $y_1 = \psi(\xi_1, c^\pm)$ and $\dot{\xi}_r = \phi(\tilde{u}, \phi^\pm(\xi, z))$ defined in (A1)–(A2) and the unboundedness of $\xi = (\xi_1, \ldots, \xi_r)^\mathsf{T}$ and $\tilde{u}$, when the state constraint in (9f) becomes active. This case corresponds to the limit problem for $\text{OCP}_\xi^\varepsilon$, if the solution $y^\varepsilon = \psi(\xi_1^\varepsilon, c^\pm)$ approaches a state-constrained optimal trajectory y^* for $\varepsilon \to 0$.

We consider a trajectory $y(t)$ with a constrained arc $y(t) = c^+$, $t \in [t_{\text{in}}, t_{\text{out}}]$ and investigate the behavior of ξ and $\tilde{u}$ at the entry point t_{in} with the entry conditions

$$y_1(t_{\text{in}}) = c^+, \quad y_i(t_{\text{in}}) = 0, \quad i = 1, \ldots, r \tag{A3a}$$

These interior boundary conditions follow from the normal form (9b)–(9d). For every admissible control u, the state vector $(y_1, \ldots, y_r, z)$ is a continuous function of time $t \in [0, T]$. Hence, the cascade structure of the dynamics (9b)–(9c) shows that all states $y_i, i = 2, \ldots, r$ have to be zero at the entry point t_{in} to a constrained arc $y_1 = c^+ = \text{const.}$ for $t \in [t_{\text{in}}, t_{\text{out}}]$.

DOI: 10.1002/oca

To allow precise statements, assume in addition that the input constraint $u \in [\bar{u}^-(y,z), \bar{u}^+(y,z)]$ is not active when entering the state constraint $y_1(t_{\text{in}}) = c^+$ and that u is continuous over t_{in}. This leads to

$$\dot{y}_r(t_{\text{in}}) = 0, \quad a^-(y,z) < 0 < a^+(y,z) \quad \text{for } t \to t_{\text{in}} \tag{A3b}$$

where $a^{\pm}(y,z)$ are the transformed input constraints (17a) for $\dot{y}_r$. The following investigations rely on a series expansions of $y^*(t)$, which we assume to exist:

A.2.1. Behavior of coordinates ξ. A power series of y_1 at $y_1(t_{\text{in}}) = c^+$ can be written as

$$y_1(t_{\text{in}} - \tau) = c^+ - (e_k \tau^k + e_{k+1}\tau^{k+1} + \mathcal{O}(\tau^{k+2})) \qquad k > r \tag{A4}$$

with $e_k \neq 0$ and $\tau \geqslant 0$ as a new time coordinate. Note that $k > r$ due to the entry conditions in (78), which imply $y_1^{(i)}(t_{\text{in}}) = 0$ for $i = 1, \ldots, r$.‡‡ The behavior of ξ_1 can be investigated by looking at (24). For $y_1 \to c^+$, the first log-term converges to $\log(c^+ - c^-)$, whereas the second term $\log(c^+ - y_1)$ becomes unbounded. Hence, ξ_1 can be approximated by

$$\begin{aligned} \xi_1(t_{\text{in}} - \tau) &\approx -\bar{c} \log(c^+ - y_1) \quad \text{for } y_1 \to c^+ \\ &= -\bar{c}k \log(e_k \tau) \\ &\quad -\bar{c} \log\left(1 + \frac{e_{k+1}}{e_k}\tau + \mathcal{O}(\tau^2)\right) \end{aligned} \tag{A5}$$

with $\bar{c} = (c^+ - c^-)/4$. The latter equation follows from (A4) and applying the sum rule $\log(p+q) = \log(p) + \log(1 + q/p)$. With the time derivatives $(\mathrm{d}^i/\mathrm{d}t^i)\xi_1(t) = (-1)^i(\mathrm{d}^i/\mathrm{d}\tau^i)\xi_1(t_{\text{in}} - \tau)$, this leads to the finite-time blowup behavior of the coordinates ξ and $\dot{\xi}_r$ for $\tau \ll 1$:

$$\begin{aligned} \xi_1(t_{\text{in}} - \tau) &\approx -\bar{c}k \log(e_k \tau) \\ \xi_{i+1}(t_{\text{in}} - \tau) &\approx (i-1)!\bar{c}k/\tau^i, \quad i = 1, \ldots, r-1 \\ \dot{\xi}_r(t_{\text{in}} - \tau) &\approx (r-1)!\bar{c}k/\tau^r \end{aligned} \tag{A6}$$

‡‡The value of k corresponds to the first derivative of $y_1(t)$, which is discontinuous over t_{in}, i.e. $y_1^{(k)}(t_{\text{in}}^-) \neq y_1^{(k)}(t_{\text{in}}^+) = 0$, and can be characterized (under certain assumptions) with respect to the order r of the constraint $y_1 = c(x)$, see [26, 27].

The faster unboundedness of the derivatives of ξ_1 can be observed in Figures 3 and 4 for Example (46).

A.2.2. Behavior of new input $\tilde{u}$. Further statements can be obtained for $\tilde{u}$ by using the saturation function (A2) with the limits (18b) to explicitly state the inverse function (25a):

$$\begin{aligned} \tilde{u} = &\frac{a^+ - a^-}{4\psi'(\xi_1, c^{\pm}) \circ \psi^{-1}} \\ &\times [\log(\dot{y}_r - a^-) - \log(a^+ - \dot{y}_r)] \end{aligned} \tag{A7}$$

where the arguments of $a^{\pm}(y,z)$ are omitted for the sake of simplicity. Note that inserting the relation (9c) for $\dot{y}_r$ leads to (26). The denominator term $\psi'(\xi_1, c^{\pm}) \circ \psi^{-1}$ in (A7) can be simplified with (24) and using (A4):

$$\begin{aligned} \psi_1'(\xi_1, \psi_1^{\pm}) \circ \psi^{-1} &= 4\frac{(y_1 - c^-)(c^+ - y_1)}{(c^+ - c^-)^2} \\ &\approx \frac{1}{\bar{c}}(e_k \tau^k + \mathcal{O}(\tau^{k+1})) \quad \text{for } \tau \ll 1 \end{aligned} \tag{A8}$$

The log-terms in (A7) are rewritten with $\dot{y}_r(t) = (-1)^r(\mathrm{d}^r/\mathrm{d}\tau^r)y_1(t_{\text{in}} - \tau)$ following from (A4):

$$\log\left(\frac{-a^- + \dot{y}_r}{a^+ - \dot{y}_r}\right)$$

$$= \log\left(\frac{-a^- - \bar{e}_k \tau^{k-r} + \mathcal{O}(\tau^{k-r+1})}{a^+ + \bar{e}_k \tau^{k-r} + \mathcal{O}(\tau^{k-r+1})}\right) \tag{A9a}$$

$$= \log\left(-\frac{a^-}{a^+} - \frac{\bar{e}_k}{a^+}\left(1 - \frac{a^-}{a^+}\right)\tau^{k-r} + \mathcal{O}(\tau^{k-r+1})\right) \tag{A9b}$$

$$= \log\left(-\frac{a^-}{a^+}\right) + \frac{\bar{e}_k}{a^-}\left(1 - \frac{a^-}{a^+}\right)\tau^{k-r} + \mathcal{O}(\tau^{k-r+1}) \tag{A9c}$$

where $\bar{e}_k = (-1)^r k!/(k-r)!e_k$. The expression (A9b) is derived by expanding the fraction in (A9a) with respect to the single numerator terms and applying the expansion rule $1/(1+p) = \sum_{i=0}^{\infty}(-1)^i p^i$. The last equation

DOI: 10.1002/oca

(A9c) follows from the standard expansion rules. In summary, (A7) can be written as

$$\tilde{u}(t_{\text{in}}-\tau)=\frac{\bar{c}}{4}(a^+-a^-)\frac{\log\left(-\dfrac{a^-}{a^+}\right)+\dfrac{\bar{e}_k}{a^-}\left(1-\dfrac{a^-}{a^+}\right)\tau^{k-r}+\mathcal{O}(\tau^{k-r+1})}{e_k\tau^k+\mathcal{O}(\tau^{k+1})} \tag{A10}$$

Owing to Assumption (A3b), $\log(-a^-/a^+)$ is bounded and $\tilde{u}$ behaves like $1/\tau^r$ corresponding to the blowup behavior of $\dot{\xi}_r$ in (A6). Hence, $\tilde{u}$ becomes unbounded (and locally non-square-summable) if $\dot{y}_r$ approaches one of the constraints $a^{\pm}(y,z)$, see (17a) and (A7). This shows that (at least if y_1 is series expandable) the penalty term on ξ_1 in the cost (27a) is actually not required, since the penalization of $\tilde{u}$ accounts for both constraints (9f), also see Remark 2. Note in particular that $\tilde{u}$ becomes even faster unbounded, if one of the constraints $a^{\pm}(y,z)$ tries to cross zero, which leads to unboundedness of $\log(-a^-/a^+)$.

A.2.3. Behavior of saturation limits $\phi^{\pm}(\xi,z)$. From (16), it can be concluded that $\gamma_{r+1}(\xi)\to 0$ for $y_1\to c^+$, since $\dot{y}_r(t_{\text{in}})=0$ and $\psi'\dot{\xi}_r\to 0$ for $\tau\to 0$, which follows from the series expansions (A6) and (A8) with $k>r$. Hence, the saturation limits (18b) behave like

$$\phi^{\pm}(\xi,z)\approx\frac{\bar{c}\tilde{a}^{\pm}(\xi,z)}{e_k\tau^k}\quad\text{for }\tau\ll 1 \tag{A11}$$

which shows that $\phi^{\pm}(\xi,z)$ becomes faster unbounded than $\dot{\xi}_r$ in (A6). Moreover, $\phi^{\pm}\to\pm\infty$ due to the inequality in (A3b). Since $\phi^{\pm}$ are the saturation limits for $\phi\in(\phi^-,\phi^+)$, the normalization of the saturation function (A2) leads to

$$\dot{\xi}_r=\phi(\tilde{u},\phi^{\pm}(\xi,z))\approx\tilde{u} \tag{A12}$$

when y_1 approaches the constraint c^+. This means that the second saturation function ϕ 'opens' and $\dot{\xi}_r$ becomes unconstrained (see Figure 4). The reason for this property is that, due to Assumption (78), y_1 can smoothly enter into the constraint c^+ without any interaction with the input constraint in (9f).

ACKNOWLEDGEMENTS

The authors would like to thank Francois Chaplais for the fruitful discussions during the postdoctoral year of the first author at the *Centre Automatique et Systèmes* in Paris.

REFERENCES

1. Hargraves C, Paris S. Direct trajectory optimization using nonlinear programming and collocation. *Journal of Guidance, Control, and Dynamics* 1987; **10**:338–342.
2. von Stryk O. Numerical solution of optimal control problems by direct collocation. *International Series of Numerical Mathematics* 1993; **111**:129–143.
3. Seywald H. Trajectory optimization based on differential inclusion. *Journal of Guidance, Control, and Dynamics* 1994; **17**(3):480–487.
4. Betts J. Survey of numerical methods for trajectory optimization. *Journal of Guidance, Control, and Dynamics* 1998; **21**:193–207.
5. Betts J. *Practical Methods for Optimal Control Using Nonlinear Programming*. Society for Industrial and Applied Mathematics (SIAM): Philadelphia, PA, 2001.
6. Nocedal J, Wright S. *Numerical Optimization*. Springer: New York, 2006.
7. Bryson A, Ho YC. *Applied Optimal Control*. Ginn & Company: Waltham, MA, 1969.
8. Pesch H. A practical guide to the solution of real-life optimal control problems. *Control and Cybernetics* 1994; **23**:7–60.
9. Kreim H, Kugelmann B, Pesch H, Breitner M. Minimizing the maximum heating of a reentry space shuttle: an optimal control problem with multiple control constraints. *Optimal Control Applications and Methods* 1996; **17**:45–69.
10. Bryson A. *Dynamic Optimization*. Addison-Wesley: Menlo Park, CA, 1999.
11. Bonnard B, Faubourg L, Launay G, Trélat E. Optimal control with state constraints and the space shuttle re-entry problem. *Journal of Dynamical and Control Systems* 2003; **9**:155–199.
12. Pontryagin L, Boltyansky V, Gamkrelidze V, Mischenko E. Mathematical Theory of Optimal Processes. Wiley-Interscience: New York, 1962.

DOI: 10.1002/oca

13. Isidori A. *Nonlinear Control Systems* (3rd edn). Springer: Berlin, 1995.
14. Graichen K. Feedforward control design for finite-time transition problems of nonlinear systems with input and output constraints. *Doctoral Thesis*, Universität Stuttgart. Shaker Verlag: Aachen, Germany, 2006. Available at http://elib.uni-stuttgart.de/opus/volltexte/2007/3004.
15. Bonnans J, Guilbaud T. Using logarithmic penalties in the shooting algorithm for optimal control problems. *Optimal Control Applications and Methods* 2003; **24**: 257–278.
16. Wright M. Interior methods for constrained optimization. *Acta Numerica* 1992; **1**:341–407.
17. Lasdon L, Waren A, Rice R. An interior penalty method for inequality constrained optimal control problems. *IEEE Transactions on Automatic Control* 1967; **12**:388–395.
18. Ascher U, Mattheij R, Russell R. *Numerical Solution of Boundary Value Problems of Ordinary Differential Equations*. Prentice-Hall: Englewood Cliffs, NJ, 1988.
19. Shampine L, Kierzenka J, Reichelt M. Solving boundary value problems for ordinary differential equations in MATLAB with bvp4c. http://www.mathworks.com/bvp_tutorial 2000.
20. Chaplais F, Petit N. Inversion in indirect optimal control of multivariable systems. *ESAIM*: *Control*, *Optimization and Calculus of Variations* 2008; **14**(2):294–317.
21. Milam M. Real-time optimal trajectory generation for constrained dynamical systems. *Ph.D. Thesis*, California Institute of Technology, Pasadena, CA, 2003.
22. Fliess M, Lévine J, Martin P, Rouchon P. Flatness and defect of nonlinear systems: introductory theory and examples. *International Journal of Control* 1995; **61**:1327–1361.
23. Graichen K, Petit N. Solving the Goddard problem with thrust and dynamic pressure constraints using saturation functions. *Proceedings of the 17th IFAC World Congress*, Seoul, Korea, 2008; 14301–14306.
24. Graichen K, Petit N. Constructive methods for initialization and handling mixed state-input constraints in optimal control. *Journal of Guidance*, *Control*, *and Dynamics* 2008; **31**(5):1334–1343.
25. Diehl M, Bock H, Schlöder J. A real-time iteration scheme for nonlinear optimization in optimal feedback control. *SIAM Journal on Control and Optimization* 2005; **43**:1714–1736.
26. Jacobson D, Lele M, Speyer J. New necessary conditions of optimality for control problems with state-variable inequality constraints. *Journal of Mathematical Analysis and Applications* 1971; **35**:255–284.
27. Bonnans J, Hermant A. No-gap second-order optimality conditions for optimal control problems with a single state constraint and control. *Mathematical Programming* 2009; **117**(1–2):21–50.

DOI: 10.1002/oca