OPTIMAL CONTROL APPLICATIONS AND METHODS
Optim. Control Appl. Meth. (2014)
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/oca.2134

An interior penalty method for optimal control problems with state and input constraints of nonlinear systems

P. Malisani[1], F. Chaplais[2] and N. Petit[2,*,†]

[1]*EDF R&D Centre des Renardières Route de Sens 77818 Moret-sur-Loing, France*
[2]*MINES ParisTech, PSL Research University, CAS - Centre automatique et systèmes, 60 bd St Michel 75006 Paris, France*

SUMMARY

This paper exposes a methodology to solve state and input constrained optimal control problems for nonlinear systems. In the presented 'interior penalty' approach, constraints are penalized in a way that guarantees the strict interiority of the approaching solutions. This property allows one to invoke simple (without constraints) stationarity conditions to characterize the unknowns. A constructive choice for the penalty functions is exhibited. The property of interiority is established, and practical guidelines for implementation are given. A numerical benchmark example is given for illustration. © 2014 The Authors. Optimal Control Applications and Methods published by John Wiley & Sons, Ltd.

Received 22 May 2013; Revised 25 November 2013; Accepted 25 June 2014

KEY WORDS: optimal control; constraints; state constraints; interior methods; penalty design

1. INTRODUCTION

This paper exposes a methodology allowing one to solve constrained optimal control problems (COCP) for general multi-input multi-output system with nonlinear dynamics. This methodology belongs to the class of *interior point methods* (IPMs), commonly considered in finite-dimensional optimization, which consists in approaching the optimum by a path strictly lying inside the constraints. A main motivation for using this approach is that, in the interior, optimality conditions are much easier to formulate explicitly.

The idea driving penalty methods (for both finite-dimensional optimization problems and optimal control problems) is as follows. An augmented performance index is considered. It is constructed as the sum of the original cost function and so-called *penalty functions* that have some diverging asymptotic behavior when the constraints are approached by any tentative solution. This augmented performance index can then be optimized in the absence of constraints, yielding a biased estimate of the solution of the original problem. Two kinds of penalty methods exist: exterior penalty and interior penalty (a.k.a. barrier methods). In both approaches, minimization of the augmented performance index favors satisfaction of the constraints, depending on the weight of the penalty. In exterior penalty methods, for each augmented problem, the solution usually violates the constraints. The weight factor of the penalty should be increased so that this violation becomes sufficiently small. When this parameter is varied, the successive solutions generate a path lying outside of the

*Correspondence to: N. Petit, MINES ParisTech, PSL Research University, CAS - Centre automatique et systèmes, 60 bd St Michel 75006 Paris, France.
†E-mail: nicolas.petit@mines-paristech.fr

constraints. In IPMs, by construction, only feasible solutions are generated, but these are biased (suboptimal). The weight factor should be set sufficiently small to reduce the bias to an acceptable value. From a practical viewpoint, exterior methods are usually considered as more robust than interior ones. On the other hand, interior methods are attractive because they generate only feasible solutions. This can be an interesting particularity in numerous applications of automatic control (especially in closed-loop receding horizon approaches where satisfaction of constraints is more important than optimality; see [1] and references therein). All penalty methods are computationally appealing, as they yield *unconstrained* problems for which a vast range of highly effective algorithms are available. In finite-dimensional optimization, outstanding algorithms have resulted from the careful analysis of the choice of penalty functions and the sequence of weights. In particular, the IPMs that are nowadays implemented in successful software packages such as KNITRO [2], OOQP [3] have their foundations in these approaches. We refer the interested reader to [4] for a historical perspective on this topic.

In this article, we apply similar penalty methods to solve COCPs in a spirit similar to [5–8]. COCPs represent a very handy formulation of objectives in numerous applications, especially because constraints are very natural in problems of engineering interest. Unfortunately, these constraints induce some serious difficulties [9–11]. In particular, it is a well-known fact (see, e.g., [11]) that constraints bearing on state variables are difficult to characterize, as they generate both constrained and unconstrained arcs along the optimal trajectory. To determine optimality conditions, it is usually necessary to know or to a priori postulate the sequence and the nature of the arcs constituting the desired optimal trajectory. Active or inactive parts of the trajectory split the optimality system in as many coupled subsets of algebraic and differential equations. Yet, not much is known on this sequence, and this often results in a high complexity. Therefore, it is often preferred to use a discretization-based approach to this problem, and to treat it, for example, through a collocation method [12], as a finite-dimensional problem [13–19]. Under this finite-dimensional form, IPMs have been applied to optimal control problems in [20–22] and [23]. This is not the path that we explore, as we want to deal with optimal control problems under their original infinite-dimensional representation.

There is a well-established literature on the mathematical foundations of IPMs for finite-dimensional mathematical programming [3]. There exist also recent works in the field of exact penalty methods for various types of optimal control problem [24–29]. These methods are of particular interest because each solution of the sequence of optimal control problem is easily computed using classical stationarity conditions of the solution. The problem can be formulated, for example, by parameterizing the control variables using a finite number of values and switching times (to be determined). The constraints, on the state and control variables, and the dynamics can be expressed in terms of the unknowns and eventually treated as penalties. After the sensitivities of the objective function with respect to these unknowns are computed, iterative descent algorithms can be used to generate the solution. Nonsmooth methods can also be employed, depending on the smoothness of the constraints transcription [29]. In IPMs, the main difficulty is to guarantee that the sequence of solution is strictly interior. This point is critical because interiority is a requirement to avoid ill-posedness and computational failure of implemented algorithms. The problem of interiority in infinite-dimensional optimization has been addressed in [30] for input-constrained optimal control, and in [31–33], respectively, for linear systems, single input single output nonlinear systems, and multi-variable nonlinear systems with cubic input constraints. These contributions provide penalty functions guaranteeing the interiority of the solutions. The purpose of this article is to propose a synthetic view by generalizing previous historic and more recent results to nonlinear systems whose inputs belong to a general convex set, using new elements of proof and to expose the method in a friendly way for practitioners.

The paper is organized as follows. Section 2 contains the problem statement and sketches the contribution. Sections 3–5 present progressively eliminate the constraints from the formulation of the penalized problems. In Section 3, it is shown that a suitable choice of the state penalty guarantees that any trajectory yielding a finite cost is interior. In Section 4, a control penalty is added. Together with the state penalty, they guarantee that any optimal control for the penalized problem lies in the interior of the control constraints with a state that is in the interior of the state constraints.

DOI: 10.1002/oca

Capitalizing on the interiority of the previous optimal controls, Section 5 uses a so-called saturation functions to transform the control constrained problem of Section 4 into a fully unconstrained optimal control problem. Section 6 recalls the convergence of the costs, and a classic convexity argument shows the convergence of the optimal controls and states for the penalized problems. Solving algorithms are presented in Section 7. For convenience, a practical guide or 'cookbook' is proposed to help the users in deriving the equations needed for implementation. Section 8 gives a numerical application of the previous methods and algorithms by solving the Goddard problem [34] with an atmospheric pressure constraint. The source code for this example is freely available for interested readers. Finally, Section 9 gives conclusions and perspectives. It is followed by appendices containing several proofs that have been omitted from the main stream of the paper, and a recall (in Appendix A) of the general convergence result by Fiacco and McCormick, which is inspirational to the presented work.

2. NOTATIONS, PROBLEM STATEMENT, AND PRESENTATION OF THE CONTRIBUTION

The COCP that we wish to determine a solution method for in this article is to determine u^* a global solution of

$$\min_{u\in\mathcal{U}\cap\mathcal{X}} \left[J(u) = \int_0^T \ell(x^u, u)dt \right] \tag{1}$$

for the dynamics

$$\dot{x}^u(t) = f(x^u(t), u(t)), \qquad x(0) = x_0 \tag{2}$$

where $\ell : \mathbb{R}^n \times \mathbb{R}^m \mapsto \mathbb{R}$ is a locally Lipschitz function of its arguments, continuously differentiable with respect to u, $x^u(t) \in \mathbb{R}^n$ and $u(t) \in \mathbb{R}^m$ are the state and the control, which satisfy the (multi-input multi-output) nonlinear dynamics (2), f is continuously differentiable, for some $D > 0$ $\|f(x,u)\| \leqslant D(1+\|x\|),\ \forall x \in \mathbb{R}^n, \forall u \in \mathcal{C}$, where $\mathcal{C}$ is a bounded closed subset of $\mathbb{R}^m$ defined by

$$\mathcal{C} \triangleq \left\{ u = (u_1, \ldots, u_p) \in \mathbb{R}^{m_1} \times \ldots \times \mathbb{R}^{m_p},\ p \geqslant 1, \text{ with } \sum m_i = m \text{ s.t. } u_i \in \mathcal{C}_i \subset \mathbb{R}^{m_i} \right\}$$

where each $\mathcal{C}_i$ is a bounded closed convex subset of $\mathbb{R}^{m_i}$, which has a nonempty interior containing 0 with continuously differentiable boundary.[‡]

The set $\mathcal{U} \cap \mathcal{X}$ is defined by *control and state constraints* that we detail below. The control $u : \mathbb{R} \mapsto \mathbb{R}^m$ is constrained to belong to the following set, which is closed and convex[§]

$$\mathcal{U} \triangleq \{u \in L^\infty([0,T], \mathbb{R}^m) \text{ s.t. } u(t) \in \mathcal{C} \text{ a.e. } t \in [0,T]\} \tag{3}$$

The set $\mathcal{X}$ is defined as follows

$$\mathcal{X} \triangleq \left\{ u \in L^\infty([0,T], \mathbb{R}^m) \text{ s.t. } x^u(t) \in X^{\text{ad}} \text{ for all } t \in [0,T] \right\} \tag{4}$$

where

$$X_{\text{ad}} \triangleq \{x \in \mathbb{R}^n \text{ s.t. } g_i(x) \leqslant 0,\ i = 1, \ldots, q\}$$

where the g_i are continuously differentiable functions [¶] $\mathbb{R}^n \mapsto \mathbb{R}$.

To implement IPMs, we shall naturally make the following assumption.

[‡]For example, this setting allows one to consider the case where $\mathcal{C}$ is $\{u \in \mathbb{R}^3 \text{ s.t. } u_1^2 + u_2^2 \leqslant 1\ , |u_3| \leqslant 1\}$. The boundary of $\mathcal{C}$ is not differentiable, yet $u \in \mathcal{C}$ can be rewritten as $u = (u_1, u_2)$, where u_1 belongs to an appropriate Euclidean disk and u_2 belongs to an appropriate segment of $\mathbb{R}$. Conveniently, the employed formalism covers the case where $\mathcal{C}$ is a hypercube.

[§]A proof of this statement is given in Appendix B.

[¶]Observe that X_{ad} is convex if the functions g_i are convex.

DOI: 10.1002/oca

Assumption 1 (strict interiority of the initial condition)
The initial condition x_0 of (2) belongs to the open set[‖]

$$\overset{\circ}{X}_{\text{ad}} \triangleq \{x \in \mathbb{R}^n \text{ s.t. } g_i(x) < 0,\ i = 1, \dots, q\}$$

2.1. Summary of the contribution

The contribution of this article is an interior penalty method, which uses a so-called *penalty functions* p_1 and p_2 and a so-called *generalized saturation function* ϕ to formulate the following problem

$$\min_{\nu \in L^\infty([0,T],\mathbb{R}^m)} \int_0^T \ell\Big(x^{\phi(\nu)}, \phi(\nu)\Big)dt + \epsilon \int_0^T p_1\Big(x^{\phi(\nu)}\Big) + p_2(\nu)dt \tag{5}$$

which is shown to generate, as $\epsilon \to 0$ a sequence of solutions converging to a solution of the COCP defined earlier. Each solution in the sequence is relatively easy to determine as the problem (5) is unconstrained. The next sections explain how p_1, p_2, and ϕ are constructed and establish equivalence and convergence results. A tutorial summary is given in the 'cookbook' of Section 7.1.

3. ELIMINATION OF THE STATE CONSTRAINT BY STATE PENALTY

A first step of the methodology we propose is to penalize the state constraints.

3.1. State penalty

We introduce a state penalty function $\gamma_g : \ (-\infty, +\infty) \ \to [0, +\infty)$, which satisfies the following assumption.

Assumption 2 (Properties of the state penalty)
The function γ_g is positive, continuously differentiable, convex, and increasing over $\{x < 0\}$. It is null over $\{x \geqslant 0\}$. It is singular in 0^- as $\lim_{x \uparrow 0} \gamma_g(x) = +\infty$.

To address the state constraints defined in (4), the following integral state penalty is introduced, for $u \in \mathcal{U}$

$$P_g(u) = \int_0^T \sum_{i=1}^{q} \gamma_g \circ g_i(x^u)dt$$

where $\circ$ denotes the function composition. To study the impact of this penalty, we introduce the notations

$$\mathcal{X}^{\text{strict}} = \left\{u \in L^\infty([0,T], \mathbb{R}^m) \text{ s.t. } x^u(t) \in \overset{\circ}{X}_{\text{ad}} \ \ \forall t \in [0,T]\right\}$$
$$\mathcal{U} \setminus \mathcal{X}^{\text{strict}} = \left\{u \in \mathcal{U} \text{ s.t. } u \notin \mathcal{X}^{\text{strict}}\right\}$$

As will appear, one can determine a sufficient condition on the state penalty γ_g ensuring that any control $u \in \mathcal{U}$ yielding a bounded penalty $P_g(u)$ necessarily belongs to $\mathcal{X}^{\text{strict}}$.

Definition 1 (Proximity to a constraint)
For any constraint g_i, we define the proximity to the constraint as

$$\alpha \mapsto \mu_{g_i}(u, \alpha) = \text{meas}\left(\{t \in [0,T] \text{ s.t. } 0 > g_i(x^u(t)) \geqslant -\alpha\}\right) \tag{6}$$

where meas(.) is the Lebesgue measure of its argument.

[‖]The set $\overset{\circ}{X}_{\text{ad}}$ is an open set. Moreover, if Assumption 1 holds and if the functions g_i are convex, then $\overset{\circ}{X}_{\text{ad}}$ is dense in X_{ad}. A proof of this result can be found in Appendix C.

DOI: 10.1002/oca

Proposition 1 (Unboundedness of integral penalty)
If, for all $u \in \mathcal{U} \setminus \mathcal{X}^{\text{strict}}$, the penalty function γ_g satisfies

$$\lim_{\alpha \downarrow 0} \gamma_g(-\alpha)\mu_{g_i}(u,\alpha) = +\infty \tag{7}$$

then $\forall u \in \mathcal{U} \setminus \mathcal{X}^{\text{strict}}$

$$P_g(u) = +\infty$$

Proof
Let $u \in \mathcal{U} \setminus \mathcal{X}^{\text{strict}}$, then there exists an index i such that $\max_{t\in[0,T]} g_i(x(t)) \geqslant 0$. Because $\gamma_g(x) = 0$ when $x \geqslant 0$, we have

$$\mathcal{I}_i \triangleq \int_0^T \gamma_g(g_i(x(t)))dt = \int_{0>g_i(x(t))} \gamma_g(g_i(x(t)))dt$$

Moreover, because $\gamma_g \geqslant 0$, we have, for $\alpha > 0$,

$$\mathcal{I}_i \geqslant \int_{0>g_i(x(t))\geqslant -\alpha} \gamma_g(g_i(x(t)))dt \triangleq \mathcal{J}_i(\alpha)$$

The state penalty satisfies $\gamma_g \geqslant 0$ on $(-\infty, 0)$; thus, $\mathcal{J}_i(\alpha)$ is a nondecreasing positive continuous function of $\alpha > 0$, which satisfies

$$\inf_{\alpha>0} \mathcal{J}_i(\alpha) = \lim_{\alpha\downarrow 0} \mathcal{J}_i(\alpha) \triangleq \mathcal{J}_i(0^+)$$

Because γ_g is increasing and because the Lebesgue measure is right-continuous (see, e.g., [35])

$$\begin{aligned}\mathcal{J}(0^+) &= \lim_{\alpha\downarrow 0} \int_{0>g_i(x(t))\geqslant -\alpha} \gamma_g(g_i(x(t)))dt \\ &\geqslant \lim_{\alpha\downarrow 0} \int_{0>g_i(x(t))\geqslant -\alpha} \gamma_g(-\alpha)dt = \lim_{\alpha\downarrow 0} \gamma_g(-\alpha)\mu_{g_i}(u,\alpha)\end{aligned}$$

with $\mu_{g_i}(u,\alpha)$ the Lebesgue measure defined in (6). If (7) holds, then $\mathcal{J}_i(0^+) = +\infty$, which implies that $\mathcal{I}_i = +\infty$. As a consequence, $P_g(u) = +\infty$. This concludes the proof. □

Because the measure $\mu_{g_i}(u,\alpha)$, which appears in (7), involves the control u, it is handy to give a lower bound of it when u spans $\mathcal{U} \setminus \mathcal{X}^{\text{strict}}$.

Proposition 2 (Lower-bound on the proximity to a constraint)
Under Assumption 1, $\alpha_0 \triangleq -\max_i(g_i(x_0)) > 0$. Then, there exists a constant $\Gamma < +\infty$ such that for all $\alpha \in [0, \alpha_0]$, for all $u \in \mathcal{U} \setminus \mathcal{X}^{\text{strict}}$ the measure $\mu_{g_i}(u,\alpha)$ defined in (6) is lower-bounded as follows

$$\mu_{g_i}(u,\alpha) \geqslant \frac{\alpha}{\Gamma}$$

The proof is given in Appendix D.

3.2. First main result

Collecting Propositions 1 and 2, one finally obtains our first main result.

Theorem 1 (Interiority of the state)
Under Assumptions 1 and 2, if the state penalty γ_g is such that

$$\lim_{\alpha\downarrow 0} \alpha\gamma_g(-\alpha) = +\infty \tag{8}$$

then for every control $u \in \mathcal{U}$, the penalty value $P_g(u)$ is finite if and only if $u \in \mathcal{X}^{\text{strict}}$.

DOI: 10.1002/oca

To satisfy (8), a possible choice for γ_g is

$$\gamma_g(x) = \begin{cases} (-x)^{-n_g}, & n_g > 1, \quad \text{for } x < 0 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

Proof
If (8) holds, then we derive from Proposition 2 that (7) holds for $u \in \mathcal{U} \setminus \mathcal{X}^{\text{strict}}$. From Proposition 1, we derive that $P_g(u) < +\infty$ for $u \in \mathcal{U}$ only if $u \in \mathcal{U} \cap \mathcal{X}^{\text{strict}}$. Conversely, if $\in \mathcal{U} \cap \mathcal{X}^{\text{strict}}$, then the trajectories are well defined and continuous; therefore, the maximum value of $g_i(t)$ for $t \in [0, T]$ and any i is strictly negative and the penalty $P_g(u)$ is finite because γ_u is bounded. Finally, one easily verifies that the penalty function defined by (9) satisfies Assumption 2 and (8). □

3.3. Introduction of a first penalized problem

Corollary 1
Under the assumptions of Theorem 1, for every $\epsilon > 0$, the three following problems are equivalent

$$\min_{u \in \mathcal{U}} J(u) + \epsilon P_g(u) \tag{10}$$

$$\min_{u \in \mathcal{U} \cap \mathcal{X}} J(u) + \epsilon P_g(u) \tag{11}$$

$$\min_{u \in \mathcal{U} \cap \mathcal{X}^{\text{strict}}} J(u) + \epsilon P_g(u) \tag{12}$$

in the sense that they have the same set of minimizers and the same optimal values. In addition, the penalty P_g in problem (12) is bounded for all of the controls in the given control set.

Proof
Obviously, the values of each minimum are in increasing order. However, to be optimal for problems (10) and (11), a control u must belong to $\mathcal{U}$ and to $\mathcal{X}^{\text{strict}}$. Therefore, the optimum for (12) is smaller or equal to the optimum for (10), and they are all equal, and their minimizers all belong to $\mathcal{U} \cap \mathcal{X}^{\text{strict}}$. Finally, (12) yields a finite penalty by application of Theorem 1. □

Furthering Corollary 1, we now wish to let $\epsilon \to 0$ in (10), or equivalently (11) or (12). However, this requires to have a minimizing sequence for which the penalty is bounded, that is, that belongs to $\mathcal{U} \cap \mathcal{X}^{\text{strict}}$. Therefore, we make the following assumption.

Assumption 3 (Existence of an approaching interior sequence)
Let $v_0 = \inf_{u \in \mathcal{U} \cap \mathcal{X}} J(u)$. For any $\omega > 0$, there exists $s(\omega) \in \mathcal{U} \cap \mathcal{X}^{\text{strict}}$ such that $v_0 \leqslant J(s(\omega)) \leqslant v_0 + \omega$.

This assumption is satisfied if, for instance, $\mathcal{U} \cap \mathcal{X}^{\text{strict}}$ is dense in $\mathcal{U} \cap \mathcal{X}$ for the sup norm, because J is continuous (Proposition 13).

Proposition 3
Assume that the g_i are convex, that the dynamics (2) is linear, and that $\mathcal{U} \cap \mathcal{X}^{\text{strict}}$ is not empty. Then, $\mathcal{X}$ and $\mathcal{X}^{\text{strict}}$ are convex, and $\mathcal{U} \cap \mathcal{X}^{\text{strict}}$ is dense in $\mathcal{U} \cap \mathcal{X}$ for the sup norm, and consequently, Assumption 3 is satisfied.

Proof
With linear dynamics, for $\lambda \in [0, 1]$, and two controls u and v, $x^{\lambda u + (1-\lambda)v} = \lambda x^u + (1 - \lambda)x^v$. Because the g_i are convex, we have

$$g_i\Big(x^{\lambda u + (1-\lambda)v}(t)\Big) = g_i(\lambda x^u(t) + (1 - \lambda)x^v(t)) \leqslant \lambda g_i(x^u(t)) + (1 - \lambda) g_i(x^v(t))$$

DOI: 10.1002/oca

If we take $u, v \in \mathcal{X}$, this proves that $\mathcal{X}$ is convex. Similarly, if $u, v \in \mathcal{X}^{\text{strict}}$, this proves that $\mathcal{X}^{\text{strict}}$ is convex. Finally, by taking $u \in \mathcal{U} \cap \mathcal{X}^{\text{strict}}$ (which is nonempty) and $v \in \mathcal{U} \cap \mathcal{X}$ and $\lambda \in (0, 1)$, we obtain

$$g_i\Big(x^{\lambda u+(1-\lambda)v}(t)\Big) \leqslant \lambda g_i(x^u(t)) < 0$$

Also, $\lambda u + (1 - \lambda)v \in \mathcal{U}$ because $\mathcal{U}$ is convex and closed. By letting λ tend to 0, this proves that $\mathcal{U} \cap \mathcal{X}^{\text{strict}}$ is dense in $\mathcal{U} \cap \mathcal{X}$ for the sup norm. □

Corollary 2
Assume also that for ϵ small enough, Penalized Optimal Control Problem (POCP) (10) has at least a minimizer $u^*(\epsilon)$, and that Assumption 3 holds. Then,

A $\lim_{\epsilon\downarrow 0} J(u^*(\epsilon)) = \inf_{u\in\mathcal{U}\cap\mathcal{X}} J(u)$
B $\lim_{\epsilon\downarrow 0} \epsilon P_g(u^*(\epsilon)) = 0$

Proof
Because of Corollary 1, $u^*(\epsilon)$ is also a solution of (11) and (12), and we can apply the classic Theorem 5 by Fiacco and McCormick, recalled for convenience in Appendix, with $S = \mathcal{U} \cap \mathcal{X}^{\text{strict}} \subset R = \mathcal{U} \cap \mathcal{X}$; observe that we have a minimizing sequence in S because of Assumption 3. So we can apply Theorem 5 with the limit bearing on the solutions of (12). The set of these solutions is equal to the set of the solutions of (10). This gives the desired conclusion. □

4. INTERIORITY OF THE CONTROL BY STATE AND CONTROL PENALTY

We now add a penalty bearing on the control to prevent it from touching the boundary of $\mathcal{C}$. The penalty uses a *gauge function* as argument. Basic properties of gauge functions necessary to understand their relevance in the present context are recalled in the next section.

4.1. Gauge functions of convex sets

Classically, one can associate a gauge function $G_{\mathcal{C}}$ to any convex set $\mathcal{C}$. Under some mild assumptions, the gauge acts almost like a norm and reveals handy in our problem formulation. Conveniently, the fact that a vector u belongs to the interior, boundary or exterior of $\mathcal{C}$ boils down to comparing $G_{\mathcal{C}}(u)$ to 1.

Definition 2 (Gauge function [36]*)*
The gauge function of $\mathcal{C}$ is the mapping $G_{\mathcal{C}} : \mathbb{R}^m \mapsto \mathbb{R}^+$ defined by

$$G_{\mathcal{C}}(u) = \inf\{\lambda \geqslant 0 \text{ s.t. } u \in \lambda\mathcal{C}\} \tag{13}$$

Proposition 4
The gauge function $G_{\mathcal{C}}$ has the following properties

(a) $G_{\mathcal{C}}(u)$ is a well defined nonnegative real for all u
(b) There exists $0 < N < M$ such that

$$\frac{\|u\|}{M} \leqslant G_{\mathcal{C}}(u) \leqslant \frac{\|u\|}{N} \ \forall u \in \mathbb{R}^m \tag{14}$$

In particular, $G_{\mathcal{C}}(u) = 0$ implies $u = 0$
(c) The gauge is positively homogeneous, that is, $G_{\mathcal{C}}(\lambda u) = \lambda G_{\mathcal{C}}(u)$ for all $\lambda \geqslant 0$
(d) $G_{\mathcal{C}}$ is a strictly convex function which is locally bounded; as a consequence, it is continuous

DOI: 10.1002/oca

(e) $G_{\mathcal{C}}$ has a directional derivative in the sense of Dini[**] at $u = 0$ along direction d and its value is $G_{\mathcal{C}}(d)$
(f) $G_{\mathcal{C}}$ is differentiable on $\mathbb{R}^m \setminus \{0\}$
(g) **[main result for later discussions]** $G_{\mathcal{C}}(u) < 1$ if and only if u belongs to the interior of $\mathcal{C}$; $G_{\mathcal{C}}(u) = 1$ if and only if u belongs to the boundary $\partial\mathcal{C}$ of $\mathcal{C}$; $G_{\mathcal{C}}(u) > 1$ if and only if u belongs to the exterior of $\mathcal{C}$.

These properties are proven in Appendix E.

As a consequence, $u(t) \in \mathcal{C}$ if and only if $G_{\mathcal{C}}(u_i(t)) \leqslant 1, \forall i$ and $u(t)$ belongs to the interior of $\mathcal{C}$ if and only if $G_{\mathcal{C}}(u_i(t)) < 1, \ \forall i$.

Remarks Under our assumptions on $\mathcal{C}$, the gauge is a norm if and only if $\mathcal{C}$ is symmetric with respect to the origin. If the origin does not belong to the interior of $\mathcal{C}$, but instead another vector u_0 belongs to this interior, then we define the gauge on the variable $u - u_0$. Finally, if $\mathcal{C}$ is equivalently defined by a set of inequations $c_j(u) \leqslant 0$, then (13) is equivalent to

$$G_{\mathcal{C}}(u) = \inf\left\{\lambda \geqslant 0 \text{ s.t. } c_j\left(\frac{u}{\lambda}\right) \leqslant 0 \ , \ \forall i\right\}$$

4.2. *Introduction of the control penalty function*

We first introduce a generic control penalty function $\gamma_u : \ [0,1] \ \to [0,+\infty)$ that satisfies the following assumption

Assumption 4 (properties of the control penalty)
The function γ_u is such that

- γ_u is continuously differentiable, strictly convex, and nondecreasing
- $\lim_{u\uparrow 1} \gamma_u(u) = +\infty$
- $\gamma_u(0) = 0$; γ_u is right continuously differentiable at $u = 0$ with $\gamma_u'(0) = 0$
- $\gamma_u'(u)$ is locally Lipschitz at $u = 0$

For $u \in \mathcal{U}$, we introduce the following integral control penalty

$$P_u(u) = \int_0^T \sum_{i=1}^{p} \gamma_u \circ G_{\mathcal{C}_i}(u_i(t))dt \tag{15}$$

From the properties of the gauge functions and of γ_u we see that $\gamma_u \circ G_{\mathcal{C}_i}(u_i(t))$ tends to infinity when $u_i(t)$ tends to the boundary of $\mathcal{C}_i$. Before defining a new penalized problem, let us give some useful properties of the penalty function.

Proposition 5 (Differentiability and convexity)
For $i = 1, \ldots, p$, the application $\gamma_u \circ G_{\mathcal{C}_i}$ is continuously differentiable on the interior of $\mathcal{C}_i$, and convex. As a consequence, the integrand in the integral penalty (15) is continuously differentiable with respect to the control u in the interior of $\mathcal{C}$ and convex.

Proof
From Proposition 4, we know that $G_{\mathcal{C}_i}$ is continuously differentiable on $\mathbb{R}^{m_i} \setminus \{0\}$ because the boundary $\partial\mathcal{C}_i$ is continuously differentiable. On the other hand, γ_u is continuously differentiable on $[0,1)$; hence, $\gamma_u \circ G_{\mathcal{C}_i}$ is continuously differentiable on the interior of $\mathcal{C}_i$ minus the origin.

Because $G_{\mathcal{C}}$ has bounded derivatives at $u = 0$ in the sense of Dini, and because $\gamma_u'(G_{\mathcal{C}}(0)) = \gamma_u'(0) = 0$, we conclude that $\gamma_u \circ G_{\mathcal{C}}$ has a zero derivative at the origin. Moreover, γ_u' being Lipschitz (with constant K) in a neighborhood of 0, one has $|\gamma_u' \circ G_{\mathcal{C}}(u)| \leqslant K|G_{\mathcal{C}}(u)|$.

[**]The Dini derivative of a function f at point $x \in \mathbb{R}^n$ along the direction $d \in \mathbb{R}^n$ is the limit (when it exists) of $\frac{f(x+hd)-f(x)}{h}$ when h tends to 0 with positive values.

DOI: 10.1002/oca

We derive that the limit of the derivative of $\gamma_u \circ G_{\mathcal{C}}(u)$ is 0 when u tends to 0. We have seen that $G_{\mathcal{C}_i}$ is convex; because γ_u is convex, and because it is nondecreasing, then $\gamma_u \circ G_{\mathcal{C}_i}$ is convex. This concludes the proof. □

4.3. Second penalized problem

We are now ready to define a second penalized problem

$$\min_{u \in \mathcal{U} \cap \mathcal{X}} K(u, \epsilon) = J(u) + \epsilon \left(P_g(u) + P_u(u) \right) \tag{16}$$

where P_g satisfies the assumptions of Theorem 1. Because of this, and because P_u is nonnegative and J is bounded over $\mathcal{U}$, problem (16) is equivalent to the two following problems

$$\min_{u \in \mathcal{U} \cap \mathcal{X}^{\text{strict}}} K(u, \epsilon) \tag{17}$$

$$\min_{u \in \mathcal{U}} K(u, \epsilon) \tag{18}$$

in the sense that they have the same set of minimizers and the same optimal values. Let us define

$$\mathcal{U}^{\text{strict}} = \left\{ u \in \mathcal{U} \text{ s.t. } \max_i \operatorname{ess\,sup}_t G_{\mathcal{C}_i}(u_i) < 1 \right\} \subset \mathcal{U}$$

We already know that any minimizer of (18) belongs to $\mathcal{X}^{\text{strict}}$. We are going to design γ_u such that it also belongs to $\mathcal{U}^{\text{strict}}$.

4.4. Construction of a neighboring interior control v

To show that optimal solutions of (16) are interior, we need to construct a neighboring control vector that are not touching the constraints. The construction of this starts with the following result.

Proposition 6
For all $u \in \mathcal{U} \cap \mathcal{X}^{\text{strict}}$, there exists $\alpha > 0$ such that, for all $v \in \mathcal{U}^{\text{strict}}$ satisfying $\|u - v\|_{L^\infty} \leqslant \alpha$, we have

$$v \in \mathcal{X}^{\text{strict}}$$

Proof
Let $u \in \mathcal{U} \cap \mathcal{X}^{\text{strict}}$ and note $-2\beta_0 = \max_{t \in [0,T], i = 1, \ldots, q} g_i(x^u(t))$. Because $u \in \mathcal{U} \cap \mathcal{X}^{\text{strict}}$, we have $\beta_0 > 0$. From Proposition 13, and the continuity of the function g, there exists $\alpha_N > 0$ and $\Lambda > 0$ such that for all $v \in \mathcal{U}^{\text{strict}}$

$$\max_i \|u_i - v_i\|_{L^\infty} \leqslant \alpha_N \Rightarrow \max_i \| \, g_i(x^u) - g_i(x^v) \, \|_{L^\infty} \leqslant \Lambda \alpha_N$$

Setting $\alpha = \beta_0 / \Lambda$, one has $\max_i \max_{t \in [0,T]} g_i(x^v(t)) \leqslant -\beta_0 < 0$. Therefore, $v \in \mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$. This concludes the proof. □

We now proceed to the construction of a control $v \in \mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$ which will be instrumental in establishing Proposition 8.

Definition 3 (De-saturated control)
For all $u \in \mathcal{U} \cap \mathcal{X}^{\text{strict}}$, for all $\alpha > 0$, we define a de-saturated control $v(u, \alpha) = (v_1 \cdots v_p)$ as follows

$$v_i(t) = \begin{cases} u_i(t) & \text{if } G_{\mathcal{C}_i}(u_i(t)) < 1 - \alpha \\ (1 - 2\alpha) u_i(t) & \text{otherwise} \end{cases} \tag{19}$$

DOI: 10.1002/oca

Proposition 7
For all $u \in \mathcal{U} \cap \mathcal{X}^{\text{strict}}$, there exists $\bar{\alpha} > 0$ such that, for all $\alpha \in (0, \bar{\alpha})$, the de-saturated control (19) satisfies

$$v \in \mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$$

As a consequence, $\mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$ is dense in $\mathcal{U} \cap \mathcal{X}^{\text{strict}}$ in the L^∞ sense.

Proof
We shall use the following definitions, inspired by Definition 1

$$\begin{aligned} E_u(\alpha) &\triangleq \{t \in [0,T] \text{ s.t. } \exists i \leqslant p \text{ s.t. } G_{\mathcal{C}_i}(u_i(t)) \geqslant 1-\alpha\} \\ \mu_u(\alpha) &\triangleq \text{meas}(E_u(\alpha)) \end{aligned} \tag{20}$$

First, let us prove that, for $\alpha \in (0, 1/2)$, $v \in \mathcal{U}^{\text{strict}}$. Assume that $\mu_u(\alpha) = 0$; in this case, for all i, $G_{\mathcal{C}_i}(u_i(t)) < 1-\alpha$ almost everywhere (a.e.). Therefore, $u \in \mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$. Using (19) yields $v = u \in \mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$.
Now, let us assume that $\mu_u(\alpha) > 0$. In this case, for all i,

$$\begin{aligned} &G_{\mathcal{C}_i}(v_i(t)) < 1-\alpha \quad \text{a.e. } t \in [0,T] \setminus E_u(\alpha) \\ &G_{\mathcal{C}_i}(v_i(t)) \leqslant (1-2\alpha) G_{\mathcal{C}_i}(u_i(t)) \leqslant 1-2\alpha \quad \forall t \in E_u(\alpha) \end{aligned}$$

because for all i, $u_i(t) \in \mathcal{C}_i$ a.e. Because $1-2\alpha \in (0,1)$, we see that $G_{\mathcal{C}_i}(v_i(t)) < 1-\alpha$ almost everywhere; therefore, $v \in \mathcal{U}^{\text{strict}}$.

We now prove that $v \in \mathcal{X}^{\text{strict}}$. Let M_i be the radius of a ball that contains $\mathcal{C}_i$; using (19), we have $\|u_i(t) - v_i(t)\| \leqslant 2\alpha \|u_i(t)\| \leqslant 2\alpha M_i$. From Proposition 6, there exists $\alpha^+ > 0$ such that if $\| u - v \|_{L^\infty} \leqslant \alpha^+$, then $v \in \mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$. Let $\bar{\alpha} = \min\{1/2, \min_i \frac{\alpha^+}{2M_i}\}$. Then, for all $\alpha \in (0, \bar{\alpha})$, we have

$$\|u_i(t) - v_i(t)\| \leqslant 2\alpha \|u_i(t)\| \leqslant \alpha^+, \quad i = 1, \ldots, p$$

Therefore, $v \in \mathcal{X}^{\text{strict}}$. Thus, $v \in \mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$. Finally, because $\mathcal{U}$ is bounded in $L^\infty[0,T]$ and $\alpha > 0$ can be chosen arbitrarily small, we see from (19) that v can be chosen arbitrarily close to u in $L^\infty[0,T]$. This concludes the proof. □

4.5. *Condition guaranteeing the strict interiority of the optimal control*

To prove that any optimal control belongs to $\mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$, it is enough to find a condition on the penalties such that for any $u \in (\mathcal{U} \setminus \mathcal{U}^{\text{strict}}) \cap \mathcal{X}^{\text{strict}}$, the de-saturated control $v \in \mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$ from Definition 3 satisfies

$$K(v, \epsilon) < K(u, \epsilon)$$

This fact contradicts the optimality of every point of $(\mathcal{U} \setminus \mathcal{U}^{\text{strict}}) \cap \mathcal{X}^{\text{strict}}$.

The following result gives an upper estimate on the difference $K(v,\epsilon) - K(u,\epsilon)$. This estimate is the sum of three terms, representing respectively an upper bound on $J(v) - J(u)$, an upper bound on the state penalty $\epsilon(P_g(v) - P_g(u))$ and the opposite of a lower bound on the difference $\epsilon(P_u(u) - P_u(v))$, which will be shown to be positive. In §4.6 we give constructive conditions on the penalties that make this upper bound strictly negative when $u \in (\mathcal{U} \setminus \mathcal{U}^{\text{strict}}) \cap \mathcal{X}^{\text{strict}}$.

Proposition 8
For any control $u \in (\mathcal{U} \setminus \mathcal{U}^{\text{strict}}) \cap \mathcal{X}^{\text{strict}}$, considering the modified control v from (19), for any $\epsilon > 0$ one has

$$K(v,\epsilon) - K(u,\epsilon) \leqslant \alpha \left[U_\ell + U_g(\epsilon) - \epsilon \gamma_u'(1-3\alpha) \right] \mu_u(\alpha) \tag{21}$$

where $\mu_u(\alpha)$ is defined by (20), U_l is a constant parameter and $U_g(\epsilon)$ depends linearly on ϵ.

DOI: 10.1002/oca

The proof of this result proceeds by calculus, notably variational calculus on convex functions. Details can be found in Appendix F.

Finally, using (21), the following result holds.

Proposition 9
If an optimal control u^* for POCP (18) belongs to $\mathcal{U} \cap \mathcal{X}^{\text{strict}}$, and if

$$\lim_{\alpha \uparrow 1} \gamma_u'(\alpha) = +\infty \tag{22}$$

then, $u^* \in \mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$.

Proof
Remember that if, for some $\alpha > 0$, $\mu_{u^*}(\alpha) = 0$, then $u^* \in \mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$. We shall now assume that $\mu_{u^*}(\alpha) > 0$ for $\alpha > 0$ in a neighborhood of 0. If u^* does not belong to $\mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$, then, using (22), for α small enough one can build a control $v \in \mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$ such that $K(v, \epsilon) < K(u, \epsilon)$ because of (21); this contradicts the assumed optimality of u^* and concludes the proof. □

4.6. Second main result

We are now ready to state our second main result.

Theorem 2 (Existence of penalties providing interior optima)
Under Assumptions 1, 2, and 4, there exists penalty functions γ_g and γ_u such that any optimal solution u^* of POCP (18) belongs to $\mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$. A constructive choice is given by (9) and

$$\gamma_u(u) = -u \log(1 - u) \qquad \text{for } u \in [0, 1) \tag{23}$$

$$\gamma_u(1) = +\infty \tag{24}$$

Proof
We know from Theorem 1 that using (9) as penalty γ_g guarantees that every optimal control u^* belongs to $\mathcal{U} \cap \mathcal{X}^{\text{strict}}$. The penalty function γ_u given by (23) and (24) satisfies Assumption 4. Further, elementary computations show that $\gamma_u'(1 - \alpha)$ satisfies (22). Because we have shown that u^* belongs to $\mathcal{U} \cap \mathcal{X}^{\text{strict}}$, Proposition 9 applies, which proves that the proposed penalties imply $u^* \in \mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$. □

Corollary 3 (Equivalence of constrained and unconstrained problems)
Define problems

$$\min_{u \in \mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}} K(u, \epsilon) \tag{25}$$

$$\min_{u \in \mathcal{U}^{\text{strict}}} K(u, \epsilon) \tag{26}$$

Under the assumptions of Theorem 2, problems (16)–(18), (25) and (26) are equivalent in the sense that they have the same optimal values and the same set of minimizers. As a consequence, if u^* is an optimal control and H the Hamiltonian, then $\frac{\partial H}{\partial u} = 0$ at $u = u^*$.

Proof
We have

$$\mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}} \subset \mathcal{U} \cap \mathcal{X}^{\text{strict}} \subset \mathcal{U} \cap \mathcal{X} \subset \mathcal{U}$$

$$\mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}} \subset \mathcal{U}^{\text{strict}} \subset \mathcal{U}$$

DOI: 10.1002/oca

which corresponds to inequalities on the minimum values in decreasing values. On the other hand, with suitable penalties, any solution of (18) must belong to $\mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$. This shows the equality of the minima and of their set of minimizers. In particular, any minimizer is a solution of (26), and hence, the ordinary calculus of variations shows that it satisfies $\frac{\partial H}{\partial u} = 0$. □
This result proves that the proposed penalties can be used to generate a sequence of interior solutions. This is discussed further by the following corollary

Corollary 4
Under the assumptions of Theorem 2, assume further that for $\epsilon > 0$ small enough, Problem (18) has at least a solution $u^*(\epsilon)$, and that Assumption 3 holds, then

A $\lim_{\epsilon \downarrow 0} J(u^*(\epsilon)) = \inf_{u \in \mathcal{U} \cap \mathcal{X}} J(u)$
B $\lim_{\epsilon \downarrow 0} \epsilon \{ P_g(u^*(\epsilon)) + P_u(u^*(\epsilon)) \} = 0$

Proof
We are going to apply the classic Theorem 5 recalled in Appendix with $S = \mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}} \subset R = \mathcal{U} \cap \mathcal{X}$. We first observe that if a control u belongs to $\mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$, both penalties $P_g(u)$ and $P_u(u)$ are bounded. Because of Assumption 3, we have a minimizing sequence for J within $\mathcal{U} \cap \mathcal{X}^{\text{strict}}$. From Proposition 7, $\mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$ is dense in $\mathcal{U} \cap \mathcal{X}^{\text{strict}}$ when using the L^∞ norm. Therefore, there exists a minimizing sequence in $\mathcal{U}^{\text{strict}} \cap \mathcal{X}^{\text{strict}}$ because of the continuity of J (Proposition 13). Hence, we can apply Theorem 5 with the previous choice of R and S, that is, with $u^*(\epsilon)$ being a solution of problem (25). From Corollary 3, the set of solutions of (25) is equal to the set of the solutions of problem (18). This gives the conclusion. □

5. REPRESENTATION OF THE CONTROL CONSTRAINT BY A CHANGE OF VARIABLES

At this stage, the fact that the optimum is necessarily interior has been established (Theorem 2). From a numerical implementation perspective, this leads to interesting possibilities. For instance, if $\mathcal{C}$ is the interval $[-1, +1]$, the change of variable $v = \text{atanh}(u)$ is a one-to-one mapping from $(0, 1)$ into $\mathbb{R}$. This change of variable allows to complete remove the constraints from the problem formulation and to work without any bounds on the unknowns. This is particularly convenient for implementation. This approach has been developed in [37–41]. The inverse of the aforementioned change of variable is called *a saturation function*. In all these references, the constraint sets have always been a cartesian product of intervals. We generalize the procedure here to our more general settings of cartesian products of convex sets, and apply it to define an equivalent unconstrained optimal control problem, well-suited for numerical implementation.

5.1. Saturation functions for convex sets

To generalize saturation functions to smooth convex sets, it is handy to first consider the mapping $\psi : \mathbb{R}^m \mapsto B^m_{\|.\|}(0, 1)$ such that

$$\psi(v) \triangleq \begin{cases} 0 \text{ if } v = 0 \\ \tanh(\|v\|) \dfrac{v}{\|v\|} \text{ otherwise} \end{cases} \tag{27}$$

where $B^m_{\|.\|}(0, 1)$ is the open unit ball of $\mathbb{R}^m$ for the norm $\|.\|$, for example, the Euclidean norm. This mapping is a homeomorphism[††] and is differentiable on $\mathbb{R}^m \setminus \{0\}$. The next proposition formulates the generalization of saturation functions[‡‡].

[††] whose inverse is $\psi^{-1}(u) \triangleq \text{atanh}(\| u \|) \frac{u}{\|u\|}$
[‡‡] it is indeed a generalization, as we recover the usual saturation function from [42] when the convex is an interval of $\mathbb{R}$.

DOI: 10.1002/oca

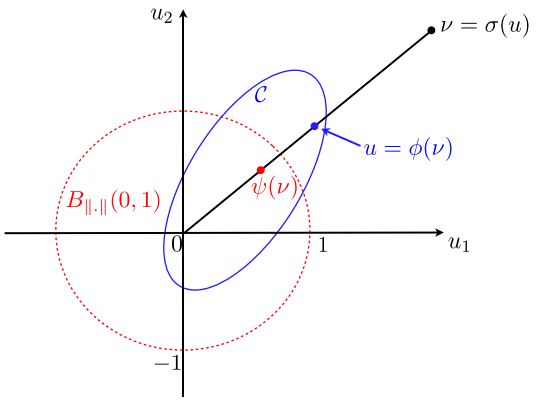

Figure 1. Example of a generalized saturation function ϕ. If u belongs to int($\mathcal{C}$), then there exists $\nu \in \mathbb{R}^m$ such that $u = \phi(\nu)$ where ϕ is defined in (28). The correspondence is one-to-one.

Proposition 10 (Generalized saturation functions)
The function $\phi : \mathbb{R}^m \mapsto \text{int}(\mathcal{C})$ defined by

$$\phi(\nu) \triangleq \begin{cases} 0 \text{ if } \nu = 0 \\ \dfrac{\tanh^2(\|\nu\|)}{G_{\mathcal{C}}(\psi(\nu))} \dfrac{\nu}{\|\nu\|} \end{cases} \quad \text{otherwise} \tag{28}$$

where ψ is defined in (27), is a homeomorphism. Moreover, this mapping is differentiable on int($\mathcal{C}$)$\setminus$ $\{0\}$. Its inverse is the function $\sigma : \text{int}(\mathcal{C}) \mapsto \mathbb{R}^m$ defined by

$$\sigma(u) \triangleq \begin{cases} 0 \text{ if } u = 0 \\ \text{atanh}(G_{\mathcal{C}}(u)) \dfrac{u}{\|u\|} \end{cases} \text{ otherwise} \tag{29}$$

Proof
See Appendix G. Notations are illustrated in Figure 1. □

Proposition 10 implies that if u belongs to int($\mathcal{C}$), then there exists $\nu \in \mathbb{R}^m$ such that $u = \phi(\nu)$ and the correspondence is one-to-one.

5.2. *Correspondence of control sets*

Let

$$\mathcal{L} \triangleq \prod_{i=1}^{p} L^{\infty}([0,T], \mathbb{R}^{m_i})$$

For each convex $\mathcal{C}_i$, define with (28) and (29) the related functions ϕ_i (28) and $\sigma_i = \phi_i^{-1}$ defined in (29).

Proposition 11
We have

$$\mathcal{L} = \{(\sigma_1(u_1), \ldots, \sigma_p(u_p)), u \in \mathcal{U}^{\text{strict}}\}$$

and

$$\mathcal{U}^{\text{strict}} = \{(\phi_1(\nu_1), \ldots, \phi_p(\nu_p)), \nu \in \mathcal{L}\}$$

DOI: 10.1002/oca

Proof
The proof is a straightforward consequence of the fact that ϕ_i and σ_i are one-to-one mappings, and that the elements u of $\mathcal{U}^{\text{strict}}$ are characterized by the existence of some $\alpha > 0$ such that $G_{\mathcal{C}_i}(u_i(t)) \leqslant 1 - \alpha$ for all i and almost every t. □

5.3. Penalized problem (final version)

Finally, we define a last penalized optimal control problem

$$\min_{\nu \in \mathcal{L}} \left[P(\nu, \epsilon) = \int_0^T \ell\left(x^{\phi(\nu)}, \phi(\nu)\right) + \epsilon \left[\sum_{i \leqslant q} \gamma_g \circ g_i\left(x^{\phi(\nu)}\right) + \sum_{i \leqslant p} \gamma_u \circ G_{\mathcal{C}_i} \circ \phi_i(\nu_i) \right] dt \right] \tag{30}$$

where the penalty functions are given by (9)–(23), and make the following assumption

Assumption 5
The penalized problem (18) has at least one optimal solution.

5.4. Third main result

We have the following equivalence theorem between problems (18) and (30), which is our third main result

Theorem 3 (Equivalence of the new representation of variables)
Under the assumptions of Theorem 2 and (existence) Assumption 5, for any $\epsilon > 0$ POCP (18) and POCP (30) are equivalent in the sense that

$$\arg\min_{u \in \mathcal{U}} K(u, \epsilon) = \phi\left(\arg\min_{\nu \in \mathcal{L}} P(\nu, \epsilon)\right)$$

where $\phi(\nu)$ denotes $(\phi_1(\nu_1), \ldots, \phi_p(\nu_p))$.

As a consequence, one can solve the POCP (18) which is constrained by $u \in \mathcal{U}$ by solving instead the unconstrained POCP (30), and then apply the operator ϕ to obtain an optimal solution for (18).

Proof
Let us consider $u^* \in \mathcal{U}$ a minimizer of $K(., \epsilon)$, which exists by Assumption 5. We have

$$K\left(u^*, \epsilon\right) \leqslant K(u, \epsilon), \quad \forall u \in \mathcal{U}^{\text{strict}}$$

Define $\nu^* = \sigma(u^*)$ and $\nu = \sigma(u)$. This definition is valid because both controls belong to $\mathcal{U}^{\text{strict}}$. Then, $u^* = \phi(\nu^*)$ and $u = \phi(\nu)$. Therefore,

$$K(\phi(\nu^*), \epsilon) \leqslant K(\phi(\nu), \epsilon)$$

or, equivalently,

$$P(\nu^*, \epsilon) \leqslant P(\nu, \epsilon)$$

From Proposition 11, we know that $\sigma(u)$ spans $\mathcal{L}$ when u spans $\mathcal{U}^{\text{strict}}$. Therefore, ν^* is optimal for POCP (30); this proves, incidentally, the existence of a solution to POCP (30). Because $u^* = \phi(\nu^*)$, this proves

$$\arg\min_{u \in \mathcal{U}} K(u, \epsilon) \subset \phi\left(\arg\min_{\nu \in \mathcal{L}} P(\nu, \epsilon)\right)$$

DOI: 10.1002/oca

Now, let us consider $v^* \in \mathcal{L}$ a minimizer of $P(.,\epsilon)$ (which has been proven to exist). From Proposition 11, $u^* \triangleq \phi(v^*) \in \mathcal{U}^{\text{strict}}$. We have

$$P(v^*,\epsilon) \leqslant P(v,\epsilon), \quad \forall v \in \mathcal{L}$$

From Proposition 11, this implies

$$P(\sigma(u^*),\epsilon) \leqslant P(\sigma(u),\epsilon), \quad \forall u \in \mathcal{U}^{\text{strict}}$$

that is,

$$K(u^*,\epsilon) \leqslant K(u,\epsilon), \quad \forall u \in \mathcal{U}^{\text{strict}} \tag{31}$$

From Theorem 2, we know that any optimal control for $K(u,\epsilon), u \in \mathcal{U}$ must belong to $\mathcal{U}^{\text{strict}}$. Therefore, we can substitute one of these optimal controls in place of u in (31); which proves that $u^* = \sigma(v^*)$ is optimal for POCP (18). Therefore,

$$\arg\min_{u\in\mathcal{U}} K(u,\epsilon) \supset \phi\left(\arg\min_{v\in\mathcal{L}} P(v,\epsilon)\right)$$

Finally, we have

$$\arg\min_{u\in\mathcal{U}} K(u,\epsilon) = \phi\left(\arg\min_{v\in\mathcal{L}} P(v,\epsilon)\right)$$

This concludes the proof. □

Corollary 5
Define

$$\bar{J}(v) = \int_0^T \ell\left(x^{\phi(v)}(t), \phi(v(t))\right) dt$$
$$\Gamma(v) = \int_0^T \sum_{i\leqslant q} \gamma_g \circ g\left(x^{\phi(v)}(t)\right) + \sum_{i\leqslant p} \gamma_u \circ G_{C_i} \circ \phi_i(v_i(t))dt$$

Assume that, for ϵ small enough, problem (30) has at least a solution $v^*(\epsilon)$. Then, under the assumptions of Theorem 2 and Assumption 3, the following holds

A $\lim_{\epsilon\downarrow 0} \bar{J}\,(v^*(\epsilon)) = \inf_{u\in\mathcal{U}\cap\mathcal{X}} J(u)$
B $\lim_{\epsilon\downarrow 0} \epsilon\Gamma(v^*(\epsilon)) = 0$

The corollary is a direct consequence of the equivalence Theorem 3 and of Corollary 4.

Remark 1
A theory similar to the theory developed in Section 5 can be constructed by replacing the tanh function by any diffeomorphism D between $\mathbb{R}$ and $(-1,+1)$ which satisfies $D(-x) = -D(x)$.

6. CONVERGENCE OF THE INTERIOR POINT METHODS

Corollaries 2, 4 and 5 establish the convergence of the optimal cost of each penalized problem to the optimal cost of the original problem (1). We can now go one step further and establish, under some extra assumptions, convergence of the control and of the state.

Assumption 6
$\mathcal{U}\cap\mathcal{X}$ is convex and the cost functional J of COCP (1) satisfies the strong convexity property

$$D \parallel u - v \parallel_{L^2}^2 \leqslant J(u) + J(v) - 2J\left(\frac{u+v}{2}\right) \ \forall u, v \in \mathcal{U}\cap\mathcal{X} \tag{32}$$

for some $D > 0$. Moreover, problem (1) has at least one optimal control u^*.

DOI: 10.1002/oca

Observe that $\mathcal{U} \cap \mathcal{X}$ is convex if the g_i are convex and the dynamics are linear.

Theorem 4 (Convergence of the optimal state and control)
Under Assumption 6, the optimal control u^* of (1) is unique. Further, under the assumptions of Corollaries 2 and 4 hold and if, for $\epsilon > 0$ small enough, (10) (respectively (18)) has at least one optimal solution $u^*(\epsilon)$, then

$$\lim_{\epsilon \downarrow 0} \| u^*(\epsilon) - u^* \|_{L^2} = 0 \tag{33}$$

$$\lim_{\epsilon \downarrow 0} \| x^{u^*(\epsilon)} - x^{u^*} \|_{L^\infty} = 0 \tag{34}$$

Proof
Let $v \in \mathcal{U} \cap \mathcal{X}$; then $\frac{u^*+v}{2} \in \mathcal{U} \cap \mathcal{X}$. As a consequence, $J(u^*) \leqslant J\left(\frac{u^*+v}{2}\right)$. From (32), we derive

$$D \| u^* - v \|_{L^2}^2 \leqslant J(u^*) + J(v) - 2J(u^*) = J(v) - J(u^*) \ \forall v \in \mathcal{U} \cap \mathcal{X} \tag{35}$$

Equation (35) proves that the optimum u^* of problem (1) is unique. Moreover, whether it be for problem (10) or (18)), any optimum $u^*(\epsilon)$ belongs to $\mathcal{U} \cap \mathcal{X}$. As a consequence we have

$$D \| u^* - u^*(\epsilon) \|_{L^2}^2 \leqslant J(u^*(\epsilon)) - J(u^*) \tag{36}$$

Using Corollary 2 (respectively Corollary 4), which proves the convergence of the costs, (36) proves the convergence (33). From Proposition 13, we derive that (34) holds. □

Corollary 6
If the conditions of Theorem 4 hold, then the solution $v^*(\epsilon)$ of the unconstrained problem (30) satisfies

$$\lim_{\epsilon \downarrow 0} \| \phi(v^*(\epsilon)) - u^* \|_{L^2} = 0$$

$$\lim_{\epsilon \downarrow 0} \| x^{\phi(v^*(\epsilon))} - x^{u^*} \|_{L^\infty} = 0$$

This result is a direct consequence of the previous convergence Theorem 4 and of the equivalence Theorem 3.

7. RESOLUTION ALGORITHMS AND 'COOKBOOK'

Thanks to the equivalence results proven earlier, we know that we 'simply' have to solve Problem (30). This will generate a solution $u^*(\epsilon)$, and one will gradually reduce ϵ to approach [§§] the solution of COCP (1).

This section explains how this can be done, what are the main difficulties that can be encountered, and what are the theoretical flexibilities one can use to simplify implementation.

[§§] Some remarks can be made on this generated sequence. Because, one has no estimate on the difference $J(u^*(\epsilon)) - J(u^*)$, it is necessary to explore the values of the cost and of the control as we solve the penalized problems for a decreasing sequence of ϵ. Most solving algorithms used to minimize the penalized costs require some initial value. It makes sense to use the output of the solving algorithm for the previous ϵ as an initial parameter for the current ϵ. If the two values of ϵ are close, it is a safe bet to say that the initial guess will be a good one. On the other hand, having close values for the ϵ increases the number of penalized problems to be solved for a given range of ϵ. Last but not the least, it is generally required to provide an trajectory at the beginning of the iterative procedure that at least belongs to X_{ad} (satisfying the differential system may not be necessary at the initialization).

DOI: 10.1002/oca

In theory, one can solve (30) using any of the classic tools of optimal control (i) dynamic programming (which we leave away from the discussion due to its relatively poor tractability in dimensions superior to 4–5), (ii) direct methods a.k.a. collocation methods, (iii) indirect methods.

In direct methods, the integrand of the cost is considered as the dynamics of an augmented state and the goal is to minimize the final value of this extra state. A collocation method is used to discretize the differential system in time by replacing the state and control by finite elements. The original problem is approximatively solved by solving this finite-dimensional optimization problem under the constraint that the dynamics are satisfied at the mesh points. A more detailed discussion on the use of direct methods in this framework with numerical illustrations can be found in [43][chapter I]. These methods are relatively fast and robust to initialization, but may reveal heavy when a high level of accuracy is desired.

In indirect methods, the conditions of the Pontryagin Minimum Principle [44] are exploited. Here, the minimizer of the Hamiltonian of (30) must satisfy $\frac{\partial H_\epsilon}{\partial v} = 0$, where v is the control variable after the change of variable. The use of the function ϕ may complicate the solving of the equation. In fact, it is not the case as is now discussed.

The Hamiltonian is

$$H_\epsilon(x, v, p) = \ell(x, \phi(v)) + \epsilon \left[\sum_{i \leqslant q} \gamma_g \circ g_i(x) + \sum_{i \leqslant p} \gamma_u \circ G_{\mathcal{C}_i} \circ \phi_i(v_i) \right] + p^T f(x, \phi(v)) \quad (37)$$

where ϕ is defined by (28). If we compare it to the Hamiltonian H_u of the original problem (1), we obtain $H_\epsilon(x, v, p) = H_u(x, \phi(v), p)$. From Theorem 2, we know that the optimal control v^* satisfies

$$\frac{\partial H_\epsilon}{\partial v}(x, v^*, p) = \frac{\partial H_u}{\partial u}\left(x, \phi(v^*), p\right) \frac{d\phi}{dv}(v^*) = 0 \quad (38)$$

Because ϕ is invertible, $\frac{d\phi}{dv}$ has full rank, and (38) is equivalent to

$$\frac{\partial H_u}{\partial u}\left(x, \phi(v^*), p\right) = 0$$

or specifically,

$$\frac{\partial \ell}{\partial u}(x, \phi(v^*)) + \epsilon \sum_{i \leqslant p} \gamma_u' \left(G_{\mathcal{C}_i}(\phi_i(v_i^*))\right) \frac{dG_{\mathcal{C}_i}}{du_i}\left(\phi_i(v_i^*)\right) + p^T \frac{\partial f}{\partial u}(\phi(v^*)) = 0 \quad (39)$$

We shall assume that (39) has a unique solution $v^*(x, p)$. Then, the two point boundary value problem (TBVP) for (30) has the dynamics

$$\frac{dx}{dt} = f(x, \phi(v^*)) \ , \ x(0) = x_0 \quad (40)$$

$$\frac{dp}{dt} = -\left(\frac{\partial \ell}{\partial x}\left(x, \phi(v^*)\right)\right)^T - \epsilon \sum_{i \leqslant q} \gamma_g'(g_i(x)) \left(\frac{dg_i}{dx}\right)^T - \left(\frac{\partial f}{\partial x}\left(x, \phi(v^*)\right)\right)^T p \ , \ p(T) = 0 \quad (41)$$

A point worth mentioning is that the stationarity (39) has to be solved simultaneously to the two differential (40) and (41). Therefore, it is of the highest interest that this equation be easy to solve; ideally, we would have a closed form for its solution. Our only degree of freedom here lies in the choice of γ_u, or, rather, of γ_u'. The thing to attempt is to find a function γ_u' such that

(a) one can prove that $\gamma_u \triangleq \int_0^{u \geqslant 0} \gamma_u'(v)dv$ satisfies the conditions of Theorem 2; this function does not need to be actually computed. Indeed, condition (22) of Proposition 9 only bears of the derivative γ_u'. By the integral construction previously mentioned, γ_u is continuously differentiable, strictly convex and nondecreasing if γ_u' is continuous, increasing and nonnegative. The only point that remains to check is that the indefinite integral tends to $+\infty$ when u tends to 1.

DOI: 10.1002/oca

(b) Equation (39) has a unique solution ν^* with $\phi(\nu^*)$ being easily computable. The existence of ν^* is essential because it guarantees that $u = \phi(\nu^*)$ belongs to the interior of $\mathcal{C}$. An attractive possibility is to design γ'_u so that the unique ν^* solution of (39) can be expressed in closed form with respect to ϵ, x and p.

Observe that (40) and (41) do not require any knowledge of ν^* but rather of $\phi(\nu^*)$, and that γ_u, nor any of its derivatives, does not appear in (40) and (41). Therefore, satisfying conditions a) and b) is an efficient way of expliciting all the terms in the TBVP (40) and (41). Such a method is implemented and explicited in the numerical example of Section 8.

Then, a solving algorithm can be described as follows

- Step 1: Initialize the continuous functions $x^{\phi(\nu)}(t)$ and $p(t)$ such that the initial values satisfy $g_i(x^{\phi(\nu)}(t)) < 0$ for all $t \in [0, T]$. Define a decreasing sequence $\epsilon_i > 0$, $i = 0, \ldots, N$ and set $\epsilon = \epsilon_0$. Note that $x^{\phi(\nu)}(t)$ and $p(t)$ need not satisfy any differential equation at this stage, even if it is better if they do.
- Step 2: Let $H_\epsilon(x, \nu, p)$ the Hamiltonian (37). Solve the $2n$ differential equations

$$\frac{dx}{dt} = f\left(x, \phi\left(\nu_\epsilon^*\right)\right) \ , \ x(0) = x_0$$
$$\frac{dp}{dt} = -\frac{\partial H_\epsilon}{\partial x}\left(x, \phi\left(\nu_\epsilon^*\right), p\right) \ , \ p(T) = 0$$

  Here ν_ϵ^* is the solution of the stationarity equation

$$\frac{\partial H_\epsilon}{\partial \nu}\left(x, \nu_\epsilon^*, p\right) = 0$$

  This can be done in Matlab by using the routines bvp5c or bvp4c (see [45]).
- Step 3: If $\epsilon = \epsilon_N$, stop. Else decrease ϵ, initialize $x^{\phi(\nu)}(t)$ and $p(t)$ with the solutions found at Step 2 and restart at Step 2.

This algorithm is detailed for Matlab in Appendix H and illustrated on a numerical example in Section 8. Other implementations, using, for example, [46] are also possible.

7.1. How-to guide 'cookbook'

Below, we give a synthetic view for the practitioners of the several steps needed to apply the presented method. It will be illustrated on a particular example in the next section.

	COOKBOOK
Input:	Consider a COCP (2)–(3) under the constraints $u \in \mathcal{U} \cap \mathcal{X}$ (3)–(4)
Step 1:	Choose $n_g > 1$ and define $\gamma_g = \begin{cases} (-x)^{-n_g > 1} \text{ for } x < 0 \\ 0 \text{ for } x \geqslant 0 \end{cases}$
Step 2:	Determine the gauge function $G_{\mathcal{C}_i}$ of each $\mathcal{C}_i$ defined in (13) and constitute the generalized saturation functions ϕ according to (27)–(28). For example, if u must belong to $[a, b]$, then pick: $G_{[-1,1]} = \lvert . \rvert$, $\phi(\nu) = \tanh\left(\frac{2\nu}{b-a}\right)$ and $u = \frac{b-a}{2}(\phi(\nu) + 1) + a$.
Step 3:	Constitute the Hamiltonian H_ϵ (37) and, therein, pick γ'_u such that the condition $\frac{\partial H_\epsilon}{\partial \nu} = 0$ is easy to solve with respect to the variable ν.
(Step 3':)	For mathematical consistency, check that $\lim_{u \uparrow 1} \gamma'_u(u) = +\infty$ and that $\gamma'(0) = 0$, $\lim_{u \uparrow 1} \gamma_u(u) = +\infty$.
Step 4:	Solve the TBVP giving the solution of the POCP (30) for decreasing values of ϵ.

DOI: 10.1002/oca

8. NUMERICAL ILLUSTRATION: GODDARD'S PROBLEM

The classical Goddard problem presented in 1919 [34] concerns maximizing the final altitude of a rocket launched in vertical direction. The problem has become a benchmark example in optimal control due to a characteristic singular arc behavior in connection with a relatively simple model structure, which makes the Goddard rocket an ideal object of study, see [38, 47].

8.1. Model equations and optimal control problem

8.1.1. Model equations. The equations of motion of the rocket are given by the ordinary differential equations

$$\dot{h} = v \tag{42}$$

$$\dot{v} = \frac{u - D(h, v)}{m} - \frac{1}{h^2} \tag{43}$$

$$\dot{m} = -\frac{u}{c} \tag{44}$$

with the altitude h from the center of Earth, the velocity v, and the mass m as the mass of the rocket. The states h, v, m, the thrust u as the input of the system, and the time t are commonly normalized and dimension free. The drag function $D(h, v)$ from the velocity dynamics is given by

$$D(h, v) = q(h, v)\frac{C_D A}{m_0 g} \tag{45}$$

as a function of the Earth's gravitational acceleration g and the dynamic pressure

$$q(h, v) = \frac{1}{2}\rho_0 v^2 e^{\beta(1-h)} \tag{46}$$

that depends on the altitude h and the velocity v. The constants in the model equations are

$$C_D \text{ drag coefficient}, \rho_0 \text{ air density at sea level},$$
$$A \text{ reference area}, \beta \text{ density decay rate},$$
$$m_0 \text{ initial mass}, c \text{ exhaust velocity}$$

The following values are taken from [38, 48]:

$$\beta = 500,\ c = 0.5,\ \frac{\rho_0 C_D A}{m_0 g} = 620$$

8.1.2. Optimal control problem. The optimal control problem is the following:

$$\min_u -h(T)$$

under the dynamics (42)–(44) and the following constraints

$$u(t) \in [0; 3.5] \text{ a.e. } t \in [0, T]$$
$$q(h(t), v(t)) \leqslant 10\ \forall t \in [0, T]$$

where the final time T is a free parameter. Let us reformulate the problem as a fixed horizon optimal control problem. To do so we make the following change of variable $\tau = \frac{t}{T}$ and we have the following augmented dynamics:

$$\dot{h} = T v$$
$$\dot{v} = T\left[\frac{u - D(h, v)}{m} - \frac{1}{h^2}\right]$$
$$\dot{m} = -T\frac{u}{c}$$
$$\dot{T} = 0$$

DOI: 10.1002/oca

and the optimal control problem becomes:

$$\min_u -\int_0^1 T v dt$$

under the aforementioned augmented dynamics and the following constraints:

$$\begin{aligned} u(t) &\in [0, 3.5] \text{ a.e. } t \in [0, 1] \\ q(h(t), v(t)) &\leqslant 10 \ \forall t \in [0, 1] \end{aligned}$$

Observe that u is constrained to belong to a convex set having the origin at its boundary. We now define $\phi(\nu)$ by

$$\phi(\nu) = \tanh\left(\frac{2\nu}{3.5}\right)$$

and

$$u(\nu) = \frac{3.5}{2}(\phi(\nu) + 1)$$

which is a one to one mapping from $\mathbb{R}$ into the interior $(0, 3.5)$ of the admissible control set. We define our control penalty function by

$$\gamma_u \circ G_{[-1,1]}(\phi(\nu)) = \gamma_u \circ G_{[-1,1]}\left(\tanh\left(\frac{2\nu}{3.5}\right)\right)$$

The Hamiltonian of the unconstrained POCP corresponding to this problem is the following:

$$\begin{aligned} H_\epsilon(h, v, m, T, \nu, p_h, p_v, p_m, p_T) \triangleq\ & T\left[-v + p_h v + p_v\left[\frac{u(\nu) - D(h, v)}{m} - \frac{1}{h^2}\right] - p_m\frac{u(\nu)}{c}\right] \\ & + \epsilon\left[\gamma_u \circ G_{[-1,1]} \circ \phi(\nu) + \gamma_g(q(h, v) - 10)\right] \end{aligned}$$

The TBVP consists in solving the followings ODEs

$$\begin{array}{cccc} \dot{h} = Tv & \dot{v} = T\left[\frac{u(\nu^*) - D(h,v)}{m} - \frac{1}{h^2}\right] & \dot{m} = -T\frac{u(\nu^*)}{c} & \dot{T} = 0 \\ \dot{p}_h = -\frac{\partial H_\epsilon}{\partial h} & \dot{p}_v = -\frac{\partial H_\epsilon}{\partial v} & \dot{p}_m = -\frac{\partial H_\epsilon}{\partial m} & \dot{p}_T = -\frac{\partial H_\epsilon}{\partial T} \\ h(0) = 1 & v(0) = 0 & m(0) = 1 & m(1) = 0.6 \\ p_h(1) = 0 & p_v(1) = 0 & p_T(0) = 0 & p_T(1) = 0 \end{array}$$

where ν^* is solution of $\frac{\partial H_\epsilon}{\partial \nu} = 0$. Observe that $G_{[-1,1]}(u) = |u|$. Thus

$$\gamma_u \circ G_{[-1,1]}(\phi(\nu)) = \gamma_u(|\phi(\nu)|)$$

where $\gamma_u : [0, 1) \mapsto \mathbb{R}^+$. Therefore, determining γ_u is equivalent to determining $\gamma_u(\phi(\nu))$ where $\gamma_u : (-1, 1) \mapsto \mathbb{R}^+$ is a symmetric function. Taking this parameterization makes the control penalization differentiable with respect to ν and we have

$$\frac{\partial H_\epsilon}{\partial \nu} = T\left[p_v\frac{u'(\nu)}{m} - p_m\frac{u'(\nu)}{c}\right] + \epsilon\gamma_u'(\phi(\nu))\phi'(\nu)$$

that is

$$\frac{3.5}{2}\left[p_v\frac{\phi'(\nu)}{m} - p_m\frac{\phi'(\nu)}{c}\right] + \epsilon\gamma_u'(\phi(\nu))\phi'(\nu)$$

DOI: 10.1002/oca

Therefore, ν^* is solution of

$$T\left[\frac{p_v}{m}-\frac{p_m}{c}\right]+\epsilon\frac{2}{3.5}\gamma_u'(\phi(\nu))=0 \tag{47}$$

From Lemma 9 and the symmetry of γ_u we know that the solution are interior if γ_u' is a bijective increasing mapping from $(-1,1)$ to $\mathbb{R}$. Moreover $\phi(\nu)$ being a increasing bijective mapping from $\mathbb{R}$ to (-1,1), one can take the following parameterization of the control penalty:

$$\frac{2}{3.5}\gamma_u'(\phi(\nu))\triangleq\sinh(\nu) \tag{48}$$

which is a bijective increasing mapping from $\mathbb{R}$ to $\mathbb{R}$. Specifically, we have

$$\gamma_u'\left(\tanh\left(\frac{2\nu}{3.5}\right)\right)=\frac{3.5}{2}\sinh(\nu) \tag{49}$$

But the inverse function of $y=\tanh(x)$ for $x\in(-1,+1)$ is $x=\frac{1}{2}\log\left(\frac{1+y}{1-y}\right)$. Hence, (49) is equivalent to

$$\gamma_u'(w)=\frac{3.5}{4}\left(\left(\frac{w+1}{w-1}\right)^{\frac{3.5}{4}}-\left(\frac{w-1}{w+1}\right)^{\frac{3.5}{4}}\right)\ ,\ \gamma(0)=0\ ,\ w\in(-1,+1) \tag{50}$$

which is a simple quadrature. It can be solved numerically and we can check that γ_u has the desired properties that make sure that the optimal control u must be interior, in particular, when x tends to 1, $\log(\gamma_u(x))/(-\log(x-1))$ increases; it reaches value 0.0762 at $x=1-10^{-14}$, which means that $\gamma_u(x)\approx\frac{1}{(1-x)^{0.0762}}$ at this point.. We see that in this example we have been able to implement the points a) and b) described in Section 7 when studying the solving of the stationarity (39).

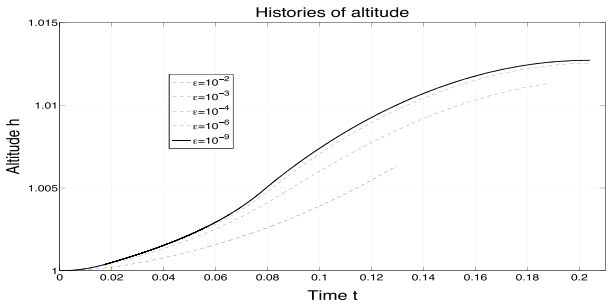

Figure 2. Histories of optimal altitude for decreasing values of ϵ.

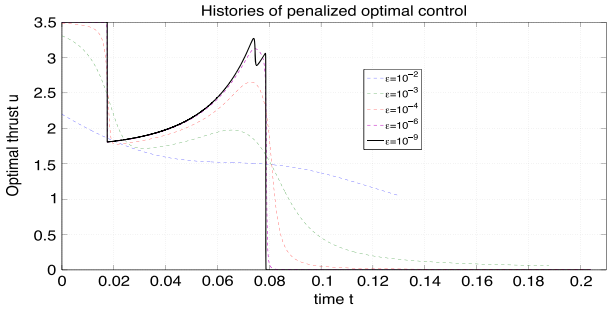

Figure 3. Histories of optimal thrust for decreasing values of ϵ.

DOI: 10.1002/oca

Equation (47) then has an analytical solution:

$$v^* = \sinh^{-1}\left(-\frac{T}{\epsilon}\left(\frac{p_v}{m} - \frac{p_m}{c}\right)\right)$$

The problem solving is initialized with constant values of the variables as follows:

$$\begin{array}{llll} h(t) = 1 & v(t) = 0.2 & m(t) = 1 & T = 0.5 \\ p_h(t) = 0 & p_v(t) = 1 & p_m(t) = 0 & p_T(t) = 0 \end{array}$$

The sequence (ϵ_n) is initialized with $\epsilon_0 = 10^{-2}$, the parameter n_g from 9 is set at $n_g = 1.1$. To solve the problem we use the MATLAB software bvp5c. The source for this example is available at [49]. Other numerical examples, notably with a two dimensional convex control set, can be found in [43].

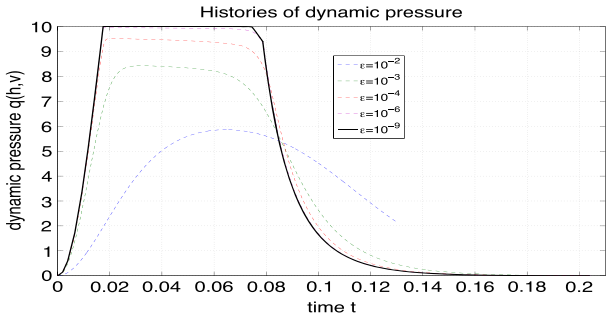

Figure 4. Histories of optimal dynamic pressure for decreasing values of ϵ.

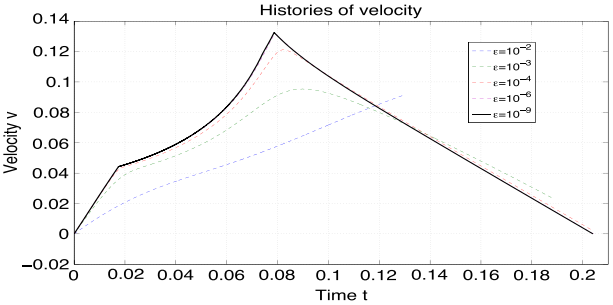

Figure 5. Histories of optimal velocity for decreasing values of ϵ.

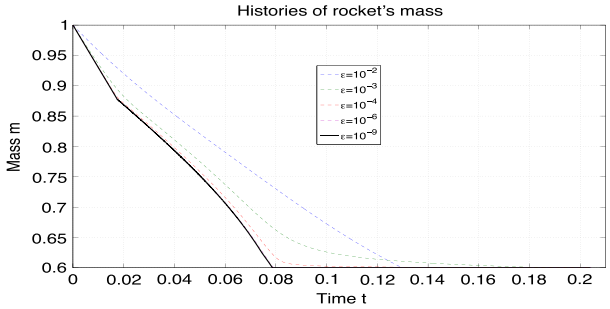

Figure 6. Histories of optimal mass for decreasing values of ϵ.

DOI: 10.1002/oca

8.2. Results

From Figures 2 to 6 histories of state variables, thrust and state constraints are given for decreasing values of the parameter ϵ. One can see that these solutions are similar to those given in [38]. Moreover, the optimal final time and the optimal value of the criterion are the following:

$$T = 0.20405546 \quad ; \quad h(T) = 1.01271747$$

9. CONCLUSION

We have constructively exhibited three classes of penalized optimal control problems whose optimal costs converge to the optimal control of a given optimal control problem with state and control constraints. Under classical convexity assumptions, we also have convergence of the control and state toward the optimal control and state of the original problem.

A notable feature of all of these penalized problems is that the state penalty is sufficient to guarantee that the state constraints are strictly satisfied; thus they do not need to be specified in the penalized problem. By adding a control penalty, we ensure that the second problem has an optimal control which is in the interior of its constraint set; therefore, it must satisfy a stationarity condition on the Hamiltonian. By making the most of this interiority property, we can finally operate a invertible change of variable on the control which defines an unconstrained penalized optimal control problem. Solving this unconstrained problem and recovering the original control by the inverse transform yields convergence of the cost and, under suitable assumptions, of the control.

Finally, we have observed on all of our numerical examples that the co-states exhibit a convergent behavior; moreover, singularities appear gradually at the points where the theory predicts they should appear, typically, entry and exit points of singular arcs. Proving the convergence of the co-states toward the co-states of the original problem is a challenging task. If it was achieved, these methods would provide an automatic method for computing the co-states and the location of their singularities in the original problem.

APPENDIX A: RECALL ON INTERIOR POINT METHODS

A.1. A result by Fiacco and McCormick

We recall a seminal result originally published in [50, 51] because of its general interest in the presented context and the simplicity of its proof. The interior methods are based on this result. We consider an abstract control set E with two subsets R and S, on which a functional J is defined, with values in $\mathbb{R}$. We wish to minimize J over the set R. This task may prove to be too difficult as-is, and instead, one can choose to solve a sequence of (easier) problems indexed by a positive parameter ϵ

$$\min_{u \in S} J(u) + \epsilon G(u) \tag{51}$$

where G is defined on S (defined later-on) with nonnegative values in $\mathbb{R}$.

Theorem 5 (Fiacco, McCormick)
We assume that problem (51) admits at least a minimizer $u^*(\epsilon)$ that belongs to $R \cap S$. Let $v_0 = \inf_{u \in R} J(u)$. We assume that v_0 is finite and that there exists a minimizing sequence in S for which J converges to v_0, that is, for any $\omega > 0$ there exists $s(\omega) \in S$ such that $v_0 \leqslant J(s(\omega)) \leqslant v_0 + \omega$. Then

A $\lim_{\epsilon \downarrow 0} J(u^*(\epsilon)) = v_0 = \inf_{u \in R} J(u)$
B $\lim_{\epsilon \downarrow 0} \epsilon G(u^*(\epsilon)) = 0$

Proof: (inspired from [50, 51]*)*
Let $\omega > 0$ an arbitrary small number , and define $\epsilon_1 = \omega / G(s(\omega))$, (if $G(s(\omega)) = 0$, we take

DOI: 10.1002/oca

$\epsilon_1 = 1$). Then we have

$$J(s(\omega)) + \epsilon_1 G(s(\omega)) \leqslant v_0 + 2\omega$$

Consider now any ϵ such that $0 < \epsilon \leqslant \epsilon_1$. We have the following inequalities

$$\begin{aligned} v_0 &\leqslant J(u^*(\epsilon)) + \epsilon G(u^*(\epsilon)) \\ &= \min_{u \in S} J(u) + \epsilon G(u) \\ &\leqslant J(s(\omega)) + \epsilon G(s(\omega)) \\ &\leqslant J(s(\omega)) + \epsilon_1 G(s(\omega)) \\ &\leqslant v_0 + 2\omega \end{aligned}$$

This proves that $\lim_{\epsilon \downarrow 0} J(u^*(\epsilon)) + \epsilon G(u^*(\epsilon)) = v_0$. Because G is nonnegative and $J(u^*(\epsilon))$ is lower bounded by v_0, this proves the limits A and B in the theorem. □

Monotonicity For the sake of completeness, we mention another result that can be found in [50]; a detailed proof of it can be found in [40].

Proposition 12
$J(u^*(\epsilon))$ decreases when ϵ decreases, and $G(u^*(\epsilon))$ increases when ϵ decreases.

A.2. Interior point methods

Without giving details on the relevant topology, IPMs take for S the interior of the set R, and we have $S \subset R$. If, for instance, E is a Hilbert space and if $J + \epsilon G$ has a minimum in the open set S, then it is characterized by the conditions of a free extremum, which are much simpler to solve than the optimality conditions for the original problem. The challenge is to design the penalty G such that the penalized problem (51) reaches a minimum on the *open* set S. This is, in general, achieved by using suitable penalties that tends to $+\infty$ when u tends to the boundary of R. The second assumption is that the infimum of J over R can be reached by elements of S, that is, the interior of S. This can be achieved by assuming that R is convex.

In these settings, Theorem 5 proves the convergence of $J(u^*(\epsilon))$ toward the optimal value v_0. The convergence of $u^*(\epsilon)$ toward a minimizer of J over R can be then proven by arguments totally unrelated to the IPM, typically by using convex analysis.

APPENDIX B: CONVEXITY OF THE SET $\mathcal{C}$

$\mathcal{C}$ is convex, so $\mathcal{U}$ is convex. Let us prove that $\mathcal{U}$ is closed. To simplify the proof, we shall use the gauge function on $\mathcal{C}$; the properties of gauge functions are detailed in Section 4.1 and are quite general and independent of the other results presented in this paper. If $G_{\mathcal{C}}$ is the gauge function of $\mathcal{C}$, it is continuous and a vector u belongs to $\mathcal{C}$ if and only if $G_{\mathcal{C}}(u) \leqslant 1$. For $u \in L^\infty[0, T]$, define

$$G_{\mathcal{C}}^{\infty}(u) = \sup \operatorname{ess}_{t \in [0,T]} G_{\mathcal{C}}(u(t))$$

Then, clearly $u \in \mathcal{U}$ if and only if $G_{\mathcal{C}}^{\infty}(u) \leqslant 1$. Moreover, it is easy to check that $G_{\mathcal{C}}^{\infty}$ is continuous for the sup norm. This proves that $\mathcal{U}$ is closed for the sup norm.

APPENDIX C: PROPERTIES OF $\overset{\circ}{X}_{\mathrm{AD}}$

The set of points that verify $g_i(x) < 0$ is open because g_i is continuous. As a consequence, $\overset{\circ}{X}_{\mathrm{ad}}$ is a finite intersection of open sets and is thus open.

DOI: 10.1002/oca

To prove the density result, consider $x \in X_{\text{ad}}$ and let $\alpha \in (0,1)$. Because g_i is convex and $x_0 \in \overset{\circ}{X}_{\text{ad}}$, we have

$$g_i(\alpha x_0 + (1-\alpha)x) \leqslant \alpha g_i(x_0) + (1-\alpha) g_i(x) \leqslant \alpha g_i(x_0) < 0$$

By letting α tend to 0, this proves that $\overset{\circ}{X}_{\text{ad}}$ is dense in X_{ad}.

Proposition 13
For all $u \in \mathcal{U}$, the maximal solution x^u of the dynamics (2) is defined on $[0, T]$ and x^u is bounded by a constant that depends only on x_0 and D. Moreover, the mapping that maps $u \in \mathcal{U}$ to the trajectory $x^u \in C^0([0,T], \mathbb{R}^n)$ is Lipschitz when $\mathcal{U}$ is equipped with the L^1 or L^∞ norm. As a consequence, the functional $J(u)$ is Lipschitz over $\mathcal{U}$ when equipped with one of these norms.

Proof
Consider x^u as the maximal solution of (2). The use of the Gronwall Lemma ([52] p. 651) for the dynamics shows that x^u is bounded on its interval of definition. Because f is continuously differentiable, the boundedness of $u \in \mathcal{U}$ and of x^u implies that the derivatives of f are bounded when $u \in \mathcal{U}$. Consider now two controls u and v in $\mathcal{U}$. Using the Gronwall Lemma on $x^u - x^v$ shows that its dynamics is sub-linear with respect to $x^u - x^v$ and $u - v$ with a zero initial condition, which proves the regularity of x^u with respect to u, both in the L^1 and L^∞ norms. □

APPENDIX D: PROOF OF PROPOSITION 2

The result is trivial if $\alpha = 0$. We now assume $\alpha > 0$. From Proposition 13 given earlier in Appendix C and from the continuous differentiability of the g_i, there exists a constant Γ such that, for all $u \in \mathcal{U}$ and any s, t in $[0, T]$

$$|g_i(x^u(t)) - g_i(x^u(s))| \leqslant \Gamma |t - s| \tag{52}$$

Let $\alpha \in (0, \alpha_0]$ and $u \in \mathcal{U} \setminus \mathcal{X}^{\text{strict}}$. Then, there exists an index i for which $g_i(x^u)$ reaches 0 in $[0, T]$. Remember that $g_i(x_0) = -\alpha_0 < 0$. Denote by t_2 the first instant at which $g_i(x^u) = 0$ and $t_1 = \max\{s < t_2 \text{ s.t. } g_i(x^u(s)) = -\alpha \in [-\alpha_0, 0)\}$. From (52), we have

$$\alpha = g_i(x^u(t_2)) - g_i(x^u(t_1)) \leqslant \Gamma |t_2 - t_1| = \Gamma (t_2 - t_1)$$

As a consequence, we have $(t_2 - t_1) \geqslant \alpha / \Gamma$. Then, we have

$$-\alpha \leqslant g_i(x^u(s)) \leqslant 0 \ \ \forall s \in [t_1, t_2]$$

and hence $\mu_{g_i}(u, \alpha) \geqslant t_2 - t_1 \geqslant \alpha / \Gamma$. This concludes the proof.

APPENDIX E: PROOF OF PROPOSITION 4

There exists two closed ball B_N and B_M such that

$$B_N \subset \mathcal{C} \subset B_M$$

with *strict* inclusions. We define $N > 0$ (respectively $M > 0$) as the radius of the ball B_N (respectively B_M).Now, if $u = 0$, then $G_{\mathcal{C}}(u)$ is well defined and is equal to 0. We now assume that $u \neq 0$. Then

$$N \frac{u}{\|u\|} \in \mathcal{C}$$

DOI: 10.1002/oca

because it has norm N; as a consequence $u \in \frac{\|u\|}{N}\mathcal{C}$ which proves that $G_{\mathcal{C}}(u)$ is well defined and upper bounded by $\frac{\|u\|}{N}$. This proves property *a*) and the right hand side inequality of (14).

On the other side, if $u \neq 0$ then

$$M\frac{u}{\|u\|} \notin \mathcal{C}$$

because its norm is M. As a consequence $u \notin \frac{\|u\|}{M}\mathcal{C}$, and $u \notin \lambda\mathcal{C}$ if $\lambda \leqslant \frac{\|u\|}{M}$. Then, $G_{\mathcal{C}}(u)$ is lower bounded by $\frac{\|u\|}{M}$; this also holds if $u = 0$. This end the proof of property *b*).

The positive homogeneity of the gauge is trivial; because it is sub-additive [36], it is convex. The continuity comes from the fact that it is convex and lower and upper bounded in the neighborhood of any point. This proves properties *c*) and *d*).

The Dini derivative at 0 is obtained by observing that $G_{\mathcal{C}}(0) = 0$ and that $\frac{G_{\mathcal{C}}(hd)}{h} = G_{\mathcal{C}}(d)$ if $h > 0$. We see that there exists a directional derivative at 0 along the direction d if and only if the Dini derivatives along the directions d and $-d$ are equal, which is equivalent to the intersection of $\mathcal{C}$ with the line directed by d being symmetrical with respect to 0. This proves property *e*). Note that, if this symmetry holds for all directions, then the gauge function is a norm.

Let us prove property *f*). Because the boundary is continuously differentiable, there exists a continuously differentiable function $\varphi : \mathbb{R}^m \mapsto \mathbb{R}$ such that $\partial\mathcal{C} = \{u \text{ s.t. } \varphi(u) = 0\}$. For all $u \in \mathbb{R}^m \setminus \{0\}$, $\lambda u \in \partial\mathcal{C} \Leftrightarrow g(u,\lambda) \triangleq \varphi(\lambda u) = 0$. In the following, for any $u \in \mathbb{R}^m \setminus \{0\}$, we consider λ such that $g(u,\lambda) = 0$. From the convexity of $\mathcal{C}$ and because 0 belongs to the interior of $\mathcal{C}$, one has $\frac{\partial g}{\partial\lambda}(u,\lambda) =< \nabla\varphi(\lambda u), u >\neq 0$ for all $u \in \mathbb{R}^m \setminus \{0\}$. Using the implicit function theorem, there exits $(-\alpha,\alpha) \subset \mathbb{R}$ and U a neighborhood of u and a C^1 function $h : U \mapsto (-\alpha,\alpha)$ such that $\forall\mu \in (\lambda-\alpha, \lambda+\alpha)$ and $\forall v \in U$ $g(v,\mu) = 0 \Leftrightarrow \mu = h(v) = G_{\mathcal{C}}(v)$. Therefore, $G_{\mathcal{C}}$ is C^1 on $\mathbb{R}^m \setminus \{0\}$. This proves *f*).

Let us now prove property *g*). We first verify easily that $u \in \mathcal{C}$ if and only if $G_{\mathcal{C}}(u) \leqslant 1$ because $\mathcal{C}$ is closed [36]. Moreover, for any $u \neq 0$, the intersection of $\mathcal{C}$ with the half axis directed by u is the segment $\left[0, \frac{u}{G_{\mathcal{C}}(u)}\right]$ because $\mathcal{C}$ is closed and $G_{\mathcal{C}}(u) > 0$ [36]. As a consequence, $G_{\mathcal{C}}(u) = 1$ implies that u is in the boundary of $\mathcal{C}$. Conversely, if $G_{\mathcal{C}}(u) = 1 - 2\alpha$ with $\alpha > 0$, because $G_{\mathcal{C}}$ is continuous, there exists a neighborhood V of u where $G_{\mathcal{C}}(u) \leqslant 1 - \alpha$. For all elements $v \in V$, the intersection of $\mathcal{C}$ with the half-axis directed by v contains $\left[0, \frac{v}{1-\alpha}\right]$. This implies the existence of a neighborhood of u that is included in $\mathcal{C}$, and hence that u is interior to $\mathcal{C}$. Similarly, if $G_{\mathcal{C}}(u) > 1$, $u \notin \mathcal{C}$, one shows the existence of a neighborhood V of u and of $\alpha > 0$ such that the intersection of $\mathcal{C}$ with the half-axis directed by $v \in V$ is included in $\left[0, \frac{v}{1+\alpha}\right]$. Therefore, u belongs to the exterior of $\mathcal{C}$. A consequence of all this is that the boundary of $\mathcal{C}$ is exactly defined by $G_{\mathcal{C}}(u) = 1$, its interior by $G_{\mathcal{C}}(u) < 1$, and its exterior by $G_{\mathcal{C}}(u) > 1$. This ends the proof.

APPENDIX F: PROOF OF PROPOSITION 8

To exhibit an upper bound on the variation of the cost, this variation is split into three additive terms, bounding respectively the variation of the original cost, of the integral of the state penalty, and the integral of the control penalty.

Define $M = \max_i M_i$. From §4.4, one readily sees that

$$\|u - v\|_{L^1} \leqslant 2\alpha M \mu_u(\alpha)$$

We now proceed to establish bounds for the various terms.

F.1. Upper bound on the variation of the original cost

Here, an upper bound on $|\int_0^T \ell(x^v, v) - \ell(x^u, u)dt|$ is exhibited. It is noted K_ℓ. From Proposition 13, there exist $\Lambda \geqslant 0$ such that

DOI: 10.1002/oca

$$K_\ell \leqslant \Lambda \int_0^T \|x^v - x^u\|_{L^\infty} + \| v(t) - u(t) \| dt \leqslant \Lambda [CT + 1] \|v - u\|_{L^1}$$
$$\leqslant \Lambda[CT + 1]2\alpha M\mu_u(\alpha)$$

Define $U_l = \Lambda(CT + 1)2M$; then

$$K_l \leqslant U_l \alpha \mu_u(\alpha) \tag{53}$$

F.2. Upper bound on the variation of the state penalty

Note $K_{\gamma_g} \triangleq \epsilon \sum_{i=1}^q \int_0^T \gamma_g \circ g_i(x^v) - \gamma_g \circ g_i(x^u) dt$. Because γ_g is increasing, the integrand is positive only when $g_i(x^v(t)) \geqslant g_i(x^u(t))$. Yet, from the construction of v in (19), one has $\max_i g_i(x^v(t)) \leqslant -\beta_0$ for all $t \in [0, T]$. Using the convexity of γ_g, and the fact that g_i is Lipschitz with constant K_g on X^{ad}, one obtains

$$\begin{aligned}
K_{\gamma_g} &\leqslant \epsilon \sum_{i=1}^q \int_{g_i(x^v(t)) \geqslant g_i(x^u(t))} \gamma_g \circ g_i(x^v) - \gamma_g \circ g_i(x^u) dt \\
&\leqslant \epsilon \sum_{i=1}^q \int_{g_i(x^v(t)) \geqslant g_i(x^u(t))} |g_i(x^u(t)) - g_i(x^v(t))| \gamma_g'(g_i(x^v(t))) dt \\
&\leqslant \epsilon q \int_0^T K_g \|x^u - x^v\|_\infty \gamma_g'(-\beta_0) dt \\
&\leqslant \epsilon q T K_g C \|u - v\|_{L^1} \gamma_g'(-\beta_0) \\
&\leqslant \epsilon q T K_g C \gamma_g'(-\beta_0) 2\alpha M \mu_u(\alpha)
\end{aligned} \tag{54}$$

Define

$$U_g(\epsilon) = \epsilon q T K_g C \gamma_g'(-\beta_0) 2M$$

then, we have

$$K_{\gamma_g} \leqslant U_g(\epsilon) \alpha \mu_u(\alpha)$$

F.3. Upper bound on the variation of the control penalty

There, we aim at getting a negative variation so that, as a whole, the cost is decreased when replacing u by v.

Define

$$K_u \triangleq \epsilon \sum_{i=1}^p \int_0^T \gamma_u \left(G_{\mathcal{C}_i}(v_i(t))\right) - \gamma_u \left(G_{\mathcal{C}_i}(u_i(t))\right) dt.$$

From the construction of v (19), we know that $G_{\mathcal{C}_i}(v_i(t)) \leqslant G_{\mathcal{C}_i}(u_i(t))$. Because γ_u is non-decreasing, this proves that the integral is negative or null. Moreover, because $u_i = v_i$ when $G_{\mathcal{C}_i}(u_i) < 1 - \alpha_i$, we have

$$K_u = \epsilon \sum_{i=1}^p \int_{G_{\mathcal{C}_i}(u_i) \geqslant 1 - \alpha_i} \gamma_u \left(G_{\mathcal{C}_i}(v_i(t))\right) - \gamma_u \left(G_{\mathcal{C}_i}(u_i(t))\right) dt$$

DOI: 10.1002/oca

Using the convexity of γ_u, one has

$$\begin{aligned}
K_u &\leqslant -\epsilon \sum_{i=1}^{p} \int_{G_{\mathcal{C}_i}(u_i)\geqslant 1-\alpha} \| G_{\mathcal{C}_i}(v_i) - G_{\mathcal{C}_i}(u_i) \|_{L^\infty} \gamma_u'\left(G_{\mathcal{C}_i}(v_i(t))\right) dt \\
&= -\epsilon \sum_{i=1}^{p} \int_{G_{\mathcal{C}_i}(u_i)\geqslant 1-\alpha} \| G_{\mathcal{C}_i}(v_i) - G_{\mathcal{C}_i}(u_i) \|_{L^\infty} \gamma_u'\left[(1-2\alpha) G_{\mathcal{C}_i}(u_i(t))\right] dt \\
&\leqslant -\epsilon \sum_{i=1}^{p} \int_{G_{\mathcal{C}_i}(u_i)\geqslant 1-\alpha} \| G_{\mathcal{C}_i}(v_i) - G_{\mathcal{C}_i}(u_i) \|_{L^\infty} \gamma_u'\left[(1-2\alpha)(1-\alpha)\right] dt \\
&\leqslant -\epsilon \sum_{i=1}^{p} \int_{G_{\mathcal{C}_i}(u_i)\geqslant 1-\alpha} \| G_{\mathcal{C}_i}(v_i) - G_{\mathcal{C}_i}(u_i) \|_{L^\infty} \gamma_u'(1-3\alpha) dt \\
&\leqslant -\epsilon \sum_{i=1}^{p} \int_{G_{\mathcal{C}_i}(u_i)\geqslant 1-\alpha} 2\alpha \| G_{\mathcal{C}_i}(u_i) \|_{L^\infty} \gamma_u'(1-3\alpha) dt \\
&\leqslant -\epsilon \sum_{i=1}^{p} \int_{G_{\mathcal{C}_i}(u_i)\geqslant 1-\alpha} 2\alpha(1-\alpha)\gamma_u'(1-3\alpha) dt \\
&= -\epsilon \sum_{i=1}^{p} \mu_{u_i}(\alpha)\alpha\gamma_u'(1-3\alpha) \\
&\leqslant -\epsilon\alpha\gamma_u'(1-3\alpha)\mu_u(\alpha)
\end{aligned} \tag{55}$$

F.4. An upper bound on $K(u_2,\epsilon) - K(u_1,\epsilon)$

Gathering (53)–(55), we obtain

$$K(v,\epsilon) - K(u,\epsilon) \leqslant \alpha\left[U_\ell + U_g(\epsilon) - \epsilon\gamma_u'(1-3\alpha)\right]\mu_u(\alpha)$$

This concludes the proof of Proposition 8. One can see that the variation is negative for α small enough if $\gamma_u'(1-\alpha)$ tends to $+\infty$ when α tends to 0.

APPENDIX G: PROOF OF PROPOSITION 10

Let us define $f : B_{\|.\|}(0,1) \mapsto \text{int}(\mathcal{C})$ as

$$f(\xi) = \begin{cases} 0 \text{ if } \xi = 0 \\ \dfrac{\|\xi\|}{G_{\mathcal{C}}(\xi)}\xi \text{ otherwise} \end{cases}$$

The differentiability of the function f on $\mathbb{R}^m \setminus \{0\}$ stems from the differentiability of both $\|.\|$ and $G_{\mathcal{C}}$. The continuity at 0 stems from (14). Its inverse is given by the following function

$$f^{-1}(\xi) = \begin{cases} 0 \text{ if } \xi = 0 \\ G_{\mathcal{C}}(\xi)\dfrac{\xi}{\|\xi\|} \text{ otherwise} \end{cases}$$

Similarly, the differentiability of the function f^{-1} on $\mathbb{R}^m \setminus \{0\}$ stems from the differentiability of both $\|.\|$ and $G_{\mathcal{C}}$. The continuity at 0 stems from (14).

DOI: 10.1002/oca

Using (27), the function

$$\phi(\nu) \triangleq f \circ \psi(\nu) = \tanh(\| \nu \|)\,[G_{\mathcal{C}} \circ \psi(\nu)]^{-1} \tanh(\| \nu \|)\frac{\nu}{\| \nu \|}$$
$$= \tanh^2(\|\nu\|)\,[G_{\mathcal{C}} \circ \psi(\nu)]^{-1}\,\frac{\nu}{\|\nu\|}$$

maps $\mathbb{R}^m$ into $\mathrm{int}(\mathcal{C})$. This mapping being the composition of two homeomorphism not differentiable only in 0, ϕ is a homeomorphism differentiable everywhere except at 0. The inverse function $\sigma : \mathrm{int}(\mathcal{C}) \mapsto \mathbb{R}^m$ is the following:

$$\sigma(u) \triangleq \psi^{-1} \circ f^{-1}(u) = \mathrm{atanh}(G_{\mathcal{C}}(u))\frac{u}{\| u \|}$$

This concludes the proof.

APPENDIX H: REMARKS ON SOLUTION ALGORITHMS USING SATURATION FUNCTIONS

We rely on the routines `bvp4c` or `bvp5c` for solving the TBVPs. We shall denote them by `bvpxc`.

Values of ϵ We have first to define a sequence of values for ϵ for which we shall solve the penalized problems. A practical choice is to choose a geometric progression using

```
EPS=logspace(a,b,n)
```

The initial exponent `a` should not be too small in order for `bvpxc` to initialize correctly. Indeed, if ϵ is very small, we are close to the solution of the optimal problem; in particular, we have observed that the co-states converge numerically to singular functions (as expected from the theory). This means that if ϵ is small at the start, for `bvpxc` to succeed, it must use a very fine mesh refinement in the neighborhood of the singularities of the co-states. Because, at the beginning of the algorithm, we have no idea where the singularities occur, this is a very difficult task. This is why we start with ϵ not too small. The number `n` of values for ϵ is a tradeoff between two considerations

- if `n` is large, then we are close to a continuation method, as we feed the result of the previous iteration as an initial guess for the next iteration. This means that the collocation algorithm will be faster because the dynamics of the two iterations are close, and hence, the initial guess will be a good one. In particular, the mesh that is used for the initial guess will yield suitable refinements in the neighborhood of the 'singularities' of the costate when ϵ is small.
- on the other hand, a large `n` means a large number of iterations in the algorithm. In practice, beyond a certain value, increasing `n` only increases the computation time without improving the results.

The final exponent `b` essentially depends on the computing power of your machine. Of course `n` must be increased if `b` is increased. We shall denote by ϵ_i, $i = 1, \ldots, n$ the sequence previously defined.

Collocation options We must also provide a maximum number of mesh-points and a (relative and/or absolute) error for the collocation method. It is mainly a trial and error process; it is advised to try a small number of meshpoints and a reasonable tolerance first. Note that, beyond a certain value, increasing the possible number of mesh-points may be counterproductive. It is also possible to adapt these parameters to ϵ. A reasonable choice is to increase the possible number of mesh-points as ϵ decreases.

Initial guess An initial guess $I_{x,1}$ and $I_{p,1}$ must be given for the state and costate. The important point is that $I_{x,1}$ must satisfy $I_{x,1}[k] \in \overset{\circ}{X}_{\mathrm{ad}}$ for all the indices k. There is no requirement on the

DOI: 10.1002/oca

costate $I_{p,1}$, except that its final value is zero. The time samples can be arithmetically distributed at this stage, because for ϵ not very small, the solving of (40) and (41) should not be stiff.

Iterations Using the initial guess $I_{x,i}$ and $I_{p,i}$, use `bvpxc` to solve the TBVP (40) and (41) for $\epsilon = \epsilon_i$, with the settings described in the initialization step. Let (x_i, p_i) the output of the solver.

If $i < n$, define $(I_{x,i+1}, I_{p,i+1}) = (x_i, p_i)$, increment i and loop into the iteration step.

If $i = n$, the state produced by the algorithm is x_n, the costate is p_n, and the control is obtained by solving (39) with respect to the unknown $\phi(v^*)$. The iteration is stopped there.

Remark: when doing $(I_{x,i+1}, I_{p,i+1}) = (x_i, p_i)$, the time samples that are produced by `bvpxc` are passed as an initial mesh for the next iteration. This means that, if, for index i, the singularities begin to appear in p_i then the mesh is adapted to these singularities, and, if the distance between two successive values of ϵ is not too big, it will provide a suitable initial mesh for the collocation method.

REFERENCES

1. Murray RM, Hauser J, Jadbabaie A, Milam MB, Petit N, Dunbar WB, Franz R. Online control customization via optimization-based control. In *Software-Enabled Control, Information Technology for Dynamical Systems*, Samad T, Balas G (eds). John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2003; 149–174.
2. Byrd RH, Nocedal J, Waltz RA. KNITRO: An Integrated Package for Nonlinear Optimization. In *Large-Scale Nonlinear Optimization*, di Pillo G, Roma M (eds). Springer Verlag, 2006; 35–59.
3. Wright MH. The interior-point revolution in optimization: History, recent developments, and lasting consequences. *Bulletin (New Series) of the American Mathematical Society* 2004; **42**:39–56.
4. Forsgren A, Gill PE, Wright MH. Interior methods for nonlinear optimization. *SIAM Review* 2002; **4**(4):525–597.
5. Hauser J, Saccon A. A barrier function method for the optimization of trajectory functionals with constraints. *45th IEEE Conference on Decision and Control*, San Diego, 2006; 864–869.
6. Fabien BC. An extended penalty function approach to the numerical solution of constrained optimal control problems. *Optimal Control Applications and Methods* 1996; **17**(5):341–355.
7. Xing AQ, Chen ZH, Wang CL, Yao YY. Exact penalty function approach to constrained optimal control problems. *Optimal Control Applications and Methods* 1989; **10**(2):173–180.
8. Xing A-Q, Wang C-L. Applications of the exterior penalty method in constrained optimal control problems. *Optimal Control Applications and Methods* 1989; **10**(4):333–345.
9. Bryson AE, Ho YC. *Applied Optimal Control*. Ginn and Company: Waltham, MA, 1969.
10. Bonnans JF, Hermant A. Stability and sensitivity analysis for optimal control problems with a first-order state constraint and application to continuation methods. *ESAIM: Control, Optimisation and Calculus of Variations* 2008; **14**(4):825–863.
11. Hartl R, Sethi S, Vickson R. A survey of the maximum principles for optimal control problems with state constraints. *SIAM Review* 1995; **37**(2):181–218.
12. Hargraves C, Paris S. Direct optimization using nonlinear programming and collocation. *AIAA Journal of Guidance, Control and Dynamics* 1987; **10**:338–342.
13. Kojima A, Morari M. LQ control for constrained continuous-time systems. *Automatica* 2004; **40**:1143–1155.
14. Yuz J, Goodwin G, Feuer A, De Doná J. Control of constrained linear systems using fast sampling rates. *Systems and Control Letters* 2005; **54**:981–990.
15. Bemporad A, Casavol A, Mosca E. A predictive reference governor for constrained control systems. *Computers in Industry* 1998; **36**:55–64.
16. Bemporad A, Morari M, Dua V, Pisikopoulos EN. The explicit linear quadratic regulator for constrained systems. *Automatica* 2002; **38**:3–20.
17. Petit N, Milam M, Murray RM. Inversion based constrained trajectory optimization. *IFAC Symposium on Nonlinear Control Systems Design*, St. Petersburg, 2001.
18. Bhattacharya R. OPTRAGEN: A Matlab toolbox for optimal trajectory generation. *45th IEEE Conference on Decision and Control*, San Diego, 2006; 6832–6836.
19. Ross IM, Fahroo F. Pseudospectral knotting methods for solving nonsmooth optimal control problems. *AIAA Journal of Guidance, Control and Dynamics* 2004; **27**(3):397–405.
20. Wright SJ. Interior point methods for optimal control of discrete time systems. *Journal of Optimization Theory and Applications* 1993; **77**:161–187.
21. Vicente LN. On interior-point Newton algorithms for discretized optimal control problems with state constraints. *Optimization Methods and Software* 1998; **8**:249–275.
22. Leibfritz F, Sachs EW. Inexact SQP interior point methods and large scale optimal control problems. *SIAM Journal on Control and Optimization* 1999; **38**:272–293.
23. Jockenhövel T, Lorenz TB, Wächter A. Dynamic optimization of the Tennessee Eastman process using the Optcontrolcentre. *Computers and Chemical Engineering* 2003; **27**:1513–1531.

DOI: 10.1002/oca

24. Yu C, Li B, Loxton RC, Teo KL. Optimal discrete-valued control computation. *Journal of Global Optimization* 2013; **56**(2):503–518.
25. Yu C, Teo KL, Zhang L, Bai Y. A new exact penalty function method for continuous inequality constrained optimization problems. *Journal of Industrial and Management Optimization (JIMO)* 2010; **6**(4):895–910.
26. Li B, Yu C, Teo KL, Duan G-R. An exact penalty function method for continuous inequality constrained optimal control problem. *Journal of Optimization Theory and Applications* 2011; **151**(2):260–291.
27. Jiang C, Lin Q, Yu C, Teo KL, Duan G-R. An exact penalty method for free terminal time optimal control problem with continuous inequality constraints. *Journal of Optimization Theory and Applications* 2012; **154**(1):30–53.
28. Loxton RC, Teo KL, Rehbock V, Yiu KFC. Optimal control problems with a continuous inequality constraint on the state and the control. *Automatica* 2009; **45**(10):2250–2257.
29. Gerdts M, Kunkel M. A nonsmooth Newton's method for discretized optimal control problems with state and control constraints. *Journal of Industrial and Management Optimization (JIMO)* 2008; **4**(2):247–270.
30. Bonnans JF, Guilbaud T. Using logarithmic penalties in the shooting algorithm for optimal control problems. *Optimal Control Applications and Methods* 2003; **24**:257–278.
31. Malisani P, Chaplais F, Petit N. Design of penalty functions for optimal control of linear dynamical systems under state and input constraints. *50th IEEE Conference on Decision and Control and European Control Conference*, Orlando, 2011; 6697–6704.
32. Malisani P, Chaplais F, Petit N. A constructive interior penalty method for non linear optimal control problems with state and input constraints. *Proceedings of IEEE American Control Conference*, Montreal, 2012; 2669–2676.
33. Malisani P, Chaplais F, Petit N. A fully unconstrained interior point algorithm for multivariable state and input constrained optimal control problems. *European Congress on Computational Methods in Applied Sciences and Engineering*, Vienna, 2012.
34. Goddard RH. *A Method for Reaching Extreme Altitudes*. Smithsonian Int. Misc. Collections 71: Washington, 1919.
35. Kolmogorov AN, Fomin SV. *Elements of the Theory of Functions and Functional Analysis*. Dover Publications: Mineloa, New York, 1999.
36. Schwartz L. *Analyse Hilbertienne*. Ecole Polytechnique: Palaiseau, France, 1978.
37. Graichen K, Petit N, Kugi A. Transformation of optimal control problems with a state constraint avoiding interior boundary conditions. *Proceedings of the 47th IEEE Conference on Decision and Control*, Cancun, 2008; 913–920.
38. Graichen K, Petit N. Solving the Goddard problem with thrust and dynamic pressure constraints using saturation functions. *Proceedings of the 2008 IFAC World Congress*, Seoul, 2008; 14301–14306.
39. 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.
40. Graichen K, Petit N. Incorporating a class of constraints into the dynamics of optimal control problems. *Optimal Control Applications and Methods* 2009; **30**(6):537–561.
41. Graichen K, Kugi A, Petit N, Chaplais F. Handling constraints in optimal control with saturation functions and system extension. *Systems and Control Letters* 2010; **59**(11):671–679.
42. Graichen K. Feedforward control design for finite-time transition problems of nonlinear systems with input and output constraints. *Ph.D. Thesis*, Universität Stuttgart, 2006.
43. Malisani P. Dynamic control of energy in buildings using constrained optimal control by interior penalty. *Ph.D. Thesis*, Mines ParisTech, Paris, France, September 2012. (Available from: http://pastel.archives-ouvertes.fr/docs/00/74/00/44/PDF/Malisani.pdf) [Accessed on 25 August 2014].
44. Pontryagin LS, Boltyanskii VG, Gamkrelidze RV, Mishchenko EF. *The Mathematical Theory of Optimal Processes*. Interscience Publishers John Wiley & Sons, Inc.: New York, London, 1962.
45. Shampine L, Kierzenka J, Reichelt M. *Solving boundary value problems for ordinary differential equations in matlab with **bvp4c***, 2000. (Available from: http://www.mathworks.com/bvp_tutorial) [Accessed on 25 August 2014].
46. *BOCOP v1.03 User Guide*. INRIA: France, 2012.
47. Milam MB. Real-time optimal trajectory generation for constrained dynamical systems. *Ph.D. Thesis*, California Institute of Technology, Pasadena, USA, 2003.
48. Seywald H. Trajectory optimization based on differential inclusion. *AIAA Journal of Guidance, Control and Dynamics* 1994; **17**:480–487.
49. Malisani P. *Source code for solving goddard's problem*, 2013. (Available from: http://cas.ensmp.fr/~petit/code_optimisation_PM/main_goddard.m) [Accessed on 25 August 2014].
50. Fiacco AV, McCormick GP. The sequential unconstrained minimization technique for nonlinear programming, a primal-dual method. *Management Science* 1964; **10**(2):360–366.
51. Fiacco AV, McCormick GP. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. Wiley: New York, 1968.
52. Khalil H. *Nonlinear Systems*. Prentice Hall, 2002.

DOI: 10.1002/oca