Discussiones Mathematicae
Differential Inclusions, Control and Optimization 27(2007) 165–193

EVOLUTION OF STRUCTURE FOR DIRECT CONTROL OPTIMIZATION

Maciej Szymkat and Adam Korytowski

AGH University of Science and Technology
30–059 Kraków, Poland

e-mail: msz@ia.agh.edu.pl

Abstract

The paper presents the Monotone Structural Evolution, a direct computational method of optimal control. Its distinctive feature is that the decision space undergoes gradual evolution in the course of optimization, with changing the control parameterization and the number of decision variables. These structural changes are based on an analysis of discrepancy between the current approximation of an optimal solution and the Maximum Principle conditions. Two particular implementations, with spike and flat generations are described in detail and illustrated with computational examples.

Keywords: optimal control, direct optimization methods.

2000 Mathematics Subject Classification: 49J15, 65K10.

1. Introduction

General numerical methods of optimal control are divided into *direct* and *indirect* [1, 2, 5]. In the direct methods, approximating finite-dimensional optimization problems are constructed and solved by nonlinear programming (NLP) algorithms. The NLP problem can be formulated in two ways. The *simultaneous* approach, represented by the *direct collocation* methods [6], constructs the decision vector of discretized controls and discretized state trajectories, thus avoiding the need for numerical integration of the state equations. It usually leads to large-scale computations. In the *sequential* approach only the control functions are discretized and the state variables are computed by numerical integration [3, 9]. The general, well established

direct methods feature large areas of convergence but they are rather slow, especially in the final stage. This approach is not particularly demanding upon the user and is considered fairly universal. It has many implementations like SOCS [1, 2], DIRCOL [18], DIRMUS (Hinsberger), or NUDOCCCS [11].

Special direct methods are often applied to dynamic optimization problems with bounded control and the hamiltonian affine in control. The control parameterization is based on *switching times* and, possibly, on the end points of singular or state-constrained arcs, which become the NLP decision variables. The derivatives of the cost or states with respect to these variables, obtained from adjoint solutions [13, 16, 19, 20, 21] or by variational and difference techniques [7, 11, 23], are used for gradient optimization and for the verification of necessary and/or sufficient NLP optimality conditions [11]. Recently, a similar approach has been applied to hybrid systems and systems with a nonlinear dependence of the r.h.s. on the control variable (see, e.g., [17]). Such a parameterization usually results in a low-dimensional decision space and relatively good convergence, but requires the knowledge of optimal control structure before starting the NLP computations. Here, as in most of the literature, the control structure is understood as the sequence of sets of constraints, simultaneously pathwise active on the pair (control, state trajectory). Note that a different definition is used in this work (see Section 3). The optimal structure is usually sought outside the NLP problem by homotopy methods or by 'try and guess' procedures. A systematic way to establish the optimal control structure inside the NLP problem was proposed in the *variable parameterization* method [13, 19, 20, 21].

In the indirect approach, the optimal solution is computed by solving the boundary value problem obtained from the Maximum Principle. Multiple shooting is frequently used, with such implementations as BNDSCO (Oberle and Grimm), MUMUS (Hiltmann *et al.*) and MUSCOD-II (Diehl). The collocation methods for indirect computations (e.g., [8]) involve large systems of algebraic equations requiring specialized algorithms. The rate of convergence of indirect methods is usually very high, but their area of convergence is small and so they require good initial guesses for the adjoint vector. Practically, the optimal control structure has to be known beforehand. This can be achieved by a direct algorithm (see [15]) or using homotopy methods where a sequence of auxiliary problems is solved by multiple shooting (e.g., [4]).

This paper describes the method of Monotone Structural Evolution (MSE), which is a generalization of the results presented in [19, 20, 21]. The MSE is a direct computational approach to optimal control problems in systems described by ordinary differential equations, with control and state constraints [22]. Its distinctive feature is that the decision space undergoes gradual evolution in the course of optimization. This evolution runs according to pre-selected rules, with changing the control parameterization and the number of decision variables. Such changes, called structural, are followed by periods of gradient optimization in a constant decision space. The changes locally increase efficacy of the gradient optimization procedures. They are based on an analysis of discrepancy between the current approximation of optimal solution and the Maximum Principle conditions, and can be continued until this discrepancy becomes negligible. The number of decision variables may thus be kept comparatively small, at least in early stages of optimization. The control is preserved by every structural change so that monotone decrease of the performance index is achieved in the whole algorithm.

The paper is organized as follows. The optimal control problem with control and state constraints is formulated in Section 2. Sections 3 and 4 explain the MSE philosophy and introduce basic notions. The general algorithm of the MSE is described in Section 5. Sections 6, 7 and 8 are devoted to two particular implementations. The technique of spike generations is described in Section 6 and illustrated with an example of forced linear oscillator. Section 7 presents the technique of optimal control approximation by interval cubic polynomials and flat generations. It is illustrated in Section 8 with a problem of optimal ascent of the F-15 aircraft. The paper ends with conclusions.

2. Optimal control problem

Consider a control system

$$\dot{x} = f(x,u), \quad t \in [0,T], \quad x(0) = x^0 \tag{2.1}$$

where the state $x(t) \in \mathbb{R}^n$. The controls u are right-continuous, piecewise-continuous functions taking values in $\boldsymbol{U}$, a given set in $\mathbb{R}^m$. The horizon T is fixed or free, $T > 0$. The performance functional

$$S(u,T) = \varphi(x(T),T) \tag{2.2}$$

is minimized on the trajectories of (2.1) subject to equality terminal conditions

$$h_i(x(T)) = 0, \quad i = 1, \ldots, n_1 \tag{2.3}$$

and inequality state constraints valid for every $t \in [0, T]$

$$g_i(x(t)) \leq 0, \quad i = 1, \ldots, n_2. \tag{2.4}$$

It is assumed that the functions f, φ, h_i and g_i are continuously differentiable. The constraints (2.3) and (2.4) are treated by the exterior penalty method. To this end, introduce state variables y_i, $i = 1, \ldots, n_2$ such that

$$\dot{y}_i(t) = \frac{1}{2}\left(g_i(x(t))_+\right)^2, \quad y_i(0) = 0, \quad i = 1, \ldots, n_2 \tag{2.5}$$

and a family of auxiliary performance functionals ($\rho_{1i}, \rho_{2i} > 0$)

$$S_{\rho_1,\rho_2}(u, T) = \varphi(x(T), T) + \frac{1}{2}\sum_{i=1}^{n_1} \rho_{1i} h_i(x(T))^2 + \sum_{i=1}^{n_2} \rho_{2i} y_i(T) \tag{2.6}$$

which are minimized on the trajectories of (2.1), (2.5). It is assumed that the optimal solution of the state-unconstrained problem (2.1), (2.5), (2.6) tends to the optimal solution of the state-constrained problem (2.1)–(2.4) as all the coefficients ρ_{1i}, ρ_{2i} tend to infinity.

In the MSE method the control u may be determined on some subset $\boldsymbol{\Theta}$ of $[0, T]$ in a state-dependent form $u(t) = P(x(t), t)$ where $P : \mathbb{R}^n \times [0, T] \to \boldsymbol{U}$ is a given function of class $\boldsymbol{C}^1$ w.r.t. its first argument and piecewise continuous w.r.t. the second. We then define

$$\hat{f}(x, u, t) = \begin{cases} f(x, u), & t \notin \boldsymbol{\Theta} \\ f(x, P(x, t)), & t \in \boldsymbol{\Theta}. \end{cases} \tag{2.7}$$

More generally, it may happen that on some subintervals of $[0, T]$ only certain components of u are predetermined functions of state, time and, possibly, other control components. We can then formally substitute $\hat{f}(x, u, t) = f(x, P(x, u, t))$ for $t \in \Theta$, with an appropriately defined P.

Introduce the hamiltonian

$$H(\psi(t), x(t), u(t), t) = \psi(t)^\top \hat{f}(x(t), u(t), t) - \frac{1}{2}\sum_{i=1}^{n_2} \rho_{2i} \left(g_i(x(t))_+\right)^2. \tag{2.8}$$

The adjoint vector ψ is a continuous solution of the adjoint boundary problem

$$\dot{\psi}(t) = -\nabla_x H(\psi(t), x(t), u(t), t), \quad t \in [0, T] \tag{2.9}$$

$$\psi(T) = -\nabla_x \varphi(x(T), T) - \sum_{i=1}^{n_1} \rho_{1i} h_i(x(T)) \nabla h_i(x(T)). \tag{2.10}$$

Here ∇_x is the operator of derivative w.r.t. state.

The celebrated Maximum Principle of Pontryagin states a necessary optimality condition for the problem under consideration. Assume that an admissible pair u, T minimizes the functional S_{ρ_1,ρ_2} subject to (2.1), (2.5). Then

$$H(\psi(t), x(t), u(t), t) \geq H(\psi(t), x(t), v, t) \quad \forall t \in [0, T[\quad \forall v \in \boldsymbol{U} \tag{2.11}$$

where x and ψ satisfy (2.1), (2.9) and (2.10). In the free-horizon problem, additionally

$$H(\psi(T), x(T), u(T-), T-) = \frac{\partial}{\partial T} \varphi(x(T), T). \tag{2.12}$$

3. CONTROL STRUCTURE

The controls used in the MSE method have *structures*. The control structure is a sequence of procedures P_i, $i = 1, 2, \ldots, N$ that determine the control $u(t)$ in successive time intervals $[\tau_{i-1}, \tau_i[$, $u(t) = P_i(x(t), t, p_i)$ where x is the solution of (2.1) generated by u and p_i is a vector of real-valued parameters. If a procedure is independent of some argument, we may skip this argument in the notation. The points $\tau_0, \tau_1, \ldots, \tau_N$ are called *structural nodes*, $0 = \tau_0 \leq \tau_1 \leq \ldots \leq \tau_N = T$. The procedures P_i are taken from a fixed, finite set $\boldsymbol{P}$. The choice of the elements of $\boldsymbol{P}$ may be suggested by general techniques of numerical approximation and, which is particularly important, by expected properties of the optimal control following from the Maximum Principle. The procedures, their number, order and parameters, as well as the nodes $\tau_1, \ldots, \tau_{N-1}$ and, possibly, τ_N are decision variables in the optimization algorithm. The restrictions of control to intervals $[\tau_{i-1}, \tau_i[$ are called *arcs*. If $\tau_i = \tau_{i-1}$, we say that the i-th arc is *of zero length*.

A convenient way of defining a structure for a multidimensional control is to determine a structure for each control component separately. Let us give a few examples of typical procedures P_i, for a single control component or a scalar control u.

(i) Let $\boldsymbol{U} = [u_{\min}, u_{\max}]$ and $f(x,u) = f^0(x) + f^1(x)u$. It is then reasonable to define two constant procedures without parameters $P_{\min} = u_{\min}$ and $P_{\max} = u_{\max}$, and put $\{P_{\min}, P_{\max}\} \subset \boldsymbol{P}$. The respective control arcs are called *boundary* (arcs that are not boundary, are called *interior*).

(ii) For $\boldsymbol{U}$ and f as above define the *switching function* $\phi = \psi^\top f^1$ and denote by $\phi^{(i)}$ its i-th Lie derivative w.r.t. x along f. Suppose that the equation $\phi^{(r)}(x,\psi,u) = 0$ can be explicitly solved with respect to u, $u = w(x,\psi) \in \boldsymbol{U}$, and all the Lie derivatives $\phi^{(i)}$, $i < r$ are constant in u. Assume also that ψ in the expression $w(x,\psi)$ can be eliminated by solving the equations $\phi^{(i)}(x,\psi,u) = 0$, $i = 0, 1, \ldots, r-1$ w.r.t. ψ. The resulting expression for u defines a control procedure in a state-feedback form, $u(t) = P(x(t))$. The respective control arc is called a *candidate singular* arc.

(iii) In the situation described in (ii) suppose that a complete elimination of the components of ψ from $w(x,\psi)$ is impossible. Still, we may construct a candidate singular control procedure in a feedback form with parameters, $u(t) = P(x(t), t, p)$. The parameter p, to be determined by optimization may be interpreted as a vector of adjoint variables at an appropriately selected moment of time.

(iv) Consider a scalar state constraint $g(x) \leq 0$. Let $g^{(i)}$ be the i-th Lie derivative of g. Assume that the equation $g^{(r)}(x,u) = 0$ can be explicitly solved with respect to u, $u = P(x) \in \boldsymbol{U}$, and all the Lie derivatives $g^{(i)}$, $i < r$ are constant in u. The resulting expression for u defines a control procedure in a state-feedback form, $u(t) = P(x(t))$. The respective control arc is called a *candidate constrained* arc.

(v) Assume that u is a hamiltonian maximizer

$$u(t) = \arg\max_{w \in \boldsymbol{U}} H(\psi(t), x(t), w, t) = P(x(t), t, \psi(t_0))$$

with t_0 fixed. One may then define a control procedure in feedback form, $u(t) = P(x(t), t, p)$. The parameter vector p is a decision variable of the optimization process. General and attractive as it may look, this technique leads to poor optimization algorithms with extremely small areas of convergence.

(vi) Interior control procedures are frequently created by means of typical, general approximation techniques, e.g., $u(t) = P(t,p)$ with P being a polynomial in t of a given degree and p, the vector of its coefficients. Of course, functions other than polynomials may also be used.

4. STRUCTURAL CHANGES

An optimal control approximation in the direct approach is a value of an *approximation mapping* $A : \boldsymbol{D}_a \to \mathcal{U}$, from the admissible set $\boldsymbol{D}_a$ in a finite-dimensional space of decision variables $\boldsymbol{D}$, $\boldsymbol{D}_a \subset \boldsymbol{D}$ into a functional control space $\mathcal{U}$. Once $\boldsymbol{D}$, $\boldsymbol{D}_a \subset \boldsymbol{D}$ and A are chosen, the performance functional (2.6) may be redefined as a function of the decision vector

$$\Sigma(d) = S_{\rho_1,\rho_2}(A(d),T), \quad d \in \boldsymbol{D}_a. \tag{4.1}$$

We assume that Σ is continuously differentiable.

It is well known that the decision space most suitable for the optimal control approximation can only be chosen if certain properties of the optimal solution are known, a condition which is seldom satisfied in the beginning of optimization. At the same time, the performance of optimization algorithms rapidly worsens with a growing dimension of the decision space. These two premises motivate the construction of methods in which the decision space in the course of optimization is gradually adapted to the accumulated knowledge on the optimal solution. Optimization is started in a decision space of a small dimension, and the dimension is increased only when this is necessary for improving the approximation of the optimal solution. The adjustment of the decision space proceeds in a series of steps called *structural changes*, separated by periods of gradient optimization in a constant space.

Gradient optimization in a constant decision space $\boldsymbol{D}$ usually produces a sequence of points asymptotically convergent to some stationary point d_∞. A characteristic property of this process is that the rate of improvement of the performance index $\Sigma(d)$ slows down more and more when d approaches d_∞. While the point d_∞ typically fulfills the necessary optimality conditions in $\boldsymbol{D}$ (e.g., the KKT conditions), the corresponding control $u_\infty = A(d_\infty)$ is frequently far from satisfying the optimality conditions of the Maximum Principle. In such a case, the optimization procedure crawling towards d_∞ may be given a new impulse by appropriately changing the decision space

so that the image of the current decision vector $\bar{d}$ is far from any stationary point in the new space $\bar{\boldsymbol{D}}$.

For more formal definitons, consider a family Δ of decision spaces $\boldsymbol{D}$ where every $\boldsymbol{D} \in \Delta$ is a real vector space of finite dimension. In general, different spaces in Δ have different dimensions. To every $\boldsymbol{D} \in \Delta$, an admissible set $\boldsymbol{D}_a$ and an approximation mapping A depending on $\boldsymbol{D}$ are assigned. Each *structural change* is determined by a mapping $(\boldsymbol{D}, d) \mapsto (\bar{\boldsymbol{D}}, \bar{d})$ where $d \in \boldsymbol{D}_a \subset \boldsymbol{D}$, $\bar{d} \in \bar{\boldsymbol{D}}_a \subset \bar{\boldsymbol{D}}$, $\boldsymbol{D}, \bar{\boldsymbol{D}} \in \Delta$. Assume that the approximation mappings A and $\bar{A}$ are assigned to $\boldsymbol{D}$ and $\bar{\boldsymbol{D}}$, respectively. It is required that the *condition of control preservation* holds

$$\bar{A}(\bar{d}) = A(d), \quad d \in \boldsymbol{D}_a, \quad \bar{d} \in \bar{\boldsymbol{D}}_a. \tag{4.2}$$

The interpretation is that d is a point reached in a (constant) decision space $\boldsymbol{D}$, and it is estimated that further optimization in another space $\bar{\boldsymbol{D}}$ will be more effective. Optimization is then continued in $\bar{\boldsymbol{D}}$, starting from an appropriately determined point $\bar{d} \in \bar{\boldsymbol{D}}$. In the MSE, this change of decision space implies a change of the sequence of procedures P_i. Thanks to condition (4.2) the control (as an element of $\mathcal{U}$) is not immediately affected, and in consequence the performance index monotonously decreases during the overall optimization. Typically, only few selected elements of the structure can be affected by a structural change. For example, one or two new procedures $\bar{P}_i$ are introduced with inserting the corresponding new nodes, or one of the procedures P_i is modified. The new arcs are often of zero length.

Two kinds of structural changes are typical for the MSE: *generations* and *reductions*. The dimension of the decision space increases in a generation, and is diminished in a reduction.

In the MSE, the structural changes aimed at speeding up the optimization are effected by *driving generations*. To explain their construction, define the *efficiency* of a generation. Assume that the generation changes the decision space from $\boldsymbol{D}$ to $\bar{\boldsymbol{D}}$. Let $d_0 \in \boldsymbol{D}$ and $\bar{d}_0 \in \bar{\boldsymbol{D}}$ be the decision vectors immediately before and after the generation. Let also $\Sigma(d)$ for $d \in \boldsymbol{D}$ be given by (4.1), and $\bar{\Sigma}(\bar{d}) = S_{\rho_1,\rho_2}(\bar{A}(\bar{d}), T)$ for $\bar{d} \in \bar{\boldsymbol{D}}$. Denote the antigradients $-\nabla\Sigma(d_0)$ and $-\nabla\bar{\Sigma}(\bar{d}_0)$ by γ and $\bar{\gamma}$, respectively. If γ and $\bar{\gamma}$ are admissible, that is, point to the interior of the respective admissible sets in $\boldsymbol{D}$ and $\bar{\boldsymbol{D}}$, the efficiency of the generation is defined as the difference of squared Euclidean norms

$$E = \|\bar{\gamma}\|^2 - \|\gamma\|^2. \tag{4.3}$$

Such a definition is justified in two ways. First, the squared norm of the gradient, multiplied by -1, is equal to the derivative of the performance index w.r.t. the line search parameter in the steepest descent direction

$$\nabla_z \Sigma(d_0 + z\gamma)|_{z=0+} = -\|\gamma\|^2, \quad \nabla_z \bar{\Sigma}(\bar{d}_0 + z\bar{\gamma})|_{z=0+} = -\|\bar{\gamma}\|^2. \tag{4.4}$$

The efficiency thus determines the increase of steepness of the performance index. Secondly, the efficiency so defined does not depend on those components of the gradient of performance index that are not affected by the generation, which simplifies computations. In the general case, the antigradients in (4.3) are replaced by their orthogonal projections onto the local conical approximations of the admissible sets.

The driving generation takes place if its relative efficiency $E/\|\gamma\|^2$ (defined for $\gamma \neq 0$) exceeds a given threshold. By choosing the threshold one can control the trade-off between the dimension of decision space and gradient magnitude. The number of simultaneously generated nodes is limited by additional rules (e.g., one or two per arc, solely at local maximizers of relative efficiency), to avoid an undesirable increase of the number of decision variables. Additional requirements can be imposed on generations to obtain controls with pre-selected regularity properties, like continuity or smoothness. The choice of particular generations is also subject to the condition that optimization should converge to the optimal control in the strong sense.

Besides the driving generations, the MSE method admits *saturation generations*, enforced by the requirement that at the moment of gradient computation each control arc has to be either purely boundary or purely interior. They are performed when the optimization process transforms an interior arc into one that contains a subarc with an active control constraint. The corresponding time interval is then divided by introducing new structural nodes.

A typical reduction consists in eliminating an arc of zero length when it is not promising. More precisely, every arc of zero length is subject to reduction if the directional derivative of the performance index w.r.t. its boundaries is nonnegative for all admissible directions. At the same time the decision variables that describe this control arc are eliminated, including at least one of the respective nodes. Such a reduction occurs each time when one of the constraints $\tau_0 \leq \tau_1 \leq \ldots \leq \tau_N$ becomes active after the line search of the gradient optimization algorithm. Another typical reduction occurs when two adjacent arcs described by identical procedures are unified.

5. General algorithm

The basic algorithm of the MSE consists of the following steps.

1^0 Selection of initial decision space $\boldsymbol{D}$ and starting point $d \in \boldsymbol{D}_a \subset \boldsymbol{D}$.

2^0 Termination, if optimality conditions in $\mathcal{U}$ are satisfied.

3^0 Generation, if it is sufficiently promising or needed.

4^0 Iteration of gradient optimization in current decision space $\boldsymbol{D}$.

5^0 Reduction, if necessary.

6^0 Return to 2^0.

The optimality conditions verified in step 2^0 may be of two types, used jointly.

(i) *Necessary conditions of the Maximum Principle.* These conditions are always included in the MSE method, though in different forms. If T is fixed, the requirement of sufficient accuracy of hamiltonian maximization may be expressed by an inequality $\|\chi\|_p \leq \eta_0$ where $\|\cdot\|_p$ denotes the norm in $\mathbb{L}^p(0,T)$ for some $p \in \{1,2,\ldots,\infty\}$, $\eta_0 \geq 0$ is a threshold, and

$$\chi(t) = \sup\{H(\psi(t),x(t),v,t) - H(\psi(t),x(t),u(t),t), v \in \boldsymbol{U}\}, \quad t \in [0,T].$$

If T is a decision variable, then condition (2.12) should also be satisfied with sufficient accuracy. Other termination conditions of this type, valid under special assumptions may be formulated with the use of the derivative $\nabla_u H$ or the switching function ϕ. It is also possible to express the termination condition 2^0 as a condition of the existence of appropriately efficient generations in step 3^0.

(ii) *Necessary conditions following from a lower bound on performance functional.* Assume that $S_{\min} = \inf\{S_{\rho_1,\rho_2}(u,T) : u, T \text{ admissible}\}$ can be evaluated. The termination condition has the form $S_{\rho_1,\rho_2}(u,T) - S_{\min} \leq \eta_1$ where $\eta_1 \geq 0$ is a threshold.

Step 3^0 is distinctive for the MSE algorithms and crucial for their convergence. The changes of structure are mainly performed to speed up optimization when a stationary point in the current decision space is being approached.

This algorithm should be equipped with special procedures for gradient computation and evaluation of efficiency of generations. These procedures are based on the solutions of the adjoint boundary problems (2.9), (2.10).

While the gradients $\nabla\Sigma(d)$ can also be computed by other techniques, like variational equations or discrete numerical approximation, the adjoint trajectories are indispensable for the estimation of accuracy of fulfillment of the Maximum Principle conditions, as well as for effectively choosing generations with satisfactory efficiency.

To treat state constraints, an outer loop of penalty modification has to be added. In the gradient optimization of step 4^0, the bounds on structural nodes and control constraints may be respected due to an appropriate organization of line search. Numerical solutions of differential equations can be conveniently obtained by the RK4 method with mesh adjusted so as to include all discontinuity points.

6. TECHNIQUE OF SPIKE GENERATIONS

The technique of spike generations will be explained with an example of forced linear oscillator. Consider the system

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -x_1 + u, \end{aligned} \tag{6.1}$$

deifined in $[0, T]$, with initial conditions

$$x_1(0) = 3, \quad x_2(0) = -3.$$

The horizon is fixed, $T = 4.4$. The performance index

$$S(u) = \frac{1}{2}\|x(T)\|^2. \tag{6.2}$$

The controls are bounded, $|u| \leq 2$, and the state is subject to a constraint

$$g(x(t)) = x_2(t) - 0.5 \leq 0, \quad t \in [0, T]. \tag{6.3}$$

We employ the penalty method described in Section 2 with

$$\dot{y}(t) = \frac{1}{2}((x_2 - 0.5)_+)^2, \quad y(0) = 0 \tag{6.4}$$

$$S_\rho(u) = \frac{1}{2}\|x(T)\|^2 + \rho y(T), \quad \rho > 0. \tag{6.5}$$

Three procedures are used to compute control values. Two of them produce boundary control arcs, $P_{\pm}(t) = \pm 2$ and one, candidate constrained arcs: $P_{\mathrm{con}}(x) = x_1$. This last formula follows from equating $g^{(1)}(x,u) = \dot{x}_2 = -x_1 + u$ to zero. Note that any candidate constrained arc becomes also singular in the optimal solution of the penalized problem (6.1), (6.4), (6.5). For a given control structure, the internal structural nodes are the only decision variables, $d = (\tau_1, \ldots, \tau_{N-1})$. The hamiltonian and the adjoint boundary problem are given by (2.8)–(2.10). Denote the functional (6.5) as a function of the decision vector by Σ. It is well known [16, 22] that its partial derivatives are given by

$$\begin{aligned} \nabla_{\tau_i}\Sigma(d) &= \psi(\tau_i)^{\top}\left(f(x(\tau_i), u(\tau_i+)) - f(x(\tau_i), u(\tau_i-))\right) \\ &= \phi(\tau_i)(u(\tau_i+) - u(\tau_i-)), \quad i = 1, \ldots, N-1 \end{aligned} \tag{6.6}$$

if $\tau_{i-1} < \tau_i < \tau_{i+1}$. Here $\phi = \psi_2$ is the switching function. In the MSE we have to generalize the formula (6.6) so that it covers also cases of arcs of zero length. For a control structure determined by a sequence of procedures $(P_1, \ldots, P_N)$, denote the value of the procedure P_j at time t by $v_j(t)$. Then

$$\begin{aligned} \nabla_{\tau_i}\Sigma(d) &= \psi(\tau_i)^{\top}\left(f(x(\tau_i), v_{i+1}(\tau_i)) - f(x(\tau_i), v_i(\tau_i))\right) \\ &= \phi(\tau_i)(v_{i+1}(\tau_i) - v_i(\tau_i)), \quad i = 1, \ldots, N-1. \end{aligned} \tag{6.7}$$

If (6.6) is applicable, (6.7) yields the same results. If $\tau_{i-1} = \tau_i < \tau_{i+1}$, $0 < i < N$, then (6.7) determines the right partial derivative, and if $\tau_{i-1} < \tau_i = \tau_{i+1}$, $0 < i < N$, it determines the left partial derivative. The case $\tau_{i-1} = \tau_i = \tau_{i+1}$ is excluded from consideration.

In this example we only use *spike generations*, in which the new control arcs are of zero length. To explain the rules for generations, assume that the control structure is defined by a sequence of procedures $(P_1, \ldots, P_N)$, $P_i \neq P_{i-1}$ for $i = 2, \ldots, N$, $P_i \in \{P_+, P_-, P_{\mathrm{con}}\}$ for $i = 1, \ldots, N$. The respective structural nodes satisfy $0 = \tau_0 < \tau_1 < \ldots < \tau_N = T$. For every $\tau \in [0,T] \backslash \{\tau_1, \ldots, \tau_{N-1}\}$, define $P_{\mathrm{ad}}(\tau) \in \{P_+, P_-, P_{\mathrm{con}}\}$:

(i) if $\phi(\tau) > 0$ and $u(\tau) < P_{\mathrm{con}}(x(\tau)) < P_+(\tau)$, then $P_{\mathrm{ad}}(\tau) = P_{\mathrm{con}}$,

(ii) if $\phi(\tau) < 0$ and $P_-(\tau) < P_{\mathrm{con}}(x(\tau)) < u(\tau)$, then $P_{\mathrm{ad}}(\tau) = P_{\mathrm{con}}$,

(iii) if $\phi(\tau) > 0$, $u(\tau) < P_+(\tau)$ and $P_{\mathrm{con}}(x(\tau)) \notin]u(\tau), P_+(\tau)[$, then $P_{\mathrm{ad}}(\tau) = P_+$,

(iv) if $\phi(\tau) < 0$, $P_-(\tau) < u(\tau)$ and $P_{\mathrm{con}}(x(\tau)) \notin]P_-(\tau), u(\tau)[$, then $P_{\mathrm{ad}}(\tau) = P_-$,

(v) if none of the above conditions is fulfilled then $P_{\mathrm{ad}}(\tau) = P_i$, where $\tau \in [\tau_{i-1}, \tau_i]$, $i = 1, \ldots, N$.

Suppose first that $\tau \in]\tau_{j-1}, \tau_j[$ and consider a generation in which the control structure is changed to $(\bar{P}_1, \ldots, \bar{P}_{N+2})$, $\bar{P}_i = P_i$ for $i \leq j$, $\bar{P}_{j+1} = P_{\mathrm{ad}}(\tau)$, and $\bar{P}_i = P_{i-2}$ for $i > j+1$. The new structural nodes are $\bar{\tau}_i$, $i = 0, \ldots, N+2$, $\bar{\tau}_i = \tau_i$ for $i < j$, $\bar{\tau}_j = \bar{\tau}_{j+1} = \tau$, and $\bar{\tau}_i = \tau_{i-2}$ for $i > j+1$. The efficiency (4.3) of this generation is equal to

$$E(\tau) = 2\phi(\tau)^2(\bar{v}_{j+1}(\tau) - v_j(\tau))^2 \tag{6.8}$$

where $\bar{v}_{j+1}(\tau)$ is the value of the procedure $\bar{P}_{j+1}$ at time τ.

Let now $\tau = 0$ and let $(\bar{P}_1, \ldots, \bar{P}_{N+1})$ be the control structure after the generation, $\bar{P}_1 = P_{\mathrm{ad}}(0)$ and $\bar{P}_i = P_{i-1}$ for $i > 1$. The new structural nodes are $\bar{\tau}_i$, $i = 0, \ldots, N+1$, $\bar{\tau}_0 = 0$ and $\bar{\tau}_i = \tau_{i-1}$ for $i \geq 1$. The efficiency equals

$$E(0) = \phi(0)^2(\bar{v}_1(0) - v_1(0))^2. \tag{6.9}$$

Consider now a spike generation at the horizon, $\tau = T$. The structure after the generation is given by $(\bar{P}_1, \ldots, \bar{P}_{N+1})$, $\bar{P}_{N+1} = P_{\mathrm{ad}}(T)$ and $\bar{P}_i = P_i$ for $i \leq N$. The new structural nodes are $\bar{\tau}_i$, $i = 0, \ldots, N+1$, $\bar{\tau}_{N+1} = T$ and $\bar{\tau}_i = \tau_i$ for $i \leq N$. The efficiency is

$$E(T) = \phi(T)^2(\bar{v}_{N+1}(T) - v_N(T))^2. \tag{6.10}$$

The generations made thus far have only been hypothetical, and served the purpose of determining the function $E : [0, T]\backslash\{\tau_1, \ldots, \tau_{N-1}\} \to \mathbb{R}$. In order to describe the generations actually used in the optimization algorithm define

$$\hat{E}(\tau) = \begin{cases} \frac{1}{2}E(\tau), & \tau \notin \{0, T\} \\ E(\tau), & \tau \in \{0, T\}. \end{cases}$$

The factor $\frac{1}{2}$ is introduced to give some preference to inserting spikes at 0 and T since the number of decision variables is then increased only by one. Let I be the set of all integers i in $\{1, \ldots, N\}$ such that $\hat{E}$ has a maximum in $[\tau_{i-1}, \tau_i]\backslash\{\tau_1, \ldots, \tau_{N-1}\}$, attained at some $\hat{\tau}_i$, and this maximum satisfies

$$\hat{E}(\hat{\tau}_i) > \varepsilon\|\gamma\|^2.$$

Here γ is the gradient of Σ immediately before the generation and $\varepsilon>0$, a given relative efficiency threshold. The control structure immediately after the generation is determined in the following way. Let θ and θ_0 be strictly increasing sequences constructed of all elements of the sets $\{\hat{\tau}_i : i \in I\}$ and $\{\hat{\tau}_i : i \in I\}\backslash\{0, T\}$, respectively. To obtain the sequence of the new structural nodes $(\bar{\tau}_0, \ldots, \bar{\tau}_{\bar{N}})$, sort the concatenation of sequences $(\tau_0, \ldots, \tau_N)$, θ and θ_0 in a nondecreasing order. The new control structure $(\bar{P}_1, \ldots, \bar{P}_{\bar{N}})$ which includes all the procedures P_i, $i \in \{1, \ldots, N\}$ and $P_{\mathrm{ad}}(\hat{\tau}_i)$, $i \in I$ is characterized as follows. Let $j \in \{1, \ldots, \bar{N}\}$. If $\bar{\tau}_{j-1} = \bar{\tau}_j = \hat{\tau}_i$ for some $i \in I$, then $\bar{P}_j = P_{\mathrm{ad}}(\hat{\tau}_i)$, $\bar{P}_{j-1} = P_i$ for $j > 1$, and $\bar{P}_{j+1} = P_i$ for $j < \bar{N}$. Otherwise, there is exactly one i in $\{1, \ldots, N\}$ such that $\bar{\tau}_j = \tau_i$. Then $\bar{P}_j = P_i$, and $\bar{P}_{j+1} = P_{i+1}$ for $j < \bar{N}$.

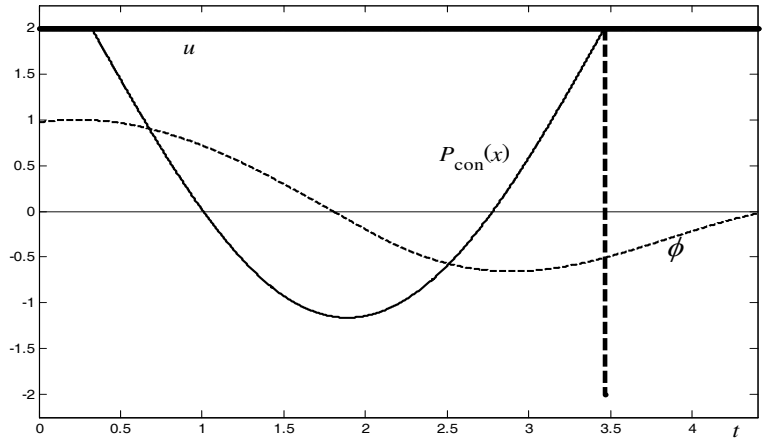

Figure 6.1. First generation (boundary)

The computations are started with an optimal control approximation $u \equiv 2$, that is, a one-element control structure P_+ and a (sufficiently) large ρ. The first generation (Figure 6.1) inserts a boundary spike, and the new structure is (P_+, P_-, P_+). After a few BFGS iterations we obtain the situation in Figure 6.2, where conditions for inserting two spikes are satisfied. The first of the generated arcs (boundary) is reduced after several iterations, and the second, a candidate constrained arc grows up (Figure 6.3) to reach the optimal solution shown in Figures 6.4 and 6.5.

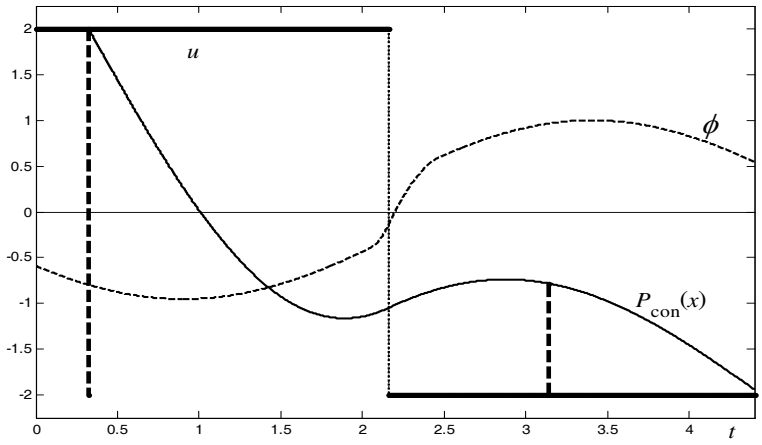

Figure 6.2. Second generation (boundary and constrained)

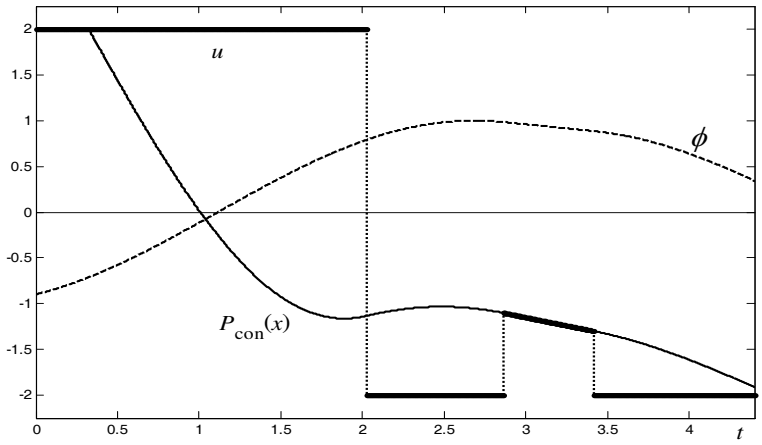

Figure 6.3. After a few more iterations

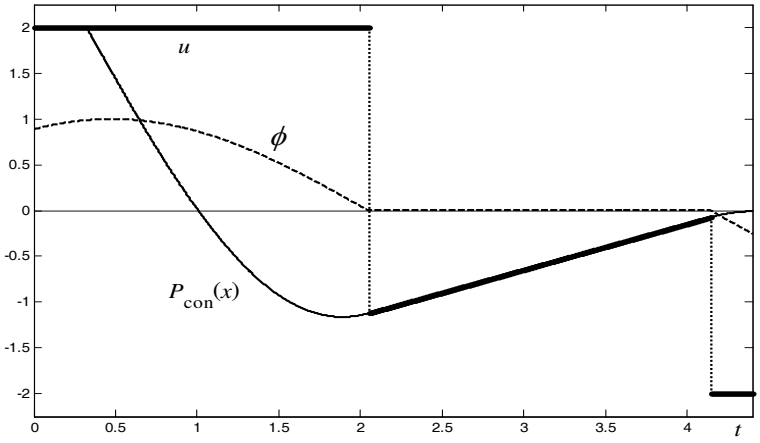

Figure 6.4. Optimal control

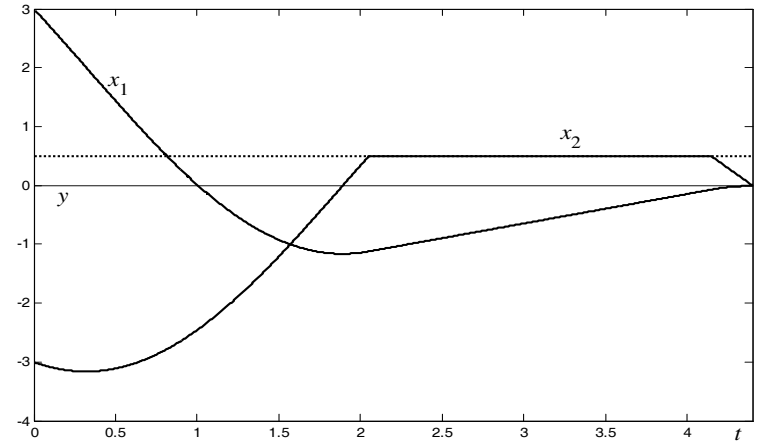

Figure 6.5. Optimal trajectories

7. INTERVAL CUBIC POLYNOMIALS AND FLAT GENERATIONS

7.1. Control approximation

Assume that the horizon T is fixed, $\boldsymbol{U} = \mathbb{R}$, and the admissible controls are interval Hermite interpolation polynomials

$$\begin{aligned} u(t) &= p^0_{i-1} V_0(t, \tau_{i-1}, \tau_i) + p^0_{i-} V_0(t, \tau_i, \tau_{i-1}) \\ &\quad + p^1_{i-1} V_1(t, \tau_{i-1}, \tau_i) + p^1_{i-} V_1(t, \tau_i, \tau_{i-1}), \\ &\quad t \in [\tau_{i-1}, \tau_i[, \quad i = 1, \dots, N \end{aligned} \tag{7.1}$$

where

$$V_0(t,a,b) = \frac{(t-b)^2(2t+b-3a)}{(b-a)^3}, \quad V_1(t,a,b) = \frac{(t-b)^2(t-a)}{(b-a)^2}. \tag{7.2}$$

For every fixed i, the r.h.s. of (7.1) is a polynomial of degree not higher than three. The advantage of this control representation is that all decision variables have obvious geometric interpretations, and it is easy to formulate continuity requirements on control and its derivative at structural nodes. Note that $u(\tau_{i-1}+) = p^0_{i-1}$, $\dot{u}(\tau_{i-1}+) = p^1_{i-1}$, $u(\tau_i-) = p^0_{i-}$, $\dot{u}(\tau_i-) = p^1_{i-}$ for $i = 1, \dots, N$. The time moments τ_0, $\tau_1, \dots,$ τ_N may be interpreted as the structural nodes in the general formulation of the MSE. The decision vector consists of the variables $\tau_1, \tau_2, \dots, \tau_{N-1}$ and p^0_{i-1}, p^1_{i-1}, p^0_{i-}, p^1_{i-} for $i = 1, \dots, N$. The nodes are subject to constraints: $0 = \tau_0 \le \tau_1 \le \dots \le \tau_N = T$. Conditions of continuity of control and/or its derivative may be imposed at certain nodes, $p^s_i = p^s_{i-}$ $\forall i \in K_s$, $s \in \{0,1\}$ where K_0 and K_1 are given subsets of $\{1, \dots, N-1\}$.

7.2. Derivatives of performance functional

Define

$$J_s(a,b) = -\int_a^b h(t) V_s(t,a,b) \mathrm{d}t \ \text{ if } a \neq b, \quad J_s(a,a) = 0$$

where $h(t) = \nabla_u H(\psi(t), x(t), u(t))$. For $s = 0, 1$, calculate the derivatives of the performance index Σ w.r.t. p^s_i and p^s_{i-}

$$\nabla_{p^s_i} \Sigma = J_s(\tau_i, \tau_{i+1}), \quad i \in \{0, \dots, N-1\} \backslash K_s \tag{7.3}$$

$$\nabla_{p^s_{i-}}\Sigma = -J_s(\tau_i,\tau_{i-1}),\quad i\in\{1,\dots,N\}\backslash K_s \tag{7.4}$$

$$\nabla_{p^s_i}\Sigma = J_s(\tau_i,\tau_{i+1}) - J_s(\tau_i,\tau_{i-1}),\quad i\in K_s. \tag{7.5}$$

The derivatives w.r.t. the nodes τ_i, $i=1,\dots,N-1$ can be computed from

$$\nabla_{\tau_i}\Sigma = \Delta H_i - \int_{\tau_{i-1}}^{\tau_{i+1}} h(t)\vartheta_i(t)\mathrm{d}t$$

where

$$\begin{aligned}
\Delta H_i &= H(\psi(\tau_i),x(\tau_i),p^0_i) - H(\psi(\tau_i),x(\tau_i),p^0_{i-})\\
\vartheta_i(t) &= p^0_{i-1}\nabla_{\tau_i}V_0(t,\tau_{i-1},\tau_i) + p^0_{i-}\nabla_{\tau_i}V_0(t,\tau_i,\tau_{i-1})\\
&\quad + p^1_{i-1}\nabla_{\tau_i}V_1(t,\tau_{i-1},\tau_i)p^1_{i-}\nabla_{\tau_i}V_1(t,\tau_i,\tau_{i-1}),\quad t<\tau_i\\
\vartheta_i(t) &= p^0_i\nabla_{\tau_i}V_0(t,\tau_i,\tau_{i+1}) + p^0_{i+1-}\nabla_{\tau_i}V_0(t,\tau_{i+1},\tau_i)\\
&\quad + p^1_i\nabla_{\tau_i}V_1(t,\tau_i,\tau_{i+1}) + p^1_{i+1-}\nabla_{\tau_i}V_1(t,\tau_{i+1},\tau_i),\quad t>\tau_i.
\end{aligned}$$

Equivalently,

$$\begin{aligned}
\nabla_{\tau_i}\Sigma &= \Delta H_i + J_0(\tau_i,\tau_{i-1})p^1_{i-} + J_1(\tau_i,\tau_{i-1})\ddot{u}(\tau_i-)\\
&\quad - J_0(\tau_i,\tau_{i+1})p^1_i - J_1(\tau_i,\tau_{i+1})\ddot{u}(\tau_i+).
\end{aligned} \tag{7.6}$$

Direct calculation yields

$$\ddot{u}(\tau_i-) = \frac{6(p^0_{i-1}-p^0_{i-})}{(\tau_i-\tau_{i-1})^2} + \frac{2(p^1_{i-1}+2p^1_{i-})}{\tau_i-\tau_{i-1}}$$

$$\ddot{u}(\tau_i+) = -\frac{6(p^0_i-p^0_{i+1-})}{(\tau_{i+1}-\tau_i)^2} - \frac{2(2p^1_i+p^1_{i+1-})}{\tau_{i+1}-\tau_i}.$$

7.3. Flat generation

Consider the generation of a new structural node $\tau\in\,]\tau_{z-1},\tau_z[$, for some $z\in\{1,\dots,N\}$. After the generation the nodes constitute a nondecreasing sequence $\bar\tau_0,\dots,\bar\tau_{\bar N}$ where $\bar N = N+1$, $\bar\tau_i=\tau_i$ for $i<z$, $\bar\tau_z=\tau$, $\bar\tau_i=\tau_{i-1}$ for $i>z$. The set of indices of all the nodes at which the s-th derivative is required to be continuous after the generation, is denoted by $\bar K_s$, $s=0,1$.

Of course, $\bar{K}_s \cap \{1,\dots,z-1\} = K_s \cap \{1,\dots,z-1\}$ and $\bar{K}_s \cap \{z+1,\dots, N+1\} = (K_s \cap \{z,\dots,N\}) + 1$.

The control before the generation has the form (7.1), and after the generation it is given by

$$\begin{aligned} u(t) &= \bar{p}^0_{i-1} V_0(t, \bar{\tau}_{i-1}, \bar{\tau}_i) + \bar{p}^0_{i-} V_0(t, \bar{\tau}_i, \bar{\tau}_{i-1}) \\ &\quad + \bar{p}^1_{i-1} V_1(t, \bar{\tau}_{i-1}, \bar{\tau}_i) + \bar{p}^1_{i-} V_1(t, \bar{\tau}_i, \bar{\tau}_{i-1}) \\ &\quad t \in [\bar{\tau}_{i-1}, \bar{\tau}_i[, \quad i = 1,\dots,\bar{N}. \end{aligned} \tag{7.7}$$

Since (7.1) and (7.7) are identical functions of time, the following relationships between the coefficients are valid for $s = 0, 1$

$$\bar{p}^s_i = p^s_i, \quad \bar{p}^s_{i-} = p^s_{i-} \text{ for } i < z \tag{7.8}$$

$$\bar{p}^s_{i+1} = p^s_i, \quad \bar{p}^s_{i+1-} = p^s_{i-} \text{ for } i \geq z$$

$$\bar{p}^s_z = \bar{p}^s_{z-} = u^{(s)}(\tau) \tag{7.9}$$

where the superscript (s) denotes the s-th dervitative w.r.t. time

$$\begin{aligned} u^{(s)}(\tau) &= p^0_{z-1} V^{(s)}_0(\tau, \tau_{z-1}, \tau_z) + p^0_{z-} V^{(s)}_0(\tau, \tau_z, \tau_{z-1}) \\ &\quad + p^1_{z-1} V^{(s)}_1(\tau, \tau_{z-1}, \tau_z) + p^1_{z-} V^{(s)}_1(\tau, \tau_z, \tau_{z-1}). \end{aligned} \tag{7.10}$$

Let $\bar{\Sigma}$ denote the performance index after the generation. The derivatives of $\bar{\Sigma}$ w.r.t. the decision variables $\bar{\tau}_i$, $i \in \{1,\dots,\bar{N}-1\}$ and $\bar{p}^s_{i-1}$, $\bar{p}^s_{i-}$, $i \in \{1,\dots,\bar{N}\}$ are determined by equalities analogous to those in Section 7.2. To use (7.6), a reindexing of the hamiltonian jumps is needed: $\Delta\bar{H}_i = \Delta H_i$ for $i < z$, $\Delta\bar{H}_i = \Delta H_{i-1}$ for $i > z$, and $\Delta\bar{H}_z = 0$. The values of derivatives, immediately before the generation and after it satisfy the relationships

$$\nabla_{\bar{p}^s_i}\bar{\Sigma} = \nabla_{p^s_i}\Sigma, \quad \nabla_{\bar{\tau}_i}\bar{\Sigma} = \nabla_{\tau_i}\Sigma \quad \text{for } i < z-1$$

$$\nabla_{\bar{p}^s_{i+1}}\bar{\Sigma} = \nabla_{p^s_i}\Sigma \quad \text{for } i > z \text{ or } i = z,\ z \notin K_s$$

$$\nabla_{\bar{p}^s_{i-}}\bar{\Sigma} = \nabla_{p^s_{i-}}\Sigma \quad \text{for } i < z,\ i \notin K_s$$

$$\nabla_{\bar{p}^s_{i+1-}}\bar{\Sigma} = \nabla_{p^s_{i-}}\Sigma \quad \text{for } i > z,\ i \notin K_s$$

$$\nabla_{\bar{\tau}_{i+1}}\bar{\Sigma} = \nabla_{\tau_i}\Sigma \quad \text{for } i > z.$$

It follows from (7.6)

$$\nabla_\tau \bar{\Sigma} \equiv \nabla_{\bar{\tau}_z} \bar{\Sigma} = (J_0(\tau, \tau_{z-1}) - J_0(\tau, \tau_z))\dot{u}(\tau)$$
$$+ (J_1(\tau, \tau_{z-1}) - J_1(\tau, \tau_z))\ddot{u}(\tau).$$

7.4. Efficiency of a flat generation

Let $\tau \in]\tau_{i-1}, \tau_i[$. Denote by $E(\tau)$ the efficiency of a flat generation at τ, understood as the difference of squared norms of gradients of the performance index immediately before the generation and after it, $E(\tau) = ||\nabla\bar{\Sigma}||^2 - ||\nabla\Sigma||^2$. This efficiency can be written as a sum

$$E(\tau) = e(\tau) + e_{i-1}(\tau) + e_i(\tau) \tag{7.11}$$

where e is the sum of squared derivatives of the performance index at the new node, and e_{i-1} and e_i are the increments of squared norms of the derivatives of the performance index w.r.t. the decision subvectors corresponding to the nodes τ_{i-1} and τ_i, respectively. The value of $\tau \in]\tau_{i-1}, \tau_i[$ should be chosen in such a way that the expression (7.11) is positive, and sufficiently large with respect to $||\nabla\Sigma||^2$.

Determine the efficiency of generation assuming that the decision vector before the generation and after it is inside the admissible set. The component related to the new node is given by

$$e(\tau) = (\nabla_\tau \bar{\Sigma})^2 + \sum_{s=0}^{1} \left\{ \begin{array}{ll} (\nabla_{\bar{p}^s_{i-}} \bar{\Sigma})^2, & i \notin \bar{K}_s \\ 0, & i \in \bar{K}_s \end{array} \right\} + \sum_{s=0}^{1} (\nabla_{\bar{p}^s_i} \bar{\Sigma})^2.$$

Calculate now the component corresponding to the node τ_{i-1}. If $i > 1$, then τ_{i-1} is a decision variable and

$$e_{i-1}(\tau) = (\nabla_{\bar{\tau}_{i-1}} \bar{\Sigma})^2 - (\nabla_{\tau_{i-1}} \Sigma)^2 + \sum_{s=0}^{1} \Big((\nabla_{\bar{p}^s_{i-1}} \bar{\Sigma})^2 - (\nabla_{p^s_{i-1}} \Sigma)^2 \Big).$$

If $i = 1$, then τ_{i-1} is not a decision variable and so

$$e_{i-1}(\tau) = \sum_{s=0}^{1} \big((\nabla_{\bar{p}^s_0} \bar{\Sigma})^2 - (\nabla_{p^s_0} \Sigma)^2 \big).$$

Finally, consider the term corresponding to τ_i. If $i < N$, then $\tau_i = \bar{\tau}_{i+1}$ is a decision variable and

$$e_i(\tau) = (\nabla_{\bar\tau_{i+1}}\bar\Sigma)^2 - (\nabla_{\tau_i}\Sigma)^2 + \sum_{s=0}^{1}\left\{\begin{array}{l}(\nabla_{\bar p^s_{i+1-}}\bar\Sigma)^2 - (\nabla_{p^s_{i-}}\Sigma)^2, i\notin K_s\\ (\nabla_{\bar p^s_{i+1}}\bar\Sigma)^2 - (\nabla_{p^s_i}\Sigma)^2, \quad i\in K_s\end{array}\right\}.$$

If $i = N$, then $\tau_i = \bar\tau_{i+1}$ is not a decision variable and

$$e_i(\tau) = \sum_{s=0}^{1}\Big((\nabla_{\bar p^s_{N+1-}}\bar\Sigma)^2 - (\nabla_{p^s_{N-}}\Sigma)^2\Big).$$

The efficiency E has a right and a left limit at every node τ_i, $i = 1,\dots,N-1$. These limits are equal to each other at τ_i, if $\Delta H_i = 0$.

7.5. On calculation of integrals

The integrals $J_s(\tau,\tau_i)$ and $J_s(\tau_i,\tau)$ are necessary to determine derivatives of Σ and $\bar\Sigma$, for $s = 0,1$, $i = 1,2,\dots,N$, $\tau\in[0,T]$. Define

$$h_{ji}(\tau) = -\int_\tau^{\tau_i} h(t)(t-\tau_i)^j\,\mathrm{d}t, \quad j = 0,1,2,3. \tag{7.12}$$

Denoting $\delta_i = \tau_i - \tau_{i-1}$, $\Delta_{ji}(\tau) = h_{ji}(\tau) - h_{ji}(\tau_{i-1})$ we have

$$\begin{aligned}
&h_{0,i-1}(\tau) = \Delta_{0i}(\tau)\\
&h_{1,i-1}(\tau) = \Delta_{1i}(\tau) + \delta_i h_{0,i-1}(\tau)\\
&h_{2,i-1}(\tau) = \Delta_{2i}(\tau) + \delta_i(\Delta_{1i}(\tau) + h_{1,i-1}(\tau))\\
&h_{3,i-1}(\tau) = \Delta_{3i}(\tau) + \delta_i(\Delta_{2i}(\tau) + 2h_{2,i-1}(\tau) - \delta_i h_{1,i-1}(\tau)).
\end{aligned} \tag{7.13}$$

To express $J_s(\tau_i,\tau)$ by $J_s(\tau,\tau_i)$ we use the identities

$$\begin{aligned}
&V_0(t,a,b) + V_0(t,b,a) = 1\\
&V_1(t,a,b) + V_1(t,b,a) = (b-a)V_0(t,a,b) + t - b.
\end{aligned}$$

Substituting $a = \tau$ and $b = \tau_i$, multiplying both sides by h and integrating from τ to τ_i, obtain

$$\begin{aligned}
&J_0(\tau_i,\tau) = h_{0i}(\tau) - J_0(\tau,\tau_i)\\
&J_1(\tau_i,\tau) = h_{1i}(\tau) + (\tau_i - \tau)J_0(\tau,\tau_i) - J_1(\tau,\tau_i).
\end{aligned} \tag{7.14}$$

The integrals $J_s(\tau, \tau_i)$ are computed by virtue of (7.2)

$$
\begin{aligned}
J_0(\tau, \tau_i) &= \frac{2h_{3i}(\tau)}{(\tau_i - \tau)^3} + \frac{3h_{2i}(\tau)}{(\tau_i - \tau)^2} \\
J_1(\tau, \tau_i) &= \frac{h_{3i}(\tau)}{(\tau_i - \tau)^2} + \frac{h_{2i}(\tau)}{\tau_i - \tau}.
\end{aligned} \tag{7.15}
$$

It is convenient to compute the derivatives of the performance index and the efficiency when solving backwards the adjoint equations. To determine the efficiency in the whole interval $[0, T]$, the nodes excluded, compute the integrals $J_s(\tau, \tau_i)$, $J_s(\tau, \tau_{i-1})$, $J_s(\tau_i, \tau)$ and $J_s(\tau_{i-1}, \tau)$, $s = 0, 1$ in the intervals $\tau_{i-1} \le \tau \le \tau_i$, successively for $i = N, N-1, \ldots, 1$. For a fixed i, perform the following steps.

(i) By numerical integration, calculate $h_{ji}(\tau)$, $j = 0, 1, 2, 3$, $\tau_{i-1} \le \tau \le \tau_i$.

(ii) Using (7.13) calculate $h_{j,i-1}(\tau)$, $j = 0, 1, 2, 3$, $\tau_{i-1} \le \tau \le \tau_i$.

(iii) Calculate $J_s(\tau, \tau_i)$, $s = 0, 1$, $\tau_{i-1} \le \tau < \tau_i$ from (7.15).

(iv) Calculate $J_s(\tau, \tau_{i-1})$, $s = 0, 1$, $\tau_{i-1} < \tau \le \tau_i$ from (7.15).

(v) Calculate $J_s(\tau_i, \tau)$ and $J_s(\tau_{i-1}, \tau)$, $s = 0, 1$, $\tau_{i-1} < \tau < \tau_i$ from (7.14).

If we only wish to determine the gradient of Σ in the current decision space, we skip steps (ii) and (iv), calculate $h_{ji}(\tau_{i-1})$, $j = 0, 1, 2, 3$ in step (i), calculate $J_s(\tau_{i-1}, \tau_i)$, $s = 0, 1$ in step (iii), and calculate $J_s(\tau_i, \tau_{i-1})$, $s = 0, 1$ in step (v).

8. Maximum range ascent of F-15 aircraft

We consider ascent of the F-15 aircraft from level flight at small altitude (5 m) with takeoff velocity (228.5 m/s) and initial mass 20244 kg, to level flight envelope. The goal is to maximize the range of flight in a given time. The longitudinal dynamics of the aircraft is described by a state equation $\dot{x} = f(x, u)$ with $x(t) \in \mathbb{R}^5$,

$$
f_1 = x_2 \sin x_3, \quad f_2 = \frac{\Theta - D}{x_4} - \sin x_3, \quad f_3 = \frac{u - \cos x_3}{x_2}
$$

$$
f_4 = \alpha\Theta, \quad f_5 = x_2 \cos x_3
$$

and with scaled state variables: altitude x_1, velocity x_2, flight path angle x_3, mass x_4, and range x_5. The vertical load factor u is the control signal.

The aircraft engines are in maximum afterburner at all conditions and the throttle setting is identically one. The thrust Θ and drag D are functions of state and control. The model of thrust, based on experimental data [12] reads

$$\Theta(x_1, M) = \sum_{s=1}^{5} \Theta_{2s-1}(x_1, M) \exp(\Theta_{2s}(x_1, M))$$

where $\Theta_p,\ p = 1, \ldots, 10$ are quadratic polynomials of two variables. The Mach number is a function of altitude and velocity

$$M = \frac{x_2}{\sqrt{a_3 x_1^3 + a_2 x_1^2 + a_1 x_1 + a_0}}.$$

The model of drag [10, 14] is given by

$$\begin{aligned}
D &= d_1 + (u x_4)^2 d_2 \\
d_1 &= C(M)\, x_2^2 \mathrm{e}^{q(x_1)}, \quad d_2 = K(M)\, x_2^{-2} \mathrm{e}^{-q(x_1)} \\
q(x_1) &= q_0(\mathrm{e}^{z(x_1)} - 1) + q_1 x_1, \quad z(x_1) = z_4 x_1^4 + z_3 x_1^3 + z_2 x_1^2 + z_1 x_1 \\
C(M) &= \frac{c_{14} M^4 + c_{13} M^3 + c_{12} M^2 + c_{11} M + c_{10}}{M^4 + c_{23} M^3 + c_{22} M^2 + c_{21} M + c_{20}} \\
K(M) &= \frac{k_{14} M^4 + k_{13} M^3 + k_{12} M^2 + k_{11} M + k_{10}}{M^5 + k_{24} M^4 + k_{23} M^3 + k_{22} M^2 + k_{21} M + k_{20}}.
\end{aligned}$$

The flight range $x_5(T)$ is to be maximized, for a fixed horizon $T = 235$ s. Thus

$$S(u) = -x_5(T).$$

The terminal conditions read

$$h_1(x(T)) = (\Theta - D)|_{t=T, u=1} = 0, \quad h_2(x(T)) = x_3(T) = 0.$$

The dynamic pressure must not exceed a given limit during the whole flight

$$\begin{aligned}
g(x) &= x_2 - Q(x_1) \leq 0 \\
Q(x_1) &= b \exp(-\tfrac{1}{2} q(x_1)), \quad b > 0.
\end{aligned}$$

The problem is reformulated using penalty functions

$$\dot{y} = \tfrac{1}{2}\left((x_2 - Q(x_1))_+\right)^2,\ y(0) = 0$$

$$S_\rho(u) = -x_5(T) + \tfrac{1}{2}\rho_1\ (\Theta - D)^2\big|_{t=T,u=1} + \tfrac{1}{2}\rho_2 x_3(T)^2 + \rho_3 y(T)$$

$$\rho_1, \rho_2, \rho_3 > 0.$$

A straightforward computation shows that the hamiltonian (2.8) is maximized by

$$u = \frac{\psi_3 x_2}{2\psi_2 x_4 K}\,\mathrm{e}^q.$$

To seek the optimal control, we use the technique of cubic Hermite polynomials and flat generations described in Section 7. Thus, all control arcs are interior. Continuity of control and its derivative is required at all internal nodes $\tau_1, \ldots, \tau_{N-1}$, that is, control approximations are smooth and $K_0 = K_1 = \{1, \ldots, N-1\}$. In every generation, only one node is added at a local maximizer of efficiency.

The optimization is started with penalty coefficients $\rho_1 = 0.01$, $\rho_2 = 1$ and $\rho_3 = 0.0001$. The initial control structure has only two nodes $\tau_0 = 0$ and $\tau_1 = T$, and one procedure P_1. The corresponding control is identically equal to one. There are four decision variables p_0^0, p_0^1, p_{1-}^0, p_{1-}^1. The control obtained after a period of gradient optimization, together with scaled efficiency of a potential flat generation is shown in Figure 8.1. There are two inherited nodes (blank circles) and one newly generated (filled circle) located at the maximum of efficiency (dashed line). The dimension of the decision space after the generation is equal to 7. Further optimization leads to the situation shown in Figure 8.2 where the efficiency exhibits two maxima, one at a node τ_1 and one in $]\tau_1, \tau_2[$. To stay within the theoretical framework of Section 7, we choose the second maximum for the next generation.

The structural evolution continues until the number of nodes reaches 12. All the penalty coefficients are then increased to 10. The final approximation resulting from optimization with the new penalty coefficients is shown in Figure 8.3 along with the exact optimal control obtained with an indirect method (dashed line). Note that the discrepancy between these curves may be arbitrarily reduced by adding more nodes. The final values of the terminal constraint functions $h_1(x(T))$ and $h_2(x(T))$ are of order 10^{-7} and 10^{-4}, respectively. The function $g(x(t))$ plotted against time in Figure 8.4 indicates that there are two state-constrained arcs in the optimal solution.

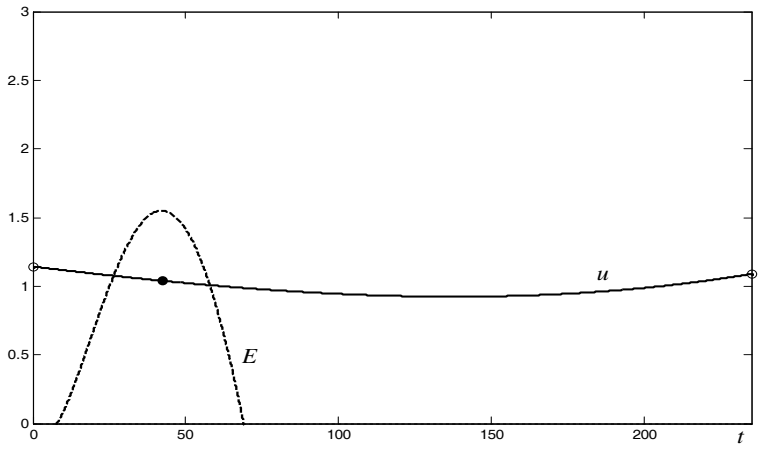

Figure 8.1. First generation

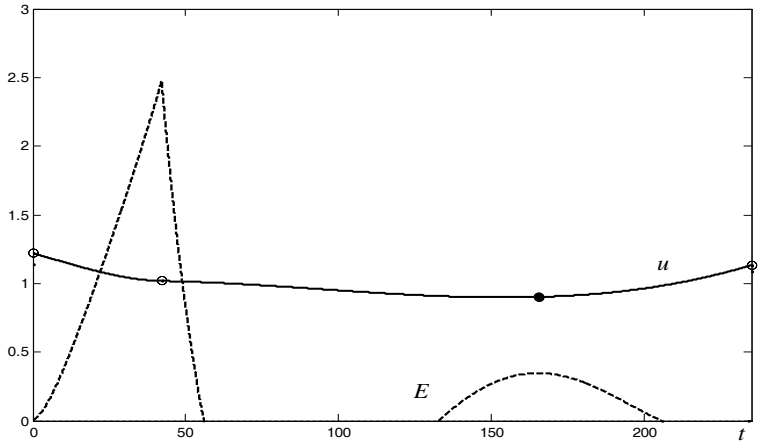

Figure 8.2. Second generation

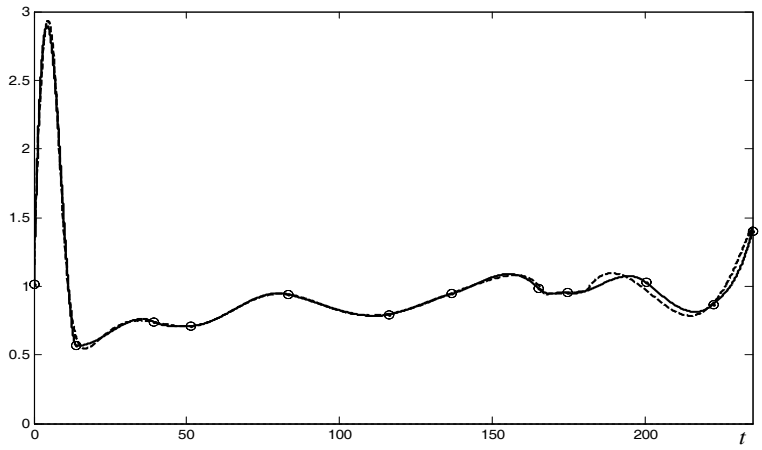

Figure 8.3. Final approximation of optimal control

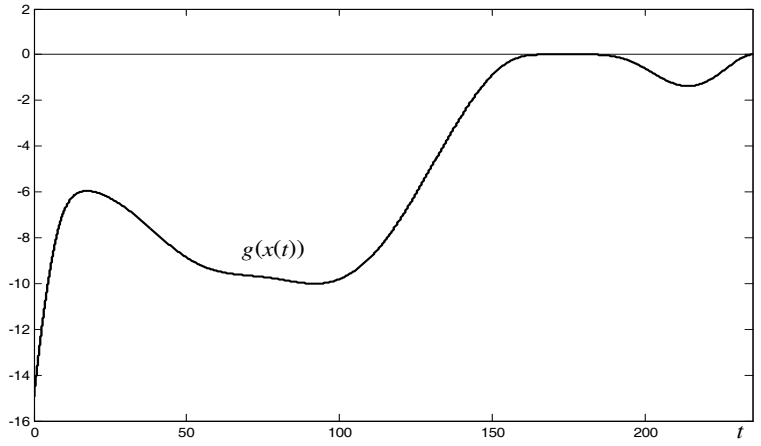

Figure 8.4. State constraint function for final approximation

Observe that the technique applied in this example has not directly produced the optimal control structure. It can be shown that even admitting the generation of candidate constrained arcs would not necessarily lead to the optimal structure, other than in the example of Section 6. This is explained by the fact that here the optimal control is continuous, and so the efficiencies of flat generations of interior arcs and candidate constrained arcs are similar. This difficulty cannot be resolved in the framework of penalty methods, and needs an approach with an explicit representation of pathwise state constraints.

9. Conclusions

The general idea of the MSE approach to control and state constrained problems of dynamic optimization has been presented, together with two computational implementations using spike and flat generations. Although the method of spike generations has been shown effective on a rather simple example, that result is in agreement with wider experience related to problems with discontinuous optimal controls. The flat generations and interval cubic control representations have been tested on a more complex problem where a good approximation of the optimal solution has been obtained in a computationally economical way.

References

[1] J.T. Betts, *Survey of numerical methods for trajectory optimization*, Journal of Guidance, Control and Dynamics **21** (2) (1998), 193–207.

[2] J.T. Betts, Practical Methods for Optimal Control Using Nonlinear Programming, SIAM (2001).

[3] H.G. Bock and K.J. Plitt, A multiple shooting algorithm for direct solution of optimal control problems, IFAC 9th World Congress, Budapest, Hungary 1984.

[4] R. Bulirsch, F. Montrone and H.J. Pesch, *Abort landing in the presence of a windshear as a minimax optimal control problem, part 2: multiple shooting and homotopy*, Journal of Optimization Theory and Applications **70** (1991), 223–254.

[5] A. Cervantes and L.T. Biegler, *Optimization Strategies for Dynamic Systems*, Encyclopedia of Optimization **4** 216–227, C. Floudas and P. Pardalos (eds.), Kluwer 2001.

[6] C.R. Hargraves and S.W. Paris, *Direct trajectory optimization using nonlinear programming and collocation*, Journal of Guidance **10** (4) (1987), 338–342.

[7] C.Y. Kaya and J.L. Noakes, *Computational algorithm for time-optimal switching control*, Journal of Optimization Theory and Applications **117** (1) (2003), 69–92.

[8] J. Kierzenka and L.F. Shampine, *A BVP solver based on residual control and the Matlab PSE*, ACM Transactions on Mathematical Software **27** (3) (2001), 299–316.

[9] D. Kraft, *On converting optimal control problems into nonlinear programming problems*, Computational Mathematics and Programming, **15** (1985), 261–280.

[10] R.R. Kumar and H. Seywald, *Should controls be eliminated while solving optimal control problems via direct methods*? Journal of Guidance, Control, and Dynamics **19** (2) (1996), 418–423.

[11] H. Maurer, C. Büskens, J.-H.R. Kim and C.Y. Kaya, *Optimization methods for the verification of second order sufficient conditions for bang-bang controls*, Optimal Control Applications and Methods **26** (2005), 129–156.

[12] J. Miller, A. Korytowski and M. Szymkat, *Two-stage construction of aircraft thrust models for optimal control computations*, Submitted to Optimal Control Applications and Methods.

[13] M. Pauluk, A. Korytowski, A. Turnau and M. Szymkat, Time optimal control of 3D crane, Proc. 7th IEEE MMAR 2001, Międzyzdroje, Poland, August 28–31 (2001), 927–932.

[14] H. Seywald, *Long flight-time range-optimal aircraft trajectories*, Journal of Guidance, Control, and Dynamics **19** (1) (1996), 242–244.

[15] H. Shen and P. Tsiotras, *Time-optimal control of axi-symmetric rigid spacecraft with two controls*, Journal of Guidance, Control and Dynamics **22** (1999), 682–694.

[16] H.R. Sirisena, *A gradient method for computing optimal bang-bang control*, International Journal of Control **19** (1974), 257–264.

[17] B. Srinivasan, S. Palanki and D. Bonvin, *Dynamic optimization of batch processes, I. Characterization of the nominal solution*, Computers and Chemical Engineering **27** (1) (2003), 1–26.

[18] O. von Stryk, User's guide for DIRCOL - a direct collocation method for the numerical solution of optimal control problems, Ver. 2.1, Technical University of Munich 1999.

[19] M. Szymkat, A. Korytowski and A. Turnau, *Computation of time optimal controls by gradient matching*, Proc. 1999 IEEE CACSD, Kohala Coast, Hawai'i, August 22–27 (1999), 363–368.

[20] M. Szymkat, A. Korytowski and A. Turnau, Variable control parameterization for time-optimal problems, Proc. 8th IFAC CACSD 2000, Salford, U.K., September 11–13, 2000, T4A.

[21] M. Szymkat, A. Korytowski and A. Turnau, Extended variable parameterization method for optimal control, Proc. IEEE CCA/CCASD 2002, Glasgow, Scotland, September 18–20, 2002.

[22] M. Szymkat and A. Korytowski, Method of monotone structural evolution for control and state constrained optimal control problems, European Control Conference ECC 2003, University of Cambridge, U.K., September 1–4, 2003.

[23] J. Wen and A.A. Desrochers, *An algorithm for obtaining bang-bang control laws*, Journal of Dynamic Systems, Measurement, and Control **109** (1987), 171–175.

Received 28 March 2006