Optimal control of ordinary differential equations

Frédéric J. Bonnans

Optimal control of ordinary differential equations[1]

J. Frédéric Bonnans[2]

August 11, 2006

[1] Lecture notes, CIMPA School on Optimization and Control, Castro Urdiales, August 28 - September 8, 2006.

[2] INRIA-Futurs and Centre de Mathématiques Appliquées (CMAP), Ecole Polytechnique, 91128 PALAISEAU Cedex, France. Email: Frederic.Bonnans@inria.fr.

Contents

s

Foreword

These notes give an introduction to the theory of optimal control of ordinary differential equations, and to some related algorithmic questions. We put the emphasis on the question of well-posedness (or not) of a local minimum.

For a system of nonlinear equations the main tool for checking well-posedness of a local solution is the implicit function theorem. We are sometimes able to reduce optimality conditions to this setting. However, there are situations when we cannot, and then several concepts of well-posedness may be used, based on the stability or uniqueness of local minimizers, solutions of optimality conditions, at different rates (strong regularity, strong stability, Hölder stability, etc.) In addition a number of functional analysis tools are needed: characterization of dual spaces, separation theorems, convex analysis.

The point of view taken in these notes is, starting from "concrete situations" (i.e. optimal control problems), to introduce gradually the needed theoretical concepts that are needed for either a numerical resolution or a sensitivity analysis of the problem. So in some sense we take the point of view of a (mathematical) engineer, but without being afraid of using abstract tools if necessary. Two chapters have been written at the occasion of the course, and the notes include also the papers [6] and [9], coauthored with A. Hermant and J. Laurent-Varin, respectively. Let me mention also the related papers [3, 8, 7].

These notes are in some sense a continuation of the book [10] written with A. Shapiro, devoted to the sensitivity analysis for general optimization problems. Many papers have since clarified the link between optimization theory and optimal control problems. A classical and still useful reference is Ioffe and Tihomirov [17]. A more recent book on optimal control is Milyutin and Osmolovskii [21].

I thank Eduardo Casas and Michel Théra for giving me the opportunity of presenting this material, and wish that these notes will motivate students for entering in this field and obtaining new results. All remarks are welcome.

s

Chapter 1

Linear quadratic control and control constrained problems

Linear quadratic optimal control problems occur in several situations:
(i) linearization of the dynamics around a stationary point (where the derivative is zero) and stabilization around that point
(ii) study of the optimality conditions of a critical point of an optimal control problem
(iii) sensitivity analysis of a local solution of an optimal control problem.

The first section of this chapter we try first present the theory of critical points, including the shooting formulation and the Riccati equation. Then we relate the notion of Legendre form to the case when we have to solve a minimization problem.

In the second section we present a no-gap theory of second-order optimality conditions as well as a sensitivity analysis, in an abstract framework: nonlinear cost function and polyhedric constraints. We show how this applies to linear quadratic optimal control problems with bound constraints.

In the third section we study the case of nonlinear local constraint on the control, of the form

$$U = \{u \in \mathbb{R}^m;\ g_i(u) \leq 0,\ i = 1, \dots, r\}, \tag{1.0.1}$$

and functions g_i are convex continuous. Then the curvature of these functions has to be taken into account.

Notations We denote the Euclidean norm of $x \in \mathbb{R}^n$ by $|x|$. The transposition of a matrix A is $A^\top$.

1.1 Unconstrained problems

1.1.1 Critical points of quadratic functionals

Consider the following dynamical system

$$\dot{y}_t = A_t y_t + B_t u_t, \quad t \in [s, T]; \quad y_s = x, \tag{1.1.2}$$

where $s \leq T$, and matrices A_t et B_t, measurable functions of time, are of size $n \times n$ and $n \times m$ respectively, and essentially bounded. Denote the control and state spaces by

$$\mathcal{U} := L^2(0, T, \mathbb{R}^m); \quad \mathcal{Y} := H^1(0, T, \mathbb{R}^n).$$

We know that with each $u \in \mathcal{U}$ is associated a unique solution in $\mathcal{Y}$ of (1.1.2), called the state and denoted $y(u)$. Define the criterion

$$F(u, y) := \tfrac{1}{2} \int_s^T [y_t \cdot C_t y_t + 2u_t \cdot D_t y_t + u_t \cdot R_t u_t] \,\mathrm{d}t + \tfrac{1}{2} y_T \cdot M y_T. \tag{1.1.3}$$

The matrices C_t, D_t and R_t are measurable, essentially bounded functions of time of appropriate dimension. The function F is therefore well-defined $\mathcal{U} \times \mathcal{Y} \to \mathbb{R}$. Denote

$$f(u) := F(u, y(u)).$$

Being quadratic and continuous, f has a gradient and the latter is an affine function of u. We say that u is a critical point of f if $Df(u) = 0$.

In order to compute the gradient, let us introduce the *adjoint state* (or *costate*) equation

$$-\dot{p}_t = A_t^\top p_t + C_t y_t + D_t^\top u_t, \quad t \in [s, T]; \quad p_T = M y_T. \tag{1.1.4}$$

The costate $p \in \mathcal{Y}$ associated with the control $u \in \mathcal{U}$ is defined as the unique solution of (1.1.4), where $y = y(u)$.

Remark 1.1 A general method for finding the costate equation is as follows: let

$$L(u, y, p) := F(u, y) + \int_s^T p(t) \cdot (A_t y_t + B_t u_t - \dot{y}_t) \mathrm{d}t$$

denote the Lagrangian associated with the cost function F and state equation (1.1.2). Then the costate equation is obtained by setting to zero the derivative of the Lagrangian with respect to the state.

Proposition 1.2 *The quadratic mapping $u \to f(u)$ is of class C^∞ from $\mathcal{U}$ to $\mathbb{R}$, and its gradient satisfies*

$$Df(u)_t = B_t^\top p_t + R_t u_t + D_t y_t, \quad t \in [0, T]. \tag{1.1.5}$$

where y and p are the state and costate associated with u.

The stationary points of f are therefore characterized by the (algebraic-differential) two-point boundary value problem (TPBVP)

$$\begin{aligned}
\dot{y}_t &= A_t y_t + B_t u_t, & t \in [s, T]; \quad y_0 = x, && (1.1.6)\\
-\dot{p}_t &= A_t^\top p_t + C_t y_t + D_t^\top u_t, & t \in [s, T]; \quad p_T = M y_T, && (1.1.7)\\
0 &= B_t^\top p_t + R_t u_t + D_t y_t. & && (1.1.8)
\end{aligned}$$

In the sequel we will often assume R_t *uniformly invertible:*

$$\exists\, \alpha > 0; \ \ |R_t v| \geq \alpha |v|, \quad \text{for all } \ v \in \mathbb{R}^m, \quad t \in (0, T). \tag{1.1.9}$$

Eliminating then the control variable from relation (1.1.8) we obtain then that the triple (u, y, p) is solution of (1.1.6)-(1.1.8) iff (y, p) is solution of the differential two-point boundary value problem

$$\begin{aligned} \dot{y}_t &= (A_t - B_t R_t^{-1} D_t) y_t - B_t R_t^{-1} B_t^\top p_t, & t \in [s, T]; \qquad (1.1.10)\\ -\dot{p}_t &= (C_t - D_t^\top R_t^{-1} D_t) y_t + (A_t^\top - D_t^\top R_t^{-1} B_t^\top) p_t, & t \in [s, T]; \qquad (1.1.11)\\ y_s &= x, \quad p_T = M y_T. & (1.1.12) \end{aligned}$$

Equations (1.1.10)-(1.1.12) may be rewritten as

$$\Psi(y, p) = 0$$

(by putting all expressions on the right-hand-side), the mapping $\Psi(y, p)$ being linear and continuous

$$\mathcal{Y} \times \mathcal{Y} \to L^2(0, T, \mathbb{R}^n) \times L^2(0, T, \mathbb{R}^n) \times \mathbb{R}^{2n}.$$

The only nonhomogeneous term is due to the given initial point x. Therefore the set of stationary points is a closed affine space, and there exists at most a stationary point iff the above system, when $x = 0$, has the only solution $y = 0$ and $p = 0$.

1.1.2 Shooting function and Hamiltonian flow

Let us introduce the *shooting function*

$$S_{s,T} : \mathbb{R}^n \to \mathbb{R}^n; \quad q \mapsto p_T - M y_T,$$

where $(y, p) \in \mathcal{Y} \times \mathcal{Y}$ is solution of (1.1.10)-(1.1.11), with initial condition (x, q) at time s. We can easily see that

Lemma 1.3 *Assume that* (1.1.9) *holds. Then the control function u is a stationary point of f iff the associated costate p is such that p_s is a zero of S.*

The problem of finding the critical points of f reduces therefore to the one of solving a linear equation in $\mathbb{R}^n$.

Denote by $\Phi_{s,t}$ the "flow" associated with (1.1.10)-(1.1.11). In other words, $\Phi_{s,t}$ associates with (x, q) the value (y_t, p_t) obtained by integrating (1.1.10)-(1.1.11) over $[s, t]$. Denote by $\Phi^y_{s,t}$ and $\Phi^p_{s,t}$ the n first and last components of $\Phi_{s,t}$. We have

$$\frac{\mathrm{d}}{\mathrm{d}t} \Phi_{s,t} = \begin{pmatrix} A_t - B_t R_t^{-1} D_t & -B_t R_t^{-1} B_t^\top \\ -C_t + D_t^\top R_t^{-1} D_t & -A_t^\top + D_t^\top R_t^{-1} B_t^\top \end{pmatrix} \Phi_{s,t} \qquad (1.1.13)$$

The *Hamiltonian function*: $\mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, asociated with the original system, is

$$H(u, y, p, t) := \tfrac{1}{2}(y \cdot C_t y + 2u \cdot D_t y + u \cdot R_t u) + p \cdot (A_t y + B_t u). \qquad (1.1.14)$$

By substituting $u = -R_t^{-1}(B_t^\top p + D_t y)$, we obtain the reduced Hamitonian

$$\mathcal{H}(y, p, t) := \tfrac{1}{2} y \cdot C_t y + p \cdot A_t y - \tfrac{1}{2}(B_t^\top p + D_t y) R_t^{-1} (B_t^\top p + D_t y). \qquad (1.1.15)$$

The matrix in (1.1.13) denoted by $M_t^{\mathcal{H}}$, is called the Hamiltonian matrix associated with the critical point problem. It satisfies the relation

$$M_t^{\mathcal{H}} = \begin{pmatrix} \frac{\partial^2 \mathcal{H}(y,p,t)}{\partial p \partial y} & \frac{\partial^2 \mathcal{H}(y,p,t)}{\partial y \partial y} \\ \frac{\partial^2 \mathcal{H}(y,p,t)}{\partial p \partial p} & \frac{\partial^2 \mathcal{H}(y,p,t)}{\partial y \partial p} \end{pmatrix} \tag{1.1.16}$$

We may write the shooting equation under the form

$$\Phi^p_{s,T}(x, p_0) = M\Phi^y_{s,T}(x, p_0). \tag{1.1.17}$$

Since $\Phi_{s,t}$ is linear, this can be rewritten as

$$\Phi^p_{s,T}(0, p_0) - M\Phi^y_{s,T}(0, p_0) = -\Phi^p_{s,T}(x, 0) + M\Phi^y_{s,T}(x, 0). \tag{1.1.18}$$

Lemma 1.4 *Assume that* (1.1.9) *holds. Then when s is close to T, $S_{s,T}$ is invertible, i.e., there exists a unique stationary point of f.*

Proof. It is easy to check that $S_{s,T}$ is a continuous function of s, and $S_{s,T}(q) \to q - Mx$ when $s \uparrow T$. Therefore $S_{s,T}$ is invertible for s close to T. The conclusion follows. ■

Definition 1.5 *We say that $s < T$ is a* conjugate point *of T if $S_{s,T}$ is not invertible. Denote by $\mathcal{T}$ the set of times $s < T$ which are not conjugate, i.e., for which $S_{s,T}$ is invertible.*

Obviously $\mathcal{T}$ is an open set. If all matrices are (real) analytic functions of time (i.e., locally expandable in power series), then the shooting function is also an analytic function, and has for each s, at most finitely many zeroes. To see this, observe that the determinant of the Jacobian of the shooting function is a nonzero analytic function of time, so that it may have only a finite number of zeroes over a bounded interval of $\mathbb{R}$. Now $\mathcal{T}$ is the set of times for which this determinant does not vanish.

We say that (y, p) is a singular solution of the two-point boundary value problem (1.1.10)-(1.1.12) if it is a nonzero solution of (1.1.10)-(1.1.12) with $x = 0$. We can express the fact that a time is a conjugate point using singular solutions.

Lemma 1.6 *A time τ is a conjugate point of T iff there exists a singular solution of* (1.1.10)-(1.1.12).

Proof. We have that τ is a conjugate point iff the shooting equation has a nonzero solution q with zero initial condition x. Integrating (1.1.10)-(1.1.12) with initial condition $(0, q)$, we derive the conclusion. ■

1.1.3 Riccati equation

Let $s \in \mathcal{T}$. Since $S_{s,T}$ is affine, with right hand side linear function of x, p_s is a linear mapping of x. So we may write

$$p_s = P_s x,$$

where P_s is a square matrix of size n. For all $\sigma \in \mathcal{T} \cap]s, T[$, (y, p) solution of (1.1.10)-(1.1.12), restricted to $[\sigma, T]$, is a stationary point with initial condition y_σ, and so

$$p(t) = P_\sigma y(t).$$

By standard results on ordinary differential equations, S_t and hence, P_t are differentiable functions of t. Substituting Py to p in (1.1.11), and factorizing by y_t, we get

$$0 = P_t \dot{y}_t + \left[\dot{P}_t + (C_t - D_t^\top R_t^{-1} D_t) + (A_t^\top - D_t^\top R_t^{-1} B_t^\top) P_t \right] y_t, \quad t \in \mathcal{T}. \tag{1.1.19}$$

Using the expression of $\dot{y}_t$ in (1.1.10) with $p_t = P_t y_t$, we obtain

$$0 = \left[\dot{P}_t + P_t A_t + A_t^\top P_t + C_t - (P_t B_t + D_t^\top) R_t^{-1} (B_t^\top P_t + D_t) \right] y_t, \quad t \in \mathcal{T}. \tag{1.1.20}$$

Since this must be satisfied for all possible values of y_t (take $s = t$ and then $y_t = x$ is arbitrary) we obtain that P is solution of the Riccati equation

$$0 = \dot{P}_t + P_t A_t + A_t^\top P_t + C_t - (P_t B_t + D_t^\top) R_t^{-1} (B_t^\top P_t + D_t) \quad t \in \mathcal{T}. \tag{1.1.21}$$

Denote by τ_0 the largest conjugate point (i.e., the first starting backwards from T). If no conjugate point exist, we set $\tau_0 = -\infty$.

Lemma 1.7 *The Riccati operator P_t (defined on $\mathcal{T}$) is symmetric.*

Proof. (i) We have that P_t is symmetric on $(\tau_0, T]$, since the final condition is symmetric, and the derivative is symmetric on the subspace of symmetric matrices[1].
(ii) We approximate the data by convolution with a smooth kernel (so as to obtain C^∞ data), and then by polynomials. In that case $\Phi_{s,T}$ is an analytic function of time, and hence the solution p_0 of (1.1.18)) too. Since each column of P_s is the solution of (1.1.18)) when w is one basis vector, we obtain that P_s is also an analytic function of time. Being symmetric for values close to T, it must be symmetric everywhere. ■

Lemma 1.8 *Assume that τ_0 is finite. Then the Riccati equation* (1.1.21)*, with final condition $P_T = M$, has a unique solution over $(\tau_0, T]$, that if τ_0 is finite, satisfies $\lim_{t\downarrow\tau_0} \|P_t\| = +\infty$.*

Proof. It is a standard result of the theory of ODEs that, since (1.1.21) is a differential equation with locally Lipschitz dynamics, it has a unique solution over a segment of the form $(\tau_1, T]$, and if τ_1 is finite, $\lim_{t\downarrow\tau_0} \|P_t\| = +\infty$.

Since (1.1.21) has a solution over $\mathcal{T}$, we obtain that $\tau_1 \leq \tau_0$. If $\tau_0 = -\infty$ the conclusion follows. Otherwise assume that $\limsup_{t\downarrow\tau_0} \|P_t\| < +\infty$. Then (1.1.21) would have a solution over $[\tau_1, T]$. But then $p_t = P_t y_t$ is solution of the two point boundary value problem over $[\tau_1, T]$, for any initial condition x. This contradicts the non invertibility of the shooting mapping. ■

Remark 1.9 Let τ be a (necessarily isolated) conjugate point. Then

$$\lim_{s\to\tau^\pm} \|P_s\| = +\infty$$

otherwise P_τ would be well-defined, and $p = P_\tau x$ would provide a solution of the shooting equations, for arbitrary x, in contradiction with the definition of a conjugate point.

[1] So that the Riccati equation may be viewed as an equation over the subspace of symmetric matrices.

1.1.4 Expression of the critical value

With every critical point u at time s is associated he critical value $f(u)$. The latter has, when $s \in \mathcal{T}$, a simple expression involving P_s. Since

$$y_T \cdot My_T = y_T \cdot p_T = x \cdot p_s + \int_s^T (\dot{y}_t \cdot p_t + y_t \cdot \dot{p}_t)\,\mathrm{d}t \tag{1.1.22}$$

we obtain, combining with (1.1.2) et (1.1.4), that

$$y_T \cdot My_T = x \cdot p_s + \int_s^T \left(p_t \cdot B_t u_t - y_t \cdot C_t y_t - y_t \cdot D_t^\top u_t\right)\mathrm{d}t \tag{1.1.23}$$

Using (1.1.23) and (1.1.8) for evaluating the critical value as a function of x, denoted $F(x)$, we obtain

$$f(x) = x \cdot p_s. \tag{1.1.24}$$

In particular, if $s \in \mathcal{T}$, then

$$f(x) = x \cdot P_s x. \tag{1.1.25}$$

Consequently, the nonnegativity f is equivalent to the positive semidefiniteness of P_s.

1.1.5 Legendre forms and minima of quadratic functions

We consider in this section the problem of minimizing the quadratic cost f. A local minimum $\bar{u}$ satisfies the second-order necessary condition[2]

$$Df(\bar{u}) = 0 \quad \text{and } D^2 f(\bar{u}) \succeq 0. \tag{1.1.26}$$

Since $D^2 f(\cdot)$ is constant, this means that $\bar{u}$ is a stationary point of f and that f is convex. In that case we know that critical points coincide with global minima.

The next step is to study the well-posedness of local minima. The latter may be defined as the invertibility of $D^2 f(\bar{u})$, so the the implicit function theorem applies to a smooth perturbation of the critical point equation $Df(\bar{u}) = 0$. The following is proved in [10, Lemma 4.124].

Lemma 1.10 *Assume that $D^2 f(\bar{u}) \geq 0$. Then $D^2 f(\bar{u})$ is invertible iff it is uniformly positive, in the following sense: there exists $\alpha > 0$ such that*

$$D^2 f(\bar{u})(h,h) \geq \alpha \|h\|^2. \tag{1.1.27}$$

Since f is quadratic, its Hessian is uniformly positive iff f satisfies the following quadratic growth condition

Definition 1.11 *Let u be a stationary point of f. We say that the* quadratic growth *property is satisfied if there exists $\alpha > 0$ such that $f(u) \geq f(\bar{u}) + \alpha \|u - \bar{u}\|_{\mathcal{U}}^2$, for all u in some neighborhood of $\bar{u}$.*

Let us now relate these notions to the one of Legendre forms [10, Sections 3.3.2 et 3.4.3].

[2] If Q is a quadratic form, $Q \succeq 0$ means that Q is nonnegative, i.e., $Q(x) \geq 0$ for all x.

Definition 1.12 *Let X be a Hilbert space. We say that $Q : X \to \mathbb{R}$ is a* Legendre form *if it is a weakly lower semi continuous (w.l.s.c.) quadratic form over X, such that, if $y^k \to y$ weakly in X and $Q(y^k) \to Q(y)$, then $y^k \to y$ strongly.*

Set $w^k := y^k - y$. Using

$$Q(y^k) = Q(y) + DQ(y)w^k + Q(w^k),$$

and since $DQ(y)w^k \to 0$ as $w^k \to 0$ weakly, we have that Q is a Legendre form iff for any sequence w^k weakly converging to 0, $Q(w^k) \to 0$ iff $w^k \to 0$ strongly.

The following examples apply easily to the quadratic costs for optimal control problems:

Example 1.13 Let Q be a quadratic form over a Hilbert space X.
(i) Let $Q(y) = \|y\|^2$ be the square of the norm. Then obviously $Q(w^k) \to 0$ iff $w^k \to 0$ strongly. Therefore Q is a Legendre form.
(ii) Assume that Q is nonnegative, and $y \mapsto \sqrt{Q(y)}$ is a norm equivalent to the one of X. Then (the weak topology being invariant by under a new equivalent norm) Q is a Legendre form.
(iii) Assume that $Q(y) = Q_1(y) + Q_2(y)$, where Q_1 is a Legendre form, and Q_2 is weakly continuous. Then Q is a Legendre form.

The notions of quadratic growth and Legendre form are related in the following way:

Lemma 1.14 *Let $Q : X \to \mathbb{R}$ be a Legendre form, and C a closed convex cone of X. Then the two statements below are equivalent:*

$$Q(h) > 0, \quad \textit{for all } \ h \in C \setminus \{0\} \tag{1.1.28}$$

$$\exists\, \alpha > 0; \quad Q(h) \geq \alpha \|h\|^2, \quad \textit{for all } \ h \in C. \tag{1.1.29}$$

Lemma 1.15 *The functional f is w.l.s.c. over $\mathcal{U}$ iff $R_t \succeq 0$ a.e., and $D^2 f$ is a Legendre form iff there exists $\alpha > 0$ such that $R_t \succeq \alpha I_d$ a.e.*

Proof. (i) We can decompose f as $f = f_1 + f_2$, where f_1 is the part that does not depend on the state (obtained by setting C_t and D_t to 0) and $f_2 = f - f_1$. It is easily checked that f_2 is weakly continuous. Therefore f is w.l.s.c. iff f_1 is w.l.s.c.
(ii) If $R_t \succeq 0$ a.e., then f_1 being convex and continuous, is w.l.s.c.; If not, it is easily shown that there exists $\beta > 0$ and a measurable set $I \subset (s, T)$ such that

$$h \cdot R_t h \leq -\beta \|h\|^2, \quad \text{for all } \ h \in \mathbb{R}^m, \ \textit{a.e. } t \in I. \tag{1.1.30}$$

Let $\mathcal{U}_I$ be the subset of $\mathcal{U}$ of functions that are zero a.e. outside I. Since $\mathcal{U}_I$ is infinite dimensional, there is an orthonormal sequence u^k in $\mathcal{U}_I$. We have that $u^k \to 0$ weakly in $\mathcal{U}$, whereas

$$\limsup_k f(u^k) = \limsup_k f_1(u^k) \leq -\beta < 0 = f(0). \tag{1.1.31}$$

Therefore f is w.l.s.c. iff $R_t \succeq 0$ a.e.
(iii) If $R_t \succeq \alpha I_d$ a.e., then $\sqrt{f_1}$ defines a norm equivalent to the one of $\mathcal{U}$, and since f_2 is weakly continous, $D^2 f$ is a Legendre form (see case (iii) example 1.13).

Otherwise there exists an orthonormal sequence u^k such that $a := \limsup f_1(u^k) \leq 0$. Since $u^k \to 0$ weakly, either $a < 0$ contradicting the weak l.s.c. of f_1, or $a = 0$ so that $f_1(u^k) \to f_1(0)$, but u^k does not strongly converge to 0, contradicting the definition of the Legendre form. ■

1.1.6 Spectral analysis

In this section, for simplicity, we assume that all matrices in the definition of the quadratic problem are constant over time, and that R is positive definite. We can make a change of variable on $\mathbb{R}^m$,

$$v = Lu$$

such that $L^\top L = R$, and then

$$|v|^2 = u \cdot Ru.$$

The corresponding change of variables on $\mathcal{U}$ has the effect of reducing R to identity. So in the sequel we assume that R is the identity matrix. Also for simplicity we assume that $D = 0$. So we may write $f = f_1 + f_2$, with

$$f_1(u) = \tfrac{1}{2} \int_s^T |u_t|^2 \mathrm{d}t = \|u\|^2 \tag{1.1.32}$$

and

$$f_2(u) = \tfrac{1}{2} \int_s^T y_t \cdot C_t y_t \mathrm{d}t + \tfrac{1}{2} y_T \cdot M y_T \tag{1.1.33}$$

Let H_s denote the Hessian of f_2, and Q_s denote the associated quadratic form.

If X, Y are Banach spaces, an operator $A \in L(X, Y)$ is said to be compact if the image of B_X (unit ball) by A has a compact closure. The following lemma is classical (see e.g. Dunford and Schwartz [13]).

Lemma 1.16 *The operator H_s is selfadjoint and compact. Consequently, there is an orthonormal basis of $\mathcal{U}_s$ composed of eigenvectors of H_s.*

Proof. The first statement is a consequence of the compactness of the mapping $\mathcal{U}_s \to \mathcal{Y}_s$, $v \mapsto z$, where z is the unique solution of the linearized equation

$$\dot{z} = Az + Bv; \quad z(s) = 0. \tag{1.1.34}$$

The second statement comes from the well-known theory of compact operators; see e.g., Balakrishnan [2, Section 3.3] ■

Lemma 1.17 *We have that*

$$\limsup_{s \uparrow T} \frac{H_s(v, v)}{\|v\|_{\mathcal{U}_s}} = 0 \tag{1.1.35}$$

Proof. The conclusion follows easily from the inequalities below, that are consequence of Gronwall's lemma and the Cauchy-Schwarz inequality:

$$\|z\|_\infty \leq C \int_s^T |v(t)| \mathrm{d}t \leq \sqrt{T - s}\, \|v\|_{\mathcal{U}_s} \tag{1.1.36}$$

■

For s close to T, the above lemma implies that the Hessian of f, i.e., $I_d + H_s$, is uniformly positive, and hence f is strongly convex, and has a unique critical point that is a minimum point. Therefore the first conjugate point τ_0 is the first for which H_s has an eigenvalue equal to -1.

1.2 Polyhedric constraints

1.2.1 Overview

Here we study problems of the form

$$\operatorname{Min} f(x); \quad x \in K, \tag{P}$$

with K closed convex subset of the Hilbert space X, and $f : X \to \mathbb{R}$ of class C^2. The essential hypothesis is that the set K is polyhedric (definition 1.22). It allows a rather complete theory of second-order optimality conditions and sensitivity.

Although the cost function is not necessarily quadratic, the application we have in view is linear quadratic optimal control problems with bound constraints on the control variable. Dealing with nonquadratic cost functions has its own interest since it suggests how to deal with nonquadratic optimal control problems (where as we will see two norms are to be used for the control space).

1.2.2 second-order necessary optimality conditions

In the statements below, X is a Hilbert space and f is of class C^2, $X \to \mathbb{R}$.

Define the (abstract) *critical cone* as

$$C(x) := \{h \in T_K(x);\ Df(x)h \leq 0\}.$$

A second-order necessary optimality condition is as follows.

Proposition 1.18 *Let $\bar{x}$, local solution of (P). Then $\bar{x}$ satisfies the first-order necessary optimality condition*

$$Df(\bar{x})h = 0, \quad \text{for all} \ \ h \in C(\bar{x}). \tag{1.2.37}$$

In addition,

$$D^2 f(\bar{x})(h,h) \geq 0, \quad \text{for all} \ \ h \in \overline{\mathcal{R}_K(\bar{x}) \cap Df(\bar{x})^{\perp}}. \tag{1.2.38}$$

Proof. Relation follows from the well-known first-order optimality condition

$$Df(\bar{x})(x - \bar{x}) \geq 0, \qquad \text{for all} \ \ x \in K \tag{1.2.39}$$

and the definition of the critical cone. If in addition $h \in \mathcal{R}_K(\bar{x}) \cap Df(\bar{x})^{\perp}$, then $\bar{x} + th \in K$ for $t > 0$ small enough, and hence

$$0 \leq \lim_{t \downarrow 0} \frac{f(\bar{x} + th) - f(\bar{x})}{\frac{1}{2}t^2} = D^2 f(\bar{x})(h,h).$$

Since $h \to D^2 f(\bar{x})(h,h)$ is continuous, this implies (1.2.38). ■

Remark 1.19 The conclusion holds even if K is nonconvex.

We now introduce a second-order sufficient optimality condition.

Proposition 1.20 *Let $\bar{x} \in K$, satisfying the second-order necessary optimality condition* (1.2.37)*. Assume that $D^2 f(\bar{x})$ is a Legendre form, and that*

$$D^2 f(\bar{x})(h,h) > 0, \quad \text{for all } \ h \in C(\bar{x}), \ h \neq 0. \tag{1.2.40}$$

Then $\bar{x}$ is a local solution of (P), that satisfies the quadratic growth condition.

Proof. If the conclusion is not satisfied, then there exists a sequence $x^k \to x$, $x^k \neq x$ for all k, such that

$$f(x^k) \leq f(\bar{x}) + o(\|x^k - \bar{x}\|^2). \tag{1.2.41}$$

Denote $t_k := \|x^k - \bar{x}\|$ and $h^k := t_k^{-1}(x^k - \bar{x})$. Then $x^k = \bar{x} + t_k h^k$, and hence,

$$f(x^k) = f(\bar{x}) + t_k Df(\bar{x})h^k + \tfrac{1}{2} t_k^2 D^2 f(\bar{x})(h^k, h^k) + o(t_k^2). \tag{1.2.42}$$

Combining with (1.2.41), get

$$Df(\bar{x})h^k + \tfrac{1}{2} t_k D^2 f(\bar{x})(h^k, h^k) \leq o(t_k). \tag{1.2.43}$$

Extracting if necessary a subsequence, we may assume that h^k weakly converges to some $\bar{h}$, and so $Df(\bar{x})h^k$ converges to $Df(\bar{x})\bar{h}$, so that with (1.2.43), $Df(\bar{x})\bar{h} \leq 0$. On the other hand, $\bar{h} \in T_K(\bar{x})$ (since a closed convex set is weakly closed), and hence, $\bar{h}$ is a critical direction.
By the first-order optimality condition $Df(\bar{x})h^k \geq 0$, so that with (1.2.43),

$$D^2 f(\bar{x})(h^k, h^k) \leq o(1),$$

and passing to the limit, $D^2 f(\bar{x})(\bar{h}, \bar{h}) \leq 0$. Condition (1.2.40) implies

$$D^2 f(\bar{x})(\bar{h}, \bar{h}) = 0, \tag{1.2.44}$$

and so $D^2 f(\bar{x})(\bar{h}, \bar{h}) = \lim_k D^2 f(\bar{x})(h^k, h^k)$. Since $D^2 f(\bar{x})$ is a Legendre form, this implies the strong convergence of h^k towards $\bar{h}$, and so $\|\bar{h}\| = 1$. Then (1.2.44) gives a contradiction with (1.2.40).

■

1.2.3 Polyhedric sets

It seems that there is an important gap between the previous necessary or sufficient second-order conditions, since they involve directions in the sets $\overline{\mathcal{R}_K(\bar{x}) \cap Df(\bar{x})^\perp}$ and $C(\bar{x})$, respectively. These two sets may be quite far one from each other, as shows the next example.

Example 1.21 Take $X = \mathbb{R}^2$, K the unit closed ball, and $f(x) = x_2$. At the minimum point $\bar{x} = (0, -1)^\top$, we have

$$C(\bar{x}) = \mathbb{R} \times \{0\}; \quad \overline{\mathcal{R}_K(\bar{x}) \cap Df(\bar{x})^\perp} = \{(0,0)\}. \tag{1.2.45}$$

That said, these two sets coincide in some important cases. Note that the first-order optimality condition may be written as

$$-Df(\bar{x}) \in N_K(\bar{x}).$$

Definition 1.22 *Let $x \in K$ and $q \in N_K(x)$. We say that K is* polyhedric *at x w.r.t. the normal direction q, if*

$$T_K(x) \cap q^\perp = \overline{\mathcal{R}_K(x) \cap q^\perp}. \tag{1.2.46}$$

If that property holds for all $x \in K$ and $q \in N_K(x)$, we say that K is polyhedric.

We will check that this applies to the case of bound constraints on the control. See section 1.2.6.

Proposition 1.23 *Assume that K is polyhedric, and that $\bar{x} \in K$ is such that $D^2 f(\bar{x})$ is a Legendre form, then $\bar{x}$ is a local minimum of (P) satisfying the quadratic growth condition iff it satisfies* (1.2.37) *and* (1.2.40).

Proof. By proposition 1.20, (1.2.37)-(1.2.40) implies local optimality with quadratic growth. Con,versely, assume that the quadratic growth condition holds. Then $\bar{x}$ satisfies the first-order condition (1.2.37), and is for $\alpha > 0$ small enough a local minimum of the problem

$$\operatorname{Min} f(x) - \tfrac{1}{2}\alpha\|x - \bar{x}\|^2; \quad x \in K.$$

Proposition 1.18 implies therefore the relation

$$D^2 f(\bar{x})(h,h) - \alpha\|h\|^2 \geq 0, \text{ for all } h \in \overline{\mathcal{R}_K(\bar{x}) \cap Df(\bar{x})^\perp},$$

implying itself (1.2.40). ■

1.2.4 Stability of solutions

Consider now a family of optimization problems of the form

$$\operatorname{Min} f(x,u); \quad x \in K, \tag{P_u}$$

with X a Hilbert space and U a Banach space, K a nonempty, closed and convex subset of X, and $f : X \times U \to \mathbb{R}$ of class C^2. We assume that $D^2_{xx} f(\bar{x}, \bar{u})$ is a Legendre form, and $\bar{x}$ local solution of $(P_{\bar{u}})$ satisfying the second-order sufficient condition

$$D_x f(\bar{x},\bar{u})h = 0 \text{ and } D^2_{xx} f(\bar{x},\bar{u})(h,h) > 0, \text{ for all } h \in C(\bar{x},\bar{u}), \ h \neq 0, \tag{1.2.47}$$

where $C(\bar{x},\bar{u})$ denotes the critical cone

$$C(\bar{x},\bar{u}) := \{h \in T_K(\bar{x}); \ D_x f(\bar{x},\bar{u})h \leq 0\}. \tag{1.2.48}$$

By proposition 1.20 the quadratic growth condition is satisfied. More precisely, define the *local problem* (around $\bar{x}$)

$$\operatorname{Min} f(x,u); \quad x \in K, \quad \|x - \bar{x}\| \leq \theta \tag{$P_{u,\theta}$}$$

with $\theta > 0$. Then for $\theta > 0$ small enough (we assume that this holds in the sequel), $\bar{x}$ is unique solution of $(P_{\bar{u},\theta})$, and there exists $\alpha > 0$ such that

$$f(x, \bar{u}) \geq f(\bar{x}, \bar{u}) + \alpha \|x - \bar{x}\|^2, \quad \text{for all} \quad x \in K, \ \|x - \bar{x}\| \leq \theta. \tag{1.2.49}$$

Let us show the stability of the local solution of (P_u) w.r.t. a perturbation.

Proposition 1.24 *Assume f w.l.s.c., $D^2_{xx} f(\bar{x}, \bar{u})$ a Legendre form, the second-order condition* (1.2.47) *satisfied, and let $\theta > 0$ be such that* (1.2.49) *holds. Then , for all $u \in U$, the local problem $(P_{u,\theta})$ has at least one solution and , if $x_u \in S(P_{u,\theta})$, we have*

$$\|x - \bar{x}\| = O(\|u - \bar{u}\|). \tag{1.2.50}$$

Proof. A minimizing sequence of problem $(P_{u,\theta})$ is bounded. Since X is a Hilbert space, there exists a limit-point (for the weak topology) x_u. The set K is weakly closed, and f is w.l.s.c.; therefore $x_u \in S(P_{u,\theta})$. Combining relations

$$\begin{array}{rcl} f(x_u, \bar{u}) & = & f(x_u, u) + \int_0^1 Df(x_u, u + \sigma(\bar{u} - u))(\bar{u} - u)\mathrm{d}\sigma \\ f(\bar{x}, \bar{u}) & = & f(\bar{x}, u) + \int_0^1 Df(\bar{x}, u + \sigma(\bar{u} - u))(\bar{u} - u)\mathrm{d}\sigma \end{array}$$

with the quadratic growth condition (1.2.49), we get

$$\begin{array}{rl} \alpha \|x_u - \bar{x}\|^2 & \leq f(x_u, \bar{u}) - f(\bar{x}, \bar{u}) \\ & \leq f(x_u, \bar{u}) - f(x_u, u) + f(\bar{x}, u) - f(\bar{x}, \bar{u}) \\ & = \int_0^1 \left[Df(x_u, u + \sigma(\bar{u} - u)) - Df(\bar{x}, u + \sigma(\bar{u} - u)) \right] (\bar{u} - u)\mathrm{d}\sigma \\ & = O\left(\|x_u - \bar{x}\| \, \|u - \bar{u}\| \right), \end{array}$$

implying (1.2.50). ■

1.2.5 Sensitivity analysis

We have a mapping $\mathbb{R}_+ \to U$, $t \to u(t)$ with $d \in U$, be such that

$$u(t) = \bar{u} + td + r(t); \quad \|r(t)\| = o(t). \tag{1.2.51}$$

Set $v(t) := \text{val}(P_{u(t),\theta})$, where $\theta > 0$ is such that (1.2.49) is satisfied. Define the subproblem

$$\underset{h \in C(\bar{x})}{\text{Min}} \ D^2 f(\bar{x}, \bar{u})((h, d), (h, d)). \tag{SP}$$

Theorem 1.25 *Assume that K is polyhedric, that f is weakly l.s.c., that $D^2 f(\bar{x})$ is a Legendre form, and that the second-order condition* (1.2.47) *is satisfied. Then the value function may be expanded as follows:*

$$v(t) = v(0) + D_u f(\bar{x}, \bar{u})(u(t) - \bar{u}) + \tfrac{1}{2} t^2 \, \text{val}(SP) + o(t^2). \tag{1.2.52}$$

In addition, any weak limit-point $\bar{h}$ of $(x_t - \bar{x})/t$ is a strong limit-point, and satisfies $\bar{h} \in S(SP)$. If (SP) has the unique solution $\bar{h}$, then the following expansion of solutions holds

$$x_t = \bar{x} + t\bar{h} + o(t). \tag{1.2.53}$$

Proof. a) *Upper estimate.* Let $\varepsilon > 0$. Since K is polyhedric, there exists $h \in \mathcal{R}_K(\bar{x}) \cap Df(\bar{x})^\perp$ such that

$$D^2 f(\bar{x}, \bar{u})((h, d), (h, d)) \leq \text{val}(SP) + \varepsilon.$$

The following holds:

$$\begin{aligned} f(\bar{x} + th, u(t)) = \;& f(\bar{x}, \bar{u}) + D_u f(\bar{x}, \bar{u})(u(t) - \bar{u}) \\ & + \tfrac{1}{2} t^2 D^2 f(\bar{x}, \bar{u})((h, d), (h, d)) + o(t^2). \end{aligned} \tag{1.2.54}$$

Since $\bar{x} + th \in K$ pour $t > 0$ small enough, we have

$$v(t) \leq f(\bar{x} + th, u(t)) \leq f(\bar{x}, \bar{u}) + D_u f(\bar{x}, \bar{u})(u(t) - \bar{u}) + \tfrac{1}{2} t^2 \left(\text{val}(SP) + \varepsilon\right) + o(t^2). \tag{1.2.55}$$

This being true for any $\varepsilon > 0$, we obtain

$$v(t) \leq f(\bar{x}, \bar{u}) + D_u f(\bar{x}, \bar{u})(u(t) - \bar{u}) + \tfrac{1}{2} t^2 \,\text{val}(SP) + o(t^2). \tag{1.2.56}$$

b) *Lower estimate.* Let $x_t \in S(P_{u(t),\theta})$. By proposition 1.24, we know that

$$\|x_t - \bar{x}\| = O(\|u(t) - \bar{u}\|) = O(t),$$

and $h_t := (x_t - \bar{x})/t$ is therefore bounded. Let $\bar{h}$ be a weak limit-point. We have

$$\begin{aligned} f(x_t, u(t)) &= f(\bar{x} + th_t, u(t)) \\ &= f(\bar{x}, \bar{u}) + Df(\bar{x}, \bar{u})(x_t - x, u(t) - \bar{u}) \\ &\quad + \tfrac{1}{2} t^2 D^2 f(\bar{x}, \bar{u})((h_t, d), (h_t, d)) + o(t^2). \end{aligned}$$

Comparing to (1.2.56), obtain after division by $\frac{1}{2} t^2$

$$2t^{-1} D_x f(\bar{x}, \bar{u}) h_t + D^2 f(\bar{x}, \bar{u})((h_t, d), (h_t, d)) \leq \text{val}(SP) + o(1). \tag{1.2.57}$$

This implies $D_x f(\bar{x}, \bar{u}) h_t \leq o(t)$, and hence, $D_x f(\bar{x}, \bar{u}) \bar{h} \leq 0$. Since $h_t \in \mathcal{R}_K(\bar{x})$, we have $\bar{h} \in T_K(\bar{x})$, therefore $\bar{h}$ is a critical direction. On the other hand, $h_t \in \mathcal{R}_K(\bar{x})$ combined with the first-order necessary condition implies $D_x f(\bar{x}, \bar{u}) h_t \geq 0$. Using the weak l.s.c. of $D^2 f(\bar{x}, \bar{u})$, get with (1.2.57)

$$D^2 f(\bar{x}, \bar{u})((\bar{h}, d), (\bar{h}, d)) \leq \liminf_{t \downarrow 0} D^2 f(\bar{x}, \bar{u})((h_t, d), (h_t, d)) \leq \text{val}(SP).$$

As $\bar{h} \in C(\bar{x})$, this im plies $\bar{h} \in S(SP)$ and hence,

$$D^2 f(\bar{x}, \bar{u})((h_t, d), (h_t, d)) \to D^2 f(\bar{x}, \bar{u})((\bar{h}, d), (\bar{h}, d)).$$

Since $\bar{h}$ is a weak limit-point of h_t, this implies $D^2_{xx} f(\bar{x}, \bar{u})(h_t, h_t) \to D^2_{xx} f(\bar{x}, \bar{u})(\bar{h}, \bar{h})$. Since $D^2_{xx} f(\bar{x}, \bar{u})$ is a Legendre form, we deduce that $\bar{h}$ is a limit-point of h_t for the strong convergence. In particular, if (SP) has a unique solution, then $h_t \to \bar{h}$, implying (1.2.53). ■

1.2.6 Bound constraints in spaces of summable square

In this section we apply the above results to the case when Ω is an open subset of $\mathbb{R}^n$, $X := L^2(\Omega)$ is the Hilbert space of summable square over Ω, and $K := L^2(\Omega)_+$ is the set of nonnegative a.e. functions of X. We recall the following result, due to Lebesgue.

Theorem 1.26 (Dominated convergence) *Let x_n a sequence of elements of $L^2(\Omega)$. Suppose that there exists $g \in L^2(\Omega)$ such that $|x_n(\omega)| \leq g(x)$, a.e. and that, for almost all ω, $x_n(\omega)$ converges. Set $x(\omega) = \lim_n x_n(\omega)$. Then $x \in L^2(\Omega)$, and $x_n \to x$ in $L^2(\Omega)$.*

Given $x \in L^2(\Omega)$, denote

$$I(x) := \{\omega \in \Omega;\ x(\omega) = 0\}; \quad J(x) := \{\omega \in \Omega;\ x(\omega) > 0\},$$

the contact set and its complement, defined up to a null measure set. The lemma below states the essential properties for the sequel.

Lemma 1.27 (i) *The cone K is a closed subset of $L^2(\Omega)$.*
(ii) *Its dual cone is $K^- = L^2(\Omega)_-$, the set of functions of X that are nonpositive a.e.*
(iii) *Let $x \in K$. Then*

$$\begin{aligned} T_K(x) &:= \{h \in X;\ h \geq 0, \quad a.e. \quad sur\ \ I(x)\}, && (1.2.58)\\ N_K(x) &:= \{h \in X_-;\ h = 0, \quad a.e. \quad sur\ \ J(x)\}. && (1.2.59)\end{aligned}$$

In addition, let $q \in N_K(x)$. Then

$$T_K(x) \cap q^\perp = \{h \geq 0, \quad a.e. \quad sur\ \ I(x);\ h(\omega)q(\omega) = 0 \quad a.e.\ \}. \qquad (1.2.60)$$

(iv) *The positive cone of $L^2(\Omega)$ is polyhedric.*

Proof. (i) Let $x_n \to \bar{x}$ in $L^2(\Omega)$, x_n nonnegative a.e. The function

$$y_n(\omega) := \min(0, x_n(\omega))$$

has value zero, and converges in $L^2(\Omega)$ towards $\min(0, \bar{x})$ in view of the dominated convergence theorem. Therefore $\min(0, \bar{x}) = 0$ in $L^2(\Omega)$, so that $\bar{x} \geq 0$ a.e., as was to be shown.
(ii) If $y \in L^2(\Omega)_-$, then clearly $\int_\Omega y(\omega)x(\omega)d\omega \leq 0$ for all $x \in K$, and hence, $L^2(\Omega)_- \subset K^-$. Conversely, if $y \in K^-$, let $x \in L^2(\Omega)$ defined by $x(\omega) := \max(0, y(\omega))$ a.e.; then $x \in K$ and hence, $0 \geq \int_\Omega y(\omega)x(\omega)d\omega = \int_\Omega (y(\omega))_+^2 d\omega$. Therefore $y(\omega) \leq 0$ a.e., implying (ii).
(iii) The expression of normal directions is a direct consequence of the formula of normal cones when the set K is a cone, see e.g. [10, Example 2.62]:

$$N_K(x) = K^- \cap x^\perp \qquad (1.2.61)$$

The one of the tangent cone follows, using the relation $T_K(x) = N_K(x)^-$, and the latter implying (1.2.60).
(iv) Let $h \in T_K(x) \cap q^\perp$. Set, for $\varepsilon > 0$, $h_\varepsilon := ((x + \varepsilon h)_+ - x)/\varepsilon$. Then $x + \varepsilon h_\varepsilon = (x + \varepsilon h)_+ \in K$, and hence, $h \in \mathcal{R}_K(x)$. By the dominated convergence theorem, $h_\varepsilon \to h$

in $L^2(\Omega)$. Point (ii) implies that $h_\varepsilon(\omega)q(\omega)$ is zero for almost all ω, and hence, $h \in q^\perp$. We have shown that K is polyhedric. ■

For problem

$$\operatorname*{Min}_{x \in L^2(\Omega)_+} f(x),$$

with f of class $C^2 : L^2(\Omega) \to \mathbb{R}$, the second-order sufficient optimality condition (1.2.40) writes, taking into account the previous lemma, when $D^2 f(\bar{x})$ is a Legendre form:

$$\begin{cases} Df(\bar{x})(\omega) \geq 0, \quad Df(\bar{x})(\omega)x(\omega) = 0, \ \text{ a.e.} \\ D^2 f(\bar{x})(h,h) > 0, \ \text{ for all } \ h \geq 0 \ \text{ over } \ I(\bar{x}), \\ Df(\bar{x})(\omega)h(\omega) = 0 \ \text{ a.e.} \end{cases} \tag{1.2.62}$$

1.3 Convex constraints on control variables

1.3.1 Framework

In this section we assume that the state equation is linear, and that the cost function is quadratic, given by (1.1.2) and (1.1.3) respectively. The problem is

$$\operatorname*{Min}_{u} f(u); \quad u \in K. \tag{P}$$

The novelty is that we have now control constraints of the form

$$u \in K,$$

where

$$K := \{u \in \mathcal{U};\ g(u(t)) \leq 0,\ a.e.\ t \in (0,T)\}. \tag{1.3.63}$$

The *convex* function $g : \mathbb{R}^m \to: \mathbb{R}^{n_g}$ is assumed to be $C^2 : \mathbb{R} \to \mathbb{R}$. For simplicity we assume that

$$g(0) = 0. \tag{1.3.64}$$

1.3.2 First-order necessary optimality conditions

Let $\bar{u}$ be a local solution of the problem

$$\operatorname*{Min}_{u} f(u); \quad u \in K.$$

Since K is convex, a first-order necessary optimality condition is

$$Df(\bar{u})(u - \bar{u}) \geq 0, \qquad \text{for all } u \in K, \tag{1.3.65}$$

or equivalently

$$Df(\bar{u}) + N_K(\bar{u}) \ni 0. \tag{1.3.66}$$

We can prove the following result of smoothness of optimal control (for which no qualification condition is needed). We denote by $\bar{y}$, $\bar{p}$ the state and costate associated with a solution or critical point $\bar{u}$.

Lemma 1.28 *Assume that R_t is uniformly positive:*

$$\exists\ \alpha > 0;\ u \cdot R_t u \geq \alpha |u|^2, \quad \textit{for almost all}\, t \in (0,T). \tag{1.3.67}$$

Then any solution of the first-order necessary optimality conditions is essentially bounded.

Proof. Let $\bar{u}$ be such a solution. Combining proposition 1.2 and (1.3.65), we obtain that the following holds:

$$(B_t^\top p_t + R_t \bar{u}_t + D_t y_t) \cdot (v - \bar{u}_t) \geq 0, \qquad \text{for all } v \in g^{-1}(]-\infty, 0]), \quad t \in [0,T]. \tag{1.3.68}$$

In view of (1.3.64), we may take $v = 0$, obtaining (using (1.3.67) and the fact that B_t, D_t, p, y are essentially bounded)

$$\alpha |\bar{u}_t|^2 \leq \bar{u}_t \cdot R_t \bar{u}_t \leq (B_t^\top p_t + D_t y_t) \cdot \bar{u}_t \leq c|\bar{u}_t| \quad t \in [0,T], \tag{1.3.69}$$

for some constant c. Then by the Cauchy Schwarz inequality, $|\bar{u}_t| \leq c/\alpha$ for a.a. t. ■

Again without any qualification condition, we can show the local nature of the tangent and normal cones to K. Denote

$$K_g := g^{-1}(\mathbb{R}_-^{n_g}).$$

Lemma 1.29 *Let $u \in K$. Then*

$$T_K(u) = \{v \in \mathcal{U};\ v_t \in T_{K_g}(u_t) \ \textit{for almost all}\, t \in (0,T)\}. \tag{1.3.70}$$

$$N_K(u) := \{\mu \in \mathcal{U};\ \mu_t \in N_{K_g}(u_t) \ \textit{for almost all}\, t \in (0,T)\}. \tag{1.3.71}$$

Proof. Denote by P_K the orthogonal projection onto K (well-defined since K is a closed convex set of the Hilbert space $\mathcal{U}$). We have that $v \in T_K(u)$ iff, given $\varepsilon > 0$,

$$v^\varepsilon := \varepsilon^{-1}(P_K(u + \varepsilon v) - u)$$

is such that $v^\varepsilon \to v$ in $\mathcal{U}$ when $\varepsilon \downarrow 0$. Obviously

$$v_t^\varepsilon = \varepsilon^{-1}(P_{K_g}(u_t + \varepsilon v_t) - u_t), \qquad \text{a.e. } t \in (0,T). \tag{1.3.72}$$

Since P_{K_g} is non expansive, $|v_t^\varepsilon| \leq |v_t|$ a.e., therefore the dominated convergence theorem implies that $v^\varepsilon \to v$ in $\mathcal{U}$ when $\varepsilon \downarrow 0$ iff $v_t^\varepsilon \to v_t$ a.e. The latter holds iff $v_t \in T_{K_g}(u_t)$ a.e.; relation (1.3.70) follows, and (1.3.71) is an easy consequence of (1.3.70). ■

We need however a qualification condition in order to relate the expression of the Lagrange multipliers to $g(u)$ and $Dg(u)$. So let us assume that

$$\exists\ \beta > 0 \ \text{ and } \ u^0 \in \mathbb{R}^m; \quad g(u^0) < -\beta. \tag{1.3.73}$$

In that case it is well-known that for all $u \in \mathbb{R}^m$:

$$T_{K_g}(u) = \{v \in \mathbb{R}^m;\ Dg_i(u)v \leq 0, \ \text{ for all } \ i;\ g_i(u) = 0\} \tag{1.3.74}$$

$$N_{K_g}(u) = \left\{ \sum_{i=1}^{n_g} \lambda_i Dg_i(u);\ \lambda \in \mathbb{R}_+^m;\ \lambda_i = 0, \ \text{ for all } \ i;\ g_i(u) < 0 \right\}. \tag{1.3.75}$$

Denote the set of active constraints at a point $u \in \mathcal{U}$ (defined up to a null measure set) by

$$I_t(u) := \{1 \leq i \leq n_g;\ g_i(u_t) = 0\}. \tag{1.3.76}$$

Lemma 1.30 *Let $u \in \mathcal{U}$, and assume that the qualification condition* (1.3.73) *holds. Then*

$$T_K(u) = \{v \in \mathcal{U};\ Dg_i(u_t)v_t \leq 0, \quad \text{for a.a. } t \in (0,T),\ \ i \in I_t(u_t)\}, \tag{1.3.77}$$

$$\begin{array}{rl} N_K(u) = & \{ u \in \mathcal{U};\ \ \mu_t = \sum_{i=1}^{n_g} \lambda_{i,t} Dg_i(u_t);\ \lambda_{i,t} \in \mathbb{R}^m_+; \\ & \lambda_{i,t} = 0, \ \ \text{for all } \ i;\ g_i(u_t) < 0 \ \ \text{a.e.} \ \ t \in (0,T)\}. \end{array} \tag{1.3.78}$$

In addition we have that if λ satisfies (1.3.78)*, then*

$$\sum_i |\lambda_{i,t}| \leq \beta^{-1} |\mu_t|\, |u^0 - u_t|. \tag{1.3.79}$$

Proof. Relations (1.3.77) and (1.3.77) are immediate consequences of the above relations. If λ satisfies (1.3.78), then since g is convex, then a.e., for all $i \in I_t(u)$:

$$-\beta \geq g(u^0) \geq Dg_i(u_t)(u^0 - u_t). \tag{1.3.80}$$

Multiplying by $\lambda_{i,t}$ and summing over i (the contribution of non active constraints is zero) we get

$$-\beta \sum_i |\lambda_{i,t}| \geq \mu_t \cdot (u^0 - u_t) \geq -|\mu_t|\, |u^0 - u_t|, \tag{1.3.81}$$

from which (1.3.79) follows. ■

Remark 1.31 Note that the above λ is not necessarily measurable. A measurable λ can be constructed as follows. Given $J \subset \{1, \dots, n_g\}$, denote the (measurable) set of times for which the set of active constraints is J (defined up to a null measure set) by

$$\mathcal{T}_J := \{t \in (0,T);\ I_t(u_t) = J\}. \tag{1.3.82}$$

Next, denote by $\varphi_J(\eta_t, \gamma_t)$ the solution of the following problem

$$\operatorname*{Min}_{\lambda \in \mathbb{R}^{n_g}_+} |\lambda|; \quad \eta_t := \sum_{i \in J} \lambda_i \gamma_i; \quad \lambda_i = 0, \ \ i \notin J. \tag{1.3.83}$$

When $t \in \mathcal{T}_J$, $\eta_t = \mu_t$ and $\gamma_t = Dg(u_t)$, the problem has a unique solution that (in view of the qualification condition) depends continuously on (η_t, γ_t); otherwise observe that if $\eta = 0$, the solution is $\lambda = 0$. Now the minimum-norm λ can be expressed as

$$\lambda_t := \sum_{J \subset \{1,\dots,n_g\}} \varphi_J(\mathbf{1}_{t \in \mathcal{T}_J} \mu_t, Dg(u_t)). \tag{1.3.84}$$

Being a sum of continuous functions of measurable mappings, this is a measurable function.

Denote the set of *Lagrange multipliers* by

$$\Lambda(u) := \left\{ \lambda \in L^2(0,T;\mathbb{R}^{n_g});\ \lambda_t \in N_{K_g}(u_t) \text{ a.e.};\ Df(u)_t + \sum_{i=1}^{n_g} \lambda_{i,t} Dg_i(u_t) = 0 \right\}. \tag{1.3.85}$$

Lemma 1.32 *The point $\bar{u}$ satisfies the first-order necessary optimality conditions iff $\Lambda(u)$ is not empty. If in addition R_t is uniformly positive, then $\Lambda(u)$ is a bounded and weakly$*$ closed subset of $L^\infty(0,T,\mathbb{R}^m)$.*

Proof. The expression of the set of Lagrange multipliers is a consequence of the expressions of the normal cone to K given before.

The L^∞ boundedness follows from lemma 1.28 and (1.3.79). It remains to show that $\Lambda(u)$ is weakly$*$ closed. Since a half-space of the form

$$H_{\psi,\beta} := \{\gamma \in L^\infty(0,T,\mathbb{R}^{n_g});\ \int_0^T \gamma_t \cdot \psi_t \mathrm{d}t \le \beta\} \tag{1.3.86}$$

is weakly$*$ closed whenever $\psi \in L^1(0,T,\mathbb{R}^{n_g})$, it suffices to show that $\Lambda(u)$ is an intersection of such spaces. Obviously $Df(u) + \sum_{i=1}^{n_g} \lambda_{i,t} Dg_i(u_t) = 0$ iff

$$\int_0^T [Df(u) + \sum_{i=1}^{n_g} \lambda_{i,t} Dg_i(u_t)\psi_{i,t}]\mathrm{d}t = 0, \qquad \text{for all } \ \psi \in L^1(0,T,\mathbb{R}^n). \tag{1.3.87}$$

That $\lambda \ge 0$ holds iff $\int_0^T \lambda_t \psi_t \mathrm{d}t \ge 0$, for all $\psi \in L^1(0,T,\mathbb{R}^{n_g})_+$. Finally the complementarity condition can be written as $\int_0^T \lambda_t g_i(u_t)\mathrm{d}t = 0$. ■

1.3.3 Second-order necessary optimality conditions

The essential ingredient here is to build paths that are "second order feasible". The set of "strongly active constraints" is defined as

$$I_t^+(u) := \{1 \le i \le n_g;\ \lambda_{i,t} > 0, \ \text{ for all } \lambda \in \Lambda(u)\}. \tag{1.3.88}$$

The critical cone is a as follows

$$C(u) := \{v \in T_K(u);\ Dg_i(u_t)v_t = 0,\ i \in I_t^+(u), \ \text{ a.a. } t\} \tag{1.3.89}$$

Let, for $\varepsilon > 0$, the ε-"almost active" constraints be defined by

$$I_t^\varepsilon(u) := \{1 \le i \le n_g;\ -\varepsilon \le g_i(u_t) < 0\}. \tag{1.3.90}$$

Denote by $C_\varepsilon(\bar{u})$ the cone of pseudo-feasible and essentially bounded critical directions, in the following sense:

$$C_\varepsilon(\bar{u}) := \{v \in C(\bar{u});\ \|v\|_\infty \le 1/\varepsilon; v_t = 0 \ \text{ if } I_t^\varepsilon(u) \ne \emptyset, \ \text{for a.a. } t\}. \tag{1.3.91}$$

Lemma 1.33 *The set $\cup_{\varepsilon>0} C_\varepsilon(\bar{u})$ is a dense subset of $C(\bar{u})$.*

Proof. Let v be a critical direction. Let $v^{1,\varepsilon}$ be the truncation

$$v_t^{1,\varepsilon} := \max(-1/\varepsilon, \min(1/\varepsilon, v_t)), \qquad \text{for all } \ \ t \in (0,T), \tag{1.3.92}$$

and v^ε be defined by

$$v_t^\varepsilon = \begin{cases} 0 & \text{if } \ I_t^\varepsilon(\bar{u}) \ne \emptyset \\ v_t^{1,\varepsilon} & \text{if not} \end{cases} \tag{1.3.93}$$

Obviously $v^\varepsilon \in C_\varepsilon(\bar{u})$. Since $\text{meas}(\cap_{\varepsilon>0} I_t^\varepsilon) = 0$, we have that $v^\varepsilon \to v$ a.e. when $\varepsilon \downarrow 0$. Since $|v_t^\varepsilon| \leq |v_t|$ a.e., the dominated convergence theorem implies that $v^\varepsilon \to v$ in $\mathcal{U}$. The result follows. ■

Let us see now, for given $v \in C_\varepsilon(\bar{u})$, build a "second-order feasible" path (this corresponds to the "primal form" of the second-order necessary conditions)

Lemma 1.34 *Given $\varepsilon > 0$ and $v \in C_\varepsilon(\bar{u})$, let $w \in L^\infty(0, T; \mathbb{R}^m)$ be such that*

$$Dg_i(\bar{u}_t)w + D^2 g_i(\bar{u}_t)(v_t, v_t) \leq -\varepsilon,\ i \in I_t(\bar{u}) \cup I_t^\varepsilon(\bar{u}). \tag{1.3.94}$$

Then for $\theta > 0$ small enough, the path u^θ defined below is contained in K:

$$u^\theta := \bar{u} + \theta v + \tfrac{1}{2}\theta^2 w. \tag{1.3.95}$$

Proof. This is an immediate consequence of a second-order expansion of $g(u^\theta)$, combined with the definitions of $I_t(\bar{u})$ and $I_t^\varepsilon(\bar{u})$. ■

Define the set of "ε-augmented Lagrange multipliers" as

$$\Lambda_\varepsilon(u) := \left\{ \lambda \in N_g(u);\ Df(u)_t + \sum_{i \in I_t(\bar{u}) \cup I_t^\varepsilon(\bar{u})} \lambda_{i,t} Dg_i(u_t) = 0 \right\}. \tag{1.3.96}$$

The qualification condition (1.3.73) implies that these sets are uniformly bounded when $\varepsilon < \beta$, and we have that $\Lambda(u) = \cap_{\varepsilon>0} \Lambda_\varepsilon(u)$.

Define the *Lagrangian* of problem (P) as $L : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$,

$$L(u, \lambda) = f(u) + \int_0^T \sum_{i=1}^{n_g} \lambda_t g(u_t)\mathrm{d}t. \tag{1.3.97}$$

Theorem 1.35 *Let $\bar{u}$ be a local solution of (P). Then for any critical direction v, there exists a Lagrange multiplier λ such that*

$$D^2_{uu} L(u, \lambda)(v, v) \geq 0. \tag{1.3.98}$$

Proof. a) Given $\varepsilon > 0$, let $v \in C_\varepsilon(\bar{u})$. Consider the subproblem

$$\begin{array}{l} \displaystyle\operatorname*{Min}_{w \in \mathcal{U}} Df(\bar{u})w + D^2 f(\bar{u})(v, v); \\ Dg_i(\bar{u}_t)w + D^2 g_i(\bar{u}_t)(v_t, v_t) \leq -\varepsilon, \quad i \in I_t(\bar{u}) \cup I_t^\varepsilon(\bar{u}). \end{array} \tag{SP_ε}$$

We choose $L^2(0, T, \mathbb{R}^{n_g})$ as constraint space. By lemma 1.34, for any feasible w in $F(SP_\varepsilon) \cap L^\infty(0, T; \mathbb{R}^m)$, the path u^θ defined in (1.3.95) is feasible. Since v is a critical direction, $Df(\bar{u})v = 0$. Using the fact that $\bar{u}$ is a local minimum of (P), we get

$$0 \leq \lim_{\theta \downarrow 0} \frac{f(u^\theta) - f(\bar{u})}{\frac{1}{2}\theta^2} = Df(\bar{u})w + D^2 f(\bar{u})(v, v) \tag{1.3.99}$$

Now let $w \in F(SP_\varepsilon)$. For $\gamma > 0$, let $w^\gamma \in F(SP_\varepsilon) \cap L^\infty(0, T; \mathbb{R}^m)$ be the unique solution of

$$\begin{array}{l} \displaystyle\operatorname*{Min}_{w \in \mathcal{U}} \int_0^T |w_t - w_t^\gamma|^2 \mathrm{d}t; \quad \|w^\gamma\|_\infty \leq 1/\gamma; \\ Dg_i(\bar{u}_t)w + D^2 g_i(\bar{u}_t)(v_t, v_t) \leq -\varepsilon,\ i \in I_t(\bar{u}) \cup I_t^\varepsilon(\bar{u}). \end{array} \tag{1.3.100}$$

Let us show that for $\varepsilon < \frac{1}{2}\beta$ and small enough γ, this problem is feasible. Indeed if $i \in I_t(\bar{u}) \cup I_t^\varepsilon(\bar{u})$, then $g_i(\bar{u}_t) \geq -\varepsilon$, so that

$$-\beta > g_i(u^0) \geq g_i(\bar{u}_t) + Dg_i(\bar{u}_t)(u^0 - \bar{u}_t) \geq -\varepsilon + Dg_i(\bar{u}_t)(u^0 - \bar{u}_t), \tag{1.3.101}$$

That is, set $\hat{w} := (u^0 - \bar{u}_t)$. Then

$$\hat{w} \in L^\infty(0, T, \mathbb{R}^n) \quad \text{and} \quad Dg_i(\bar{u}_t)\hat{w}_t \leq -\tfrac{1}{2}\beta, \quad \text{for a.a.} \quad t \in (0, T). \tag{1.3.102}$$

Since $\bar{u}$ is essentially bounded, this proves that the linear constraints may be satisfied by some essentially bounded w such that $\|w\|_\infty \leq c(\|v\|_\infty^2 + \varepsilon)$, for some $c > 0$ not depending on v or ε. Finally if $1/\gamma \geq c(\|v\|_\infty + \varepsilon)$, feasibility of (1.3.100) holds.

Now $w_t^\gamma = w_t$ if $|w_t| \leq 1/\gamma$, and $|w_t^\gamma| \leq |w_t|$ a.e.; it follows that when $\gamma \downarrow 0$, $w^\gamma \to w$ in $\mathcal{U}$. Passing to the limit in (1.3.99) (in which w is w^γ) we obtain that

$$Df(\bar{u})w + D^2 f(\bar{u})(v, v) \geq 0, \quad \text{for all} \quad w \in F(SP_\varepsilon). \tag{1.3.103}$$

In other words, $F(SP_\varepsilon)$ has a nonnegative value.
b) The dual (in the sense of convex analysis) of (SP_ε) is the problem

$$\operatorname*{Max}_{\lambda \in \Lambda_\varepsilon(\bar{u})} D_{uu}^2 L(u, \lambda)(v, v) + \varepsilon \|\lambda\|_{L^1}. \tag{SD_ε}$$

The problem obtained by an additive perturbation of the constraints, i.e.,

$$\begin{array}{l} \operatorname{Min}_{w \in \mathcal{U}} Df(\bar{u})w + D^2 f(\bar{u})(v, v); \\ Dg_i(\bar{u}_t)w + D^2 g_i(\bar{u}_t)(v_t, v_t) \leq -\varepsilon + \eta, \; i \in I_t(\bar{u}) \cup I_t^\varepsilon(\bar{u}), \end{array} \tag{1.3.104}$$

where $\eta \in \mathcal{U}$, is feasible; indeed, using $\hat{w}$ satisfying (1.3.102), it suffices to take w of the form

$$w_t = c(1 + |\eta_t|)\hat{w}_t, \quad \text{for large enough} \quad c > 0. \tag{1.3.105}$$

It follows that the primal and dual values are equal. In addition, we know that the set of dual solutions is bounded and weakly$*$ compact. In view of step a), we obtain that $\operatorname{val}(SD_\varepsilon) \geq 0$.
c) It is easily checked that $\Lambda_\varepsilon(\bar{u})$ is bounded in $L^\infty(0, T, \mathbb{R}^{n_g})$. We may check that it is a weakly$*$ compact subset of $L^\infty(0, T, \mathbb{R}^{n_g})$, using arguments similar to those of the proof of lemma 1.32.
d) Let $v \in C(\bar{u})$, and for $\varepsilon > 0$, $v^\varepsilon \in C_\varepsilon(\bar{u})$ be such that $v^\varepsilon \to v$ in $\mathcal{U}$. It follows that $D^2 f(\bar{u})(v^\varepsilon, v^\varepsilon) \to D^2 f(\bar{u})(v, v)$ and $D^2 g(\bar{u})(v^\varepsilon, v^\varepsilon) \to D^2 g(\bar{u})(v, v)$ in $L^1(0, T, \mathbb{R}^{n_g})$. For each $\varepsilon > 0$ there exists $\lambda^\varepsilon \in \Lambda_\varepsilon(\bar{u})$ such that $D_{uu}^2 L(\bar{u}, \lambda^\varepsilon)(v^\varepsilon, v^\varepsilon) + \varepsilon \|\lambda^\varepsilon\|_{L^1} \geq 0$. Given $\varepsilon_0 > 0$, λ^ε belongs to $\Lambda_{\varepsilon_0}(\bar{u})$ when $\varepsilon < \varepsilon_0$. Since $\Lambda_{\varepsilon_0}(\bar{u})$ is weakly$*$ compact, there exists a sequence $\varepsilon_k \downarrow 0$, such that there exists $\lambda^k \in \Lambda_{\varepsilon_k}(\bar{u})$ that weakly$*$ converges to some $\bar{\lambda}$, and denoting by v^k the corresponding sequence extracted from v^ε,

$$D_{uu}^2 L(\bar{u}, \lambda^{\varepsilon_k})(v^k, v^k) + \varepsilon_k \|\lambda^k\|_{L^1} \geq 0.$$

We obtain that $\bar{\lambda} \in \Lambda_\varepsilon(\bar{u})$ for all $\varepsilon > 0$, and hence $\bar{\lambda} \in \Lambda(\bar{u})$, and

$$D_{uu}^2 L(\bar{u}, \bar{\lambda})(v, v) = \lim D_{uu}^2 L(\bar{u}, \lambda^\varepsilon)(v^\varepsilon, v^\varepsilon) \geq 0 \tag{1.3.106}$$

as was to be proved. ■

1.4 Notes

The theory of unconstrained linear quadratic problems is classical and can be found in many textbooks. We have taken the point of view of studying the critical points. Also we emphasize the role of Legendre form in the case of minimization problems. The concept of polyhedricity is due to Haraux [16] and Mignot [20]. Our presentation in section 1.2 follows [5]. Various extensions are presened in [10]. Section 1.3 is an adaptation to the case of the control of ODEs of results obtained when dealing with the optimal control of a semilinear elliptic system [4].

Chapter 2

Nonlinear optimal control,

2.1 Unconstrained nonlinear optimal control

2.1.1 Setting

We consider in this section unconstrained optimal controls problems, with nonlinear dynamics and cost functions. Due to this we restrict the analysis to the case of essentially bounded control variables. So the function spaces for the control and state variables will be

$$\mathcal{U} := L^\infty(0,T;\mathbb{R}^m); \quad \mathcal{Y} := W^{1,\infty}(0,T;\mathbb{R}^n). \tag{2.1.1}$$

The optimal control problem is as follows

$$(\mathcal{P}) \qquad \min_{(u,y)\in\mathcal{U}\times\mathcal{Y}} F(u,y) := \int_0^T \ell(u(t),y(t))\mathrm{d}t + \phi(y(T)) \tag{2.1.2}$$

$$\text{subject to} \qquad \dot{y}(t) = f(u(t),y(t)), \ \text{ a.e. } \ t\in(0,T) \quad ; \quad y(0)=y_0 \tag{2.1.3}$$

The functions involved in this setting, all of class C^∞, are:

- $\ell : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}$, distributed cost,
- $\phi : \mathbb{R}^n \to \mathbb{R}$, final cost,
- $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n$, dynamics (assumed to be Lipschitz).

Remark 2.1 The existence of solutions in this setting is a difficult question. A coercivity hypothesis on ℓ of the type

$$\exists\,\beta\in\mathbb{R},\ \alpha>0; \quad \ell(u,y) \geq \alpha|u|^2 - \beta \tag{2.1.4}$$

implies that minimizing sequences are bounded in $L^2(0,T,\mathbb{R}^m)$. Therefore a subsequence weakly converges. However, we cannot pass to the limit in the state equation, using the above functional framework. One has to rely on the theory of relaxed controls, see e.g. Ekeland and Temam [14]. In the sequel we assume the existence of a (locally) optimal control.

2.1.2 First-order optimality conditions

We may apply the implicit function theorem to the state equation, viewed as written in the space $L^\infty(0,T,\mathbb{R}^n)$. It follows that the mapping $u \mapsto y_u$ (solution of the state equation) is of class C^∞, $\mathcal{U} \to \mathcal{Y}$. Denote the cost function, expressed a depending on the control only, as

$$J(u) := \int_0^T \ell(u(t), y_u(t))\mathrm{d}t + \phi(y_u(T)) \tag{2.1.5}$$

Then $J(\cdot)$ is of class C^∞ over $\mathcal{U}$. We next show how to compute its first derivative. We define first the *Hamiltonian function* $H : \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ by

$$H(u, y, p) := \ell(u, y) + pf(u, y). \tag{2.1.6}$$

Observe that the state equation may be written as

$$\dot{y}(t) = H_p(u(t), y(t), p(t)) = f(u(t), y(t)) \quad \text{a.e. } t \in [0, T] \quad ; \quad y(0) = y_0. \tag{2.1.7}$$

Next, the adjoint state equation is defined as

$$-\dot{p}(t) = H_y(u(t), y(t), p(t)) \quad \text{a.e. } t \in [0, T], \quad p(T) = D\phi(y(T)). \tag{2.1.8}$$

Introduce the *linearized state equation*

$$\dot{z}(t) = Df(u(t), y(t))(v(t), z(t)) \quad \text{a.e. } t \in [0, T] \quad ; \quad z(0) = 0. \tag{2.1.9}$$

Then for all u and v in $\mathcal{U}$, using the chain rule:

$$DJ(u)v := \int_0^T D\ell(u(t), y_u(t))(v(t), z(t))\mathrm{d}t + D\phi(y_u(T))z(T). \tag{2.1.10}$$

Use

$$\begin{aligned} D\phi(y_u(T))z(T) &= p(T)z(T) = \int_0^T [\dot{p}(t)z(t) + p(t)\dot{z}(t)]\mathrm{d}t \\ &= \int_0^T [-H_y(u(t), y(t), p(t))z(t) + p(t)Df(u(t), y(t))(v(t), z(t))]\mathrm{d}t \\ &= \int_0^T [-\ell_y(u(t), y(t))z(t) + p(t)D_u f(u(t), y(t))v(t)]\mathrm{d}t. \end{aligned} \tag{2.1.11}$$

We deduce that

$$DJ(u)v := \int_0^T H_p(u(t), y(t), p(t))v(t)\mathrm{d}t. \tag{2.1.12}$$

In other words, $H_p(u(t), y(t), p(t))$ is the derivative of J at point u. Therefore

Proposition 2.2 *Let J attain a local minimum at the point $u \in \mathcal{U}$. Then, denoting by y and p the state and costate associated with u, we have*

$$H_p(u(t), y(t), p(t))v(t) = 0, \qquad a.e. \;\; t \in (0, T). \tag{2.1.13}$$

Remark 2.3 The above relations are reminiscent of classical *Hamiltonian systems*, introduced by Hamilton in [15]. The latter are defined as follows. Given a smooth function (the Hamiltonian) $\mathcal{H} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, the associated (dynamical) Hamiltonian system is

$$\dot{y}(t) = \mathcal{H}_p(y(t), p(t)); \quad -\dot{p}(t) = \mathcal{H}_y(y(t), p(t)). \tag{2.1.14}$$

An obvious invariant of the Hamiltonian system is the value of the Hamiltonian itself, since $\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}_y(y(t), p(t)) = \mathcal{H}_y(y(t), p(t))\dot{y}(t) + \mathcal{H}_p(y(t), p(t))\dot{p}(t) = 0$. For mechanical conservative systems, the Hamiltonian function represents the mechanical energy (sum of potential and cinetic energy). In (2.1.7)-(2.1.8) we have the additional "algebraic" variable u, and if u is locally optimal, the additional "algebraic" relation (2.1.13). We show in section 2.1.4 that in some cases u can be eliminated from the algebraic relation.

2.1.3 Pontryaguin's principle

Let $z \in L^1(0,T)$. We say that $t_0 \in]0,T[$ is a *Lebesgue point* of z if

$$z(t_0) = \lim_{\gamma \downarrow 0} \frac{1}{2\gamma} \int_{t_0-\gamma}^{t_0+\gamma} z(t)\mathrm{d}t. \tag{2.1.15}$$

This property is satisfied almost everywhere, see e.g. Rudin [24, theorem 7.7].

Definition 2.4 *We say that $(u,y) \in \mathcal{U} \times \mathcal{Y}$ is a* Pontryagin extremal *it the following holds:*

$$u(t) \in \operatorname*{argmin}_{w \in \mathbb{R}^m} H(w, y(t), p(t)), \qquad a.e. \quad t \in (0,T). \tag{2.1.16}$$

Theorem 2.5 *Let $\bar{u}$ and $\bar{y}$ be an optimal control and the associated optimal state. Then $(\bar{u}, \bar{y})$ is a Pontryagin extremal.*

Proof.
a) Let v be a feasible control, with associated state y. Denote $w := y - \bar{y}$. Since f is Lipschitz, we have that

$$\begin{aligned} \|\dot{w}(t)\| &\leq |f(u(t), y(t)) - f(\bar{u}(t), y(t))| + |f(\bar{u}(t), y(t)) - f(\bar{u}(t), \bar{y}(t))| \\ &\leq O(\|u(t) - \bar{u}(t)\|) + O(\|w(t)\|). \end{aligned}$$

We deduce that

$$\|y_u - \bar{y}\|_\infty = O(\|u - \bar{u}\|_1). \tag{2.1.17}$$

b) Denote by $\bar{p}$ the costate associated with $\bar{u}$. Let v be a feasible control, with associated state y. Set $\Delta := J(v) - J(\bar{u})$. Adding to Δ the null amount

$$\int_0^T \bar{p}(t) \cdot [f(v(t), y(t)) - f(\bar{u}(t), \bar{y}(t)) - \dot{y} + \dot{\bar{y}}]\, dt,$$

obtain $\Delta = A + B$, where

$$\begin{aligned} A &:= \int_0^T [H(v(t), \bar{y}(t), \bar{p}(t)) - H(\bar{u}(t), \bar{y}(t), \bar{p}(t))]\, dt, \\ B &:= \int_0^T [H(v(t), y(t), \bar{p}(t)) - H(v(t), \bar{y}(t), \bar{p}(t))]\, dt + \int_0^T \bar{p}(t) \cdot [\dot{\bar{y}} - \dot{y}]\, dt \\ &\quad + \Phi(y(T)) - \Phi(\bar{y}(T)). \end{aligned}$$

Since $-\frac{\mathrm{d}}{\mathrm{d}t}\bar{p}(t) = H_y(\bar{u}(t), \bar{y}(t))$ and $p(T) = \Phi'(\bar{y}(T))$, integrating by parts the term $\int_0^T \bar{p}(t) \cdot [\dot{\bar{y}} - \dot{y}]\, dt$, we can write $B = B_1 + B_2$, with

$$\begin{aligned} B_1 &= \int_0^T [H(v(t), y(t), \bar{p}(t)) - H(v(t), \bar{y}(t), \bar{p}(t)) - H_y(\bar{u}(t), \bar{y}(t))(y(t) - \bar{y}(t))]\, dt \\ &= \int_0^T [H_y(v(t), \hat{y}(t), \bar{p}(t)) - H_y(\bar{u}(t), \bar{y}(t), \bar{p}(t))(\, y(t) - \bar{y}(t))dt, \\ B_2 &= \Phi(y(T)) - \Phi(\bar{y}(T)) - \Phi'(\bar{y}(T))(y(T) - \bar{y}(T)) \\ &= (\Phi'(\tilde{y}(T))) - \Phi'(\bar{y}(T)))\,(y(T) - \bar{y}(T)), \end{aligned}$$

where (by the mean value theorem) $\hat{y}(t) \in [\bar{y}(t), y(t)]$ for all t, and $\tilde{y} \in [\bar{y}(T), y(T)]$. By (2.1.17), $|B_2| = o(\|v - u\|_1)$. On the other hand, by Lebesgue's theorem,

$$H_y(v(t), \hat{y}(t), \bar{p}(t)) \to H_y(\bar{u}(t), \bar{y}(t), \bar{p}((t)) \quad \text{in } L^1(0, T).$$

Combining with (2.1.17), get

$$|B_1| \leq \|H_y(v, \hat{y}, p) \to H_y(\bar{u}, \bar{y}, p)\|_1 \|\hat{y} - \bar{y}\|_\infty = o(\|v - u\|_1).$$

We have proved that

$$\Delta = A + o(\|v - \bar{u}\|_1). \tag{2.1.18}$$

c) Consider now the *spike perturbations*, i.e.,fix $\gamma > 0$, $t_0 \in]0, T[$, $w \in U$ and

$$v_\gamma(t) = w \;\text{ if }\; |t - t_0| \leq \gamma, \;\; \bar{u}(t) \;\text{ sinon.}$$

Then

$$A = \int_{t_0 - \gamma}^{t_0 + \gamma} [H(w, \bar{y}(t), \bar{p}(t)) - H(\bar{u}(t), \bar{y}(t), \bar{p}(t))]\, dt,$$

and $\|v_\gamma - \bar{u}\|_1 = O(\gamma)$.

Almost each $t_0 \in]0, T[$ is a Lebesgue point of $t \to H(\bar{u}(t), \bar{y}(t), \bar{p}(t))$. Therefore, by (2.1.18), we have, for almost all $t_0 \in]0, T[$,

$$0 \leq \lim_{\gamma \downarrow 0} \frac{J(v_\gamma) - J(\bar{u})}{2\gamma} = H(w, \bar{y}(t_0), p(t_0)) - H(\bar{u}(t_0), \bar{y}(t_0), p(t_0)) \tag{2.1.19}$$

as was to be proved. ■

In addition, it is easy to prove that each Pontryagin extremal is such that the Hamiltonian is constant over the trajectory:

Lemma 2.6 *Let (u, y) be a Pontryagin extremal, and p be the associated costate. Then $t \mapsto H(u(t), y(t), p(t))$ is a constant function (up to a set of measure 0 !).*

Proof.
a) Set $g(t) := \min_{u \in U} H(u, y(t), p(t))$. For $R > \|u\|_\infty$, let $U_R := U \cap B_R$, where B_R is the ball of radius R and center 0 in $\mathbb{R}^m$. Using

$$\begin{array}{lll} |g(t') - g(t)| & \leq & \sup_{u \in \bar{U}_R} |H(u, y(t'), p(t')) - H(u, y(t), p(t))| \\ & \leq & c\left(\|y(t') - y(t)\| + \|p(t') - p(t))\|\right), \end{array} \tag{2.1.20}$$

(with c independent of t and t') as well as the absolute continuity of y and p, we deduce thta g is absolutely continuous. So there exists a set $\mathcal{T} \subset [0, T]$, of full measure in $[0, T]$, such that (2.1.16) is satisfied, and y, p and g are differentiable, for all $t \in \mathcal{T}$. Let $t_0 \in \mathcal{T}$. By (2.1.16), for $t > t_0$, we have

$$\frac{g(t) - g(t_0)}{t - t_0} \leq \frac{H(u(t_0), y(t), p(t)) - H(u(t_0), y(t_0), p(t_0))}{t - t_0}$$

and so with the state and costate equations:

$$\begin{array}{ll} \dot{g}(t_0) & \leq \lim\limits_{t \downarrow t_0} \dfrac{H(u(t_0), y(t), p(t)) - H(u(t_0), y(t_0), p(t_0))}{t - t_0} \\ & = H_y(u(t_0), y(t_0), p(t_0))\dot{y}(t_0) + H_p(u(t_0), y(t_0), p(t_0))\dot{p}(t_0) = 0. \end{array}$$

Taking $t < t_0$, we would prove in a similar way that $\dot{g}(t_0) \geq 0$. Therefore $\dot{g}(t) = 0$ a.e., which since g is absolutely continuous, implies that g is constant. ■

Remark 2.7 We have stated Pontryaguin's principle for a global minimum. However, the proof indicates that it also holds for a local minimum in the topology of $L^1(0, T, \mathbb{R}^m)$. It also holds for a *strong relative minimum* in the sense of calculus of variations, i.e., a point at which the cost function is less or equal than for every other control whose associated state is close in the uniform topology.

2.1.4 Legendre-Clebsch conditions

If $(\bar{u}, y_{\bar{u}})$ is a Pontryaguin extremal, denoting $\bar{y} = y_{\bar{u}}$ and $\bar{p} = p_{\bar{u}}$, then obviously the so-called *weak Legendre-Clebsch condition* holds:

$$D^2_{uu} H(\bar{u}(t), \bar{y}(t), \bar{p}(t)) \succeq 0 \quad \text{a.e.} \tag{2.1.21}$$

It is easily seen that this condition also holds for local minima in $\mathcal{U}$.

We say that a stationary point $\bar{u}$ of J satisfies the *strong Legendre-Clebsch condition* whenever

$$\exists\, \alpha > 0;\; D^2_{uu} H(\bar{u}(t), \bar{y}(t), \bar{p}(t))(v, v) \geq \alpha |v|^2, \quad \text{for all } \; v \in \mathbb{R}^m, \quad \text{a.e. } \; t \in (0, T). \tag{2.1.22}$$

From the proof of Pontryaguin's principle it can be checked that the strong Legendre-Clebsch condition is a necessary condition for quadratic growth (in the sense of proposition 2.16).

Another consequence of the strong Legendre-Clebsch condition is that we can apply the IFT (implicit function theorem) to the stationarity equation

$$D_u H(\bar{u}(t), \bar{y}(t), \bar{p}(t)) = 0. \tag{2.1.23}$$

Since the IFT has a local nature, the strong Legendre-Clebsch condition allows the control to have large jumps, but not small ones. Therefore the following holds.

Proposition 2.8 *Let $\bar{u}$ be a stationary point of J satisfying the strong Legendre-Clebsch condition. Then there exists $\varepsilon > 0$, such that for all $t_0 \in [0,T]$, and $t \in V_\varepsilon(t_0) := [t_0 - \varepsilon, t_0 + \varepsilon] \cap [0,T]$, there exists a C^∞ function $\Upsilon : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m$ such that, either $\bar{u}(t) = \Upsilon(\bar{y}(t), \bar{p}(t))$, or ess sup$\{|\bar{u}(t) - \bar{u}(t')|;\ t, t' \in V_\varepsilon(t_0)\} > \varepsilon$.*

Remark 2.9 If in addition $H(., \bar{y}(t), \bar{p}(t))$ is pseudo-convex (i.e., has convex level sets) for all $t \in [0,T]$, then we obtain that $t \to \bar{u}(t)$ is of class C^∞.

2.1.5 Abstract second-order necessary optimality conditions

For the sake of clarity, we introduce first the second-order optimality conditions in an abstract setting. Let in this subsection $\mathcal{U}$, $\mathcal{Y}$ and $\mathcal{W}$ be arbitrary Banach spaces. Consider a C^2 mapping $\mathcal{A} : \mathcal{U} \times \mathcal{Y} \to \mathcal{W}$. Define the *state equation* as

$$\mathcal{A}(u, y) = 0. \tag{2.1.24}$$

Let (u_0, y_0) be a zero of $\mathcal{A}$ (a solution of (2.1.24)). Assume that $D_y\mathcal{A}(u_0, y_0)$ is invertible. Then by the Implicit Function Theorem, (2.1.24) is locally equivalent to $y = y_u$, where the function $y_u : \mathcal{U} \to \mathcal{Y}$ is of class C^2, and we have for all $v \in \mathcal{U}$

$$y_{u_0+v} = y_0 + z + o(\|v\|), \tag{2.1.25}$$

where $z \in \mathcal{Y}$ is the unique solution of

$$D\mathcal{A}(u_0, y_0)(v, z) = D_u\mathcal{A}(u_0, y_0)v + D_y\mathcal{A}(u_0, y_0)z = 0. \tag{2.1.26}$$

Consider a C^2 cost function $F(u, y)$, with $F : \mathcal{U} \times \mathcal{Y} \to \mathbb{R}$. In a neighborhood of u_0, the *reduced cost* function $J(u) := F(u, y_u)$ is well defined. Let the *Lagrangian function* be defined a s

$$\mathcal{L}(u, y, p) := F(u, y) + \langle p, \mathcal{A}(u, y)\rangle \tag{2.1.27}$$

with here $p \in \mathcal{W}^*$. Let the *costate* $p_u \in \mathcal{W}^*$ be defined as the unique solution of

$$0 = D_y\mathcal{L}(u, y_u, p_u) = D_yF(u, y_u) + D_y\mathcal{A}(u, y_u)^\top p_u. \tag{2.1.28}$$

Locally, $J(u + v)$ is well-defined and equal to $\mathcal{L}(u + v, y_{u+v}, p_u)$. It follows that

$$J(u + v) = \mathcal{L}(u + v, y_{u+v}, p_u) = \mathcal{L}(u, y_u, p_u) + D_u\mathcal{L}(u, y_u, p_u)v + o(\|v\|), \tag{2.1.29}$$

and therefore an expression of the derivative of J is

$$DJ(u) = D_u\mathcal{L}(u, y_u, p_u). \tag{2.1.30}$$

In particular, if J attains a local minimum over a convex set K at the point $\bar{u}$, then the following first-order necessary optimality condition holds:

$$\langle D_u\mathcal{L}(u, y_u, p_u), v - \bar{u}\rangle \geq 0, \quad \text{for all } v \in K. \tag{2.1.31}$$

Remark 2.10 We easily recover of course as a particular case the results of the previous section. We proved there a very interesting regularity result: the derivative of the cost function happens to be (identifiable to) a function in $\mathcal{U}$ (instead of $\mathcal{U}^*$).

Now we compute second-order expansions. Using again $J(u+v) = \mathcal{L}(u+v, y_{u+v}, p_u)$, (2.1.25), and the convention $((x))^2 \equiv (x,x)$:

$$\begin{array}{rcl} J(u+v) & = & \mathcal{L}(u, y_u, p_u) + D_u\mathcal{L}(u, y_u, p_u)v \\ & & +\frac{1}{2}D^2_{((u,y))^2}\mathcal{L}(u, y_u, p_u)((v, y_{u+v} - y_u))^2 + o(|v\|^2), \\ & = & J(u) + D_u\mathcal{L}(u, y_u, p_u)v + \frac{1}{2}D^2_{((u,y))^2}\mathcal{L}(u, y_u, p_u)((v, z))^2 + o(\|v\|^2). \end{array} \tag{2.1.32}$$

Therefore:

Lemma 2.11 *The second-order dervative of J is characterized by*

$$D^2J(\bar{u})(v,v) = D^2_{((u,y))^2}\mathcal{L}(u, y_u, p_u)((v,z))^2, \quad \textit{for all } \ v \in \mathcal{U}. \tag{2.1.33}$$

An immediate consequence is the following second-order necessary optimality condition:

Proposition 2.12 *Let J attain a local (unconstrained) minimum at $\bar{u}$. Then for all $v \in \mathcal{U}$ and z solution of* (2.1.26)*, the following holds:*

$$D^2_{((u,y))^2}\mathcal{L}(u, y_u, p_u)((v,z))^2 \geq 0. \tag{2.1.34}$$

Of course this is nothing else that the condition $D^2J(\bar{u}) \succeq 0$, where "$\succeq 0$" means that the associated quadratic form is nonnegative.

Remark 2.13 As is well-known, a second-order *sufficient* optimality condition is that there exists $\alpha > 0$ such that for all $v \in \mathcal{U}$ and z solution of (2.1.26), the following holds:

$$D^2_{((u,y))^2}\mathcal{L}(u, y_u, p_u)((v,z))^2 \geq \alpha\|v\|^2. \tag{2.1.35}$$

Note however that then the function $v \rightarrow \sqrt{D^2_{((u,y))^2}\mathcal{L}(u, y_u, p_u)((v,z))^2}$ is a norme equivqlent to the one of $\mathcal{U}$. This means that $\mathcal{U}$ is *Hilbertisable* (i.e., endowed with an equivalent norm, is a Hilbert space). So we see that (2.1.35) *never holds* for a non Hilbertisable space like L^s for $s \neq 2$. In particular, it never holds in our application to optimal control ! We wil have to rely on *two norms* second-order sufficient optimality conditions.

2.1.6 Specific second-order necessary optimality condition

We just apply the previous results. The expression of the Lagrangian is

$$\begin{array}{rl} L(u,y,p) & = F(u,y) + \displaystyle\int_0^T p(t)(\ell(u(t), y(t)) - \dot{y}(t))\mathrm{d}t \\ & = \displaystyle\int_0^T H(u(t), y(t), p(t))\mathrm{d}t + \phi(y(T)) - \int_0^T p(t)\dot{y}(t))\mathrm{d}t. \end{array} \tag{2.1.36}$$

Here we may take the multiplier p in $\mathcal{U}$, since we know that the costates associated with control variables are in this space. The last term in the r.h.s. of (2.1.36) being linear in y, has no contribution to the Hessian of the Lagrangian, and it remains

$$D^2J(u)(v,v) = \int_0^T D^2_{((u,y))^2} H(u(t), y_u(t), p_u(t))((v,z))^2 \mathrm{d}t + D^2\phi(y_u(T))(v,v). \quad (2.1.37)$$

Therefore the expression of the second-order necessary optimality condition is as follows:

Proposition 2.14 *Let J attain a local (unconstrained) minimum at $\bar{u}$. Then for all $v \in \mathcal{U}$, z being the solution of the linearized state equation* (2.1.9)*, the expression in the r.h.s. of* (2.1.37) *is nonnegative.*

2.1.7 Second-order sufficient optimality conditions

We know that $u \mapsto J(u)$ is of class C^∞, $\mathcal{U} \to \mathbb{R}$. Therefore, we may write

$$J(u+v) = J(u) + DJ(u)v + \tfrac{1}{2}D^2J(u)(v,v) + r(u,v) \quad (2.1.38)$$

where for fixed u we have, denoting by $\|\cdot\|_s$ the norm in L^s ($s \in [1, +\infty[$):

$$r(u,v) = O(\|v\|^3_\infty). \quad (2.1.39)$$

For the theory of second-order sufficient conditions we need to check that (under appropriate hypotheses) the second-order term of the expansion of J dominates the remainder $r(u,v)$. Since this second-order term involves "integrals of squares" it will be of the order of the L^2 norm. Therefore it is useful to check that $r(u,v)$ is small with respect to the L^2 norm of v. Note that (2.1.39) gives no guarantee in this respect, since no inequality of the type $\|\cdot\|_\infty \leq C\|\cdot\|_2$ holds.

Lemma 2.15 *For any $M > 0$, there exists $c_M > 0$ such that, if $\|u\|_\infty \leq M$ and $\|v\|_\infty \leq M$, then*

$$|r(u,v)| \leq C_M\|v\|_3^3 \leq C_M\|v\|_\infty\|v\|_2^2. \quad (2.1.40)$$

Proof. The last inequality being obvious, we just have to prove the first one. In the sequel we use Gronwall's lemma several times, and often omit the time argument. Using Taylor's expansions up to order q with integral remainders, and since derivatives of any order are Lipschitz on bounded sets, we see that the remainder over a bounded set is uniformly of order $q+1$.

We first obtain an expansion of the mapping y_u. Set $\delta = (v, y_{u+v} - y_u)$, $\delta_y = y_{u+v} - y_u$. Since

$$\dot{\delta}_y(t) = f(u+v, y_{u+v}) - f(u,y) = O(|v(t)| + |y_{u+v}(t) - y_u(t)|) \quad (2.1.41)$$

(with $O(\cdot) \leq c|\cdot|$ uniformly whenever $\|u\|_\infty \leq M$ and say $\|v\|_\infty \leq 1$, we obtain that

$$\|y_{u+v} - y_u\|_\infty = O(\|v\|_1). \quad (2.1.42)$$

Next, set

$$\delta_{yz} := y_{u+v} - y_u - z.$$

We have that

$$\begin{array}{ll} \dot{\delta}_{yz} & = f(u+v, y_{u+v}) - f(u,y) - Df(u,y)(v,z) \\ & = f(u+v, y_{u+v}) - f(u,y) - Df(u,y)(v, y_{u+v} - y_u) + D_y f(u,y)\delta_{yz} \\ & = D_y f(u,y)\delta_{yz} + \frac{1}{2}D^2 f(u,y)((v,\delta_y))^2 + O(|v(t)|^3 + |y_{u+v}(t) - y_u(t)|^3). \end{array} \quad (2.1.43)$$

This proves that

$$y_{u+v} = y_u + z + z_{v,v} + r_{v,v} \quad (2.1.44)$$

where $z_{v,v}$ is solution of

$$\dot{z}_{v,v} = D_y f(u,y) z_{v,v} + \frac{1}{2}D^2 f(u,y)((v,\delta_y))^2 \quad (2.1.45)$$

and

$$r_{v,v}(t) = O(|v(t)|^3 + \|v\|_1^3). \quad (2.1.46)$$

Note that, since $v \to z_{v,v}$ is a quadratic mapping, $z_{v,v}$ is nothing but the second derivative of y_u in direction v. Omitting the time argument, get

$$\ell(u+v, y_{u+v}) \ \ = \ell(u,y_u) + D\ell(u,y_u)(z + z_{v,v}) + \frac{1}{2}D^2\ell(u,y_u)((v,z))^2 + r_\ell(u,v) \quad (2.1.47)$$

and $r_\ell(u,v)$ is the remainder in the second-order expansion (since it includes no linear or quadratic term), and satisfies

$$r_L(u,v)(t) = O(|v(t)|^3 + \|v\|_1^3) = O(|v(t)|^3 + \|v\|_3^3). \quad (2.1.48)$$

Integrating the above relation over time, we obtain the desired result. ■

Proposition 2.16 *Let $u \in \mathcal{U}$ satisfy the second-order sufficient condition:*

$$DJ(u) = 0 \ \ \textit{and} \ \ D^2 J(u)(v,v) \geq \alpha\|v\|_2^2, \ \ \textit{for all} \ \ v \in \mathcal{U}. \quad (2.1.49)$$

Then for all $\alpha' < \alpha$, there exists $\varepsilon > 0$ such that u satisfies the (two-norms) quadratic growth property

$$J(u+v) \geq J(u) + \frac{1}{2}\alpha'\|v\|_2^2, \ \ \textit{for all} \ \ v; \ \|v\|_\infty \leq \varepsilon. \quad (2.1.50)$$

Remark 2.17 The statement of the second-order sufficient condition uses two norms: the L^2 norm for the estimate of increase of the cost function, and the L^∞ norm for the neighborhood.

Remark 2.18 The above results correspond to the following abstract situation. Let the Banach space $\mathcal{U}$ be included in a Hilbert space X, and denote by $\|\cdot\|_\mathcal{U}$, $\|\cdot\|_X$ the norms of $\mathcal{U}$ and X resp. Assume that J is a C^2 function over $\mathcal{U}$, and set $r(u,v) := J(u+v) - J(u) - \frac{1}{2}D^2 J(u)(v,v)$. If $\bar{u} \in \mathcal{U}$ is such that $DJ(\bar{u}) = 0$, and there exist constants $\alpha > 0$, $\varepsilon \in (0,\alpha)$, and $\varepsilon' > 0$ such that

$$\left\{ \begin{array}{ll} D^2 J(\bar{u})(v,v) \geq \alpha\|v\|_X^2, & \text{for all} \ \ v \in \mathcal{U}; \\ |r(u,v)| \leq \frac{1}{2}(\alpha - \varepsilon)\|v\|_X^2, & \text{when} \ \ \|v\|_\mathcal{U} < \varepsilon', \end{array} \right. \quad (2.1.51)$$

then J has a local minimum at $\bar{u}$, and the following quadratic growth condition is satisfied:

$$J(\bar{u}+v) \geq J(\bar{u}) + \frac{1}{4}\varepsilon\|v\|_X^2, \quad \text{when} \ \ \|v\|_\mathcal{U} < \varepsilon'. \quad (2.1.52)$$

2.2 Control constrained problems

In this section we briefly indicate how to deal with control constrained problems, when the control space is $\mathcal{U} = L^\infty(0, T, \mathbb{R}^m)$.

2.2.1 Bound constraints: necessary conditions

We consider here the case when we have the constraint $u \in K$, where

$$K := \{u \in \mathcal{U}; \quad u \geq 0 \text{ a.e. } \} = \mathcal{U}_+ \tag{2.2.53}$$

We first check that the polyhedricity theory applies.

Definition 2.19 *Let C be a closed convex cone of a Banach space X. We assume that C is pointed, i.e., $c \cap (-C) = \{0\}$. The induced order relation over X defined by $a \succeq_C b$, means that $b - a \in C$. We say that w is the* least upper bound *of a and b if $a \preceq_K w$, $b \preceq_K w$, and if $a \preceq_K u$, $b \preceq_K u$ for some $u \in X$, then $w \preceq_K u$.*

We say that C induces a lattice structure *on X if, for any a and b in X, the least upper bound $a \vee b$ exists and the operator $\vee : Y \times Y \to Y$ is continuous.*

We quote the following result [10, Thm. 3.58]:

Proposition 2.20 *Suppose that C induces a lattice structure on X. Then C is polyhedric.*

It immediatly follows that the positive cone of $L^s(0, T, \mathbb{R}^m)$ is, for all $s \in [0, +\infty]$, polyhedric (the same conclusion holds for $C([0, T])$). In particular, $\mathcal{U}_+$ is polyhedric. Therefore:

Proposition 2.21 *Let J attain a local minimum on $\mathcal{U}_+$ at $\bar{u}$. Then*

$$D^2 J(\bar{u})(v, v) \geq 0, \qquad \textit{for all } v \in C(\bar{u}). \tag{2.2.54}$$

We remind that $C(\bar{u})$ is the critical cone, defined by

$$C(\bar{u}) = \{v \in T_{\mathcal{U}_+}(\bar{u}); \quad DJ(\bar{u})v = 0\}. \tag{2.2.55}$$

In the case of the control space $L^2(0, T, \mathbb{R}^m)$, we have given in lemma 1.27 the expression of tangent and normal cones. Unfortunately no such simple expressions hold in the case of $L^\infty(0, T, \mathbb{R}^m)$. Still we have the following, see Cominetti and Penot [11] (our formulation is slightly different, but equivalent):

Proposition 2.22 *Let $u \in \mathcal{U}_+$. For $v \in \mathcal{U}$, and $\varepsilon > 0$, set*

$$a_\varepsilon(v) := \textit{ess sup}\{v(t); \;\; u(t) \geq -\varepsilon\}. \tag{2.2.56}$$

Then $v \in T_K(u)$ iff $\lim_{\varepsilon \downarrow 0} a_\varepsilon(v) \leq 0$.

We will now obtain a stronger second-order necessary condition based on the following observation. Since $v \mapsto D^2J(\bar{u})(v,v)$ is continuous $L^2(0,T,\mathbb{R}^m) \to \mathbb{R}$, obviously (2.2.54) implies

$$D^2J(\bar{u})(v,v) \geq 0, \quad \text{for all } v \in C_2(\bar{u}), \tag{2.2.57}$$

where $C_2(\bar{u})$ is the closure in $L^2(0,T,\mathbb{R}^m)$ of $C(\bar{u})$. We obtain the result below:

Lemma 2.23 *Let J attain a local minimum on $\mathcal{U}$ at $\bar{u}$. Then*

$$D^2J(\bar{u})(v,v) \geq 0, \quad \textit{for all } v \in C_2(\bar{u}), \tag{2.2.58}$$

and

$$C_2(\bar{u}) = \{L^2(0,T,\mathbb{R}^m)_+; \quad v(t)DJ(\bar{u})(t) = 0 \quad \textit{a.e. } \}. \tag{2.2.59}$$

Proof. We only have to prove (2.2.59). So let $\hat{C}_2(\bar{u})$ denote the r.h.s. of (2.2.59). Given $v \in \hat{C}_2(\bar{u})$ and $\varepsilon > 0$, let $v^\varepsilon \in \mathcal{U}$

$$v^\varepsilon := \left\{ \begin{array}{ll} 0 & \text{if } \bar{u}(t) > -\varepsilon, \\ \max(1/\varepsilon, \min(1/\varepsilon, v(t))) & \text{otherwise.} \end{array} \right. \tag{2.2.60}$$

Then $v^\varepsilon \in C(\bar{u})$ and $\lim_{\varepsilon \downarrow 0} v^\varepsilon = v$ in $L^2(0,T,\mathbb{R}^m)$. It follows that $C_2(\bar{u}) \supset \hat{C}_2(\bar{u})$. Since $\hat{C}_2(\bar{u})$ is a closed subset of $L^2(0,T,\mathbb{R}^m)$ containing $C(\bar{u})$, the converse also holds. ■

Remark 2.24 Of course the "stronger" second-order necessary condition of lemma 2.23 can be obtained directly, without refering to the polyhedricity theory. We prefered, however, to show how these concepts are linked.

2.2.2 General sufficient second-order conditions

It is more instructive to state sufficient second-order conditions with (general control) constraints of the type $u \in K$, where here K is any nonempty closed convex subset of $\mathcal{U}$. Let $\bar{u}$ be a stationary point, i.e.

$$DJ(\bar{u})(v - \bar{u}) \geq 0, \quad \text{for all } v \in K. \tag{2.2.61}$$

Define the critical cone

$$C(\bar{u}) := \{v \in T_K(\bar{u}); \ DJ(\bar{u})v = 0\} \tag{2.2.62}$$

as well as its closure in $X := L^2(0,T,\mathbb{R}^m)$:

$$C_2(\bar{u}) := \overline{C(\bar{u})}^{L^2(0,T,\mathbb{R}^m)}. \tag{2.2.63}$$

Proposition 2.25 *Let $u \in \mathcal{U}$ satisfy the second-order sufficient condition:*

$$DJ(u) = 0 \ \textit{ and } \ D^2J(u)(v,v) \geq \alpha \|v\|_2^2, \ \textit{ for all } \ v \in C_2(\bar{u}). \tag{2.2.64}$$

Then u satisfies for some $\alpha' > 0$ a (two-norms) quadratic growth property of the form

$$J(u+v) \geq J(u) + \tfrac{1}{2}\alpha' \|v\|_2^2, \ \textit{ for all } \ v; \ \|v\|_\infty \leq \varepsilon. \tag{2.2.65}$$

The proof is a variant of the one of proposition 1.20. We leave it as an exercice.

2.3 Notes

The stability of solutions to control constrained nonlinear optimal control problems is discussed in Alt [1] The two-norm approach for stability and sensitivity analysis was considered in Dontchev and Hager [12], and Malanowski [18]. Related results can be found in Pales and Zeidan [22, 23]. It is possible to check in certain cases the positiveness of the Hessian of the reduced cost, by solving a differential Riccati equation; see Maurer and Oberle [19].

Bibliography

[1] W. Alt. Stability of solutions to control constrained nonlinear optimal control problems. *Applied Mathematics and Optimization*, 21:53–68, 1990.

[2] A.V. Balakrishnan. *Applied functional analysis*, volume 3 of *Applications of Mathematics*. Springer-Verlag, New York, second edition, 1981.

[3] N. Bérend, J.F. Bonnans, J. Laurent-Varin, M. Haddou, and C. Talbot. An interior-point approach to trajectory optimization. Rapport de Recherche 5613, INRIA, 2005. AIAA J. Guidance, Control and Dynamics, to appear.

[4] J.F. Bonnans. Second order analysis for control constrained optimal control problems of semilinear elliptic systems. *Journal of Applied Mathematics & Optimization*, 38:303–325, 1998.

[5] J.F. Bonnans. *Optimisation Continue*. Dunod, Paris, 2006.

[6] J.F. Bonnans and A. Hermant. No gap second order optimality conditions for optimal control problems with a single state constraint and control. *Mathematical Programming, Series B*, 2006. To appear. Published as INRIA Report 5837.

[7] J.F. Bonnans and A. Hermant. Stability and sensitivity analysis for optimal control problems with a first-order state constraint. Technical report, 2006. Available at http://hal.inria.fr/inria-00087573.

[8] J.F. Bonnans and A. Hermant. Well-posedness of the shooting algorithm for state constrained optimal control problems with a single constraint and control. Technical report, 2006. INRIA Report 5889.

[9] J.F. Bonnans and J. Laurent-Varin. Computation of order conditions for symplectic partitioned runge-kutta schemes with application to optimal control. Rapport de recherche RR-5398, INRIA, 2004. Numerische Mathematik, to appear.

[10] J.F. Bonnans and A. Shapiro. *Perturbation analysis of optimization problems*. Springer-Verlag, New York, 2000.

[11] R. Cominetti and J.P. Penot. Tangent sets to unilateral convex sets. *Comptes Rendus de l'Académie des Sciences de Paris, Série I*, 321:1631–1636, 1995.

[12] A.L. Dontchev and W.W. Hager. Lipschitzian stability in nonlinear control and optimization. *SIAM J. on Control and Optimization*, 31:569–603, 1993.

[13] N. Dunford and J. Schwartz. *Linear operators, Vol I and II.* Interscience, New York, 1958, 1963.

[14] I. Ekeland and R. Temam. *Analyse convexe et problèmes variationnels.* Dunod, Paris, 1974.

[15] W.R. Hamilton. Second essay on a general method in dynamics. *Philosophical Transactions of the Royal Society, Part I*, pages 95–144, 1835. `http://www.emis.de/classics/Hamilton/SecEssay.pdf`.

[16] A. Haraux. How to differentiate the projection on a convex set in Hilbert space. Some applications to variational inequalities. *Journal Mathematical Society of Japan*, 29:615–631, 1977.

[17] A.D. Ioffe and V.M. Tihomirov. *Theory of Extremal Problems.* North-Holland Publishing Company, Amsterdam, 1979. Russian Edition: Nauka, Moscow, 1974.

[18] K. Malanowski. Two-norm approach in stability and sensitivity analysis of optimization and optimal control problems. *Advances in Mathematical Sciences and Applications*, 2:397–443, 1993.

[19] H. Maurer and H.J. Oberle. Second order sufficient conditions for optimal control problems with free final time: the Riccati approach. *SIAM J. Control and Optimization*, 41:380–403 (electronic), 2002.

[20] F. Mignot. Contrôle dans les inéquations variationnelles elliptiques. *Journal of Functional Analysis*, 22:130–185, 1976.

[21] A.A. Milyutin and N. N. Osmolovskii. *Calculus of Variations and Optimal Control.* American Mathematical Society, Providence, 1998.

[22] Z. Páles and V. Zeidan. First- and second-order necessary conditions for control problems with constraints. *Transactions of the American Mathematical Society*, 346:421–453, 1994.

[23] Z. Páles and V. Zeidan. Optimum problems with certain lower semicontinuous set-valued constraints. *SIAM J. on Optimization*, 8:707–727 (electronic), 1998.

[24] W. Rudin. *Real and complex analysis.* Mc Graw-Hill, New York, 1987.

Mathematical Programming manuscript No.
(will be inserted by the editor)

J. Frédéric Bonnans · Audrey Hermant

No-gap Second-order Optimality Conditions for Optimal Control Problems with a Single State Constraint and Control

Dedicated to Professor S.M. Robinson on the occasion of his 65th birthday.

Received: date / Accepted: date

Abstract The paper deals with optimal control problems with only one control variable and one state constraint, of arbitrary order. We consider the case of finitely many boundary arcs and touch times. We obtain a no-gap theory of second-order conditions, allowing to characterize second-order quadratic growth.

Keywords Optimal control · second-order optimality conditions · state constraint · quadratic growth · regular junctions

Mathematics Subject Classification (2000) 49K15 · 34H05 · 90C48

1 Introduction

Considerable efforts have been done in the past for reducing the gap between second-order necessary and sufficient optimality conditions for optimization problems in Banach spaces, with so-called cone constraint (i.e. the constraint mapping must be in a convex cone, or more generally in a convex set). This framework includes many optimal control problems. The theory of second-order necessary optimality conditions involves a term taking into account the curvature of the convex set, see Kawasaki [20], Cominetti [12]. By contrast, second-order sufficient optimality conditions typically involve no such term; see e.g. Maurer and Zowe [29]. We say that a no-gap condition holds, when the only change between necessary or sufficient second-order optimality conditions is between a strict and non strict inequality. In that case it is usually possible to obtain a characterization of the second-order growth condition. There are essentially two cases when no-gap conditions were obtained: (i) the

J. Frédéric Bonnans · Audrey Hermant
INRIA, BP 105, 78153 Le Chesnay, France
E-mail: frederic.bonnans@inria.fr, audrey.hermant@inria.fr

polyedric framework, in the case when the Hessian of Lagrangian is a Legendre form, originating in the work by Haraux [14] and Mignot [30], applied to optimal control problems in e.g. Sokolowski [38] and Bonnans [4], and the extended polyhedricity framework in [9, Section 3.2.3]; this framework essentially covers the case of control constraints (and finitely many final state constraints); and (ii) the second-order regularity framework, introduced in [6] and [5], with applications to semi definite optimization. We refer to [9] for an overview of these theories.

Our paper deals with state-constrained optimal control problems. This occurs in many applications, see e.g. [2,3,10,1]. In optimal control theory, no-gap second-order optimality conditions were known for *mixed* control-state constraints, see e.g. Milutyin-Osmolovskiĭ [31, Part. 2], Osmolovskiĭ [32, 33], and Zeidan [39], whose results use conjugate point theory and Ricatti equations.

Generally speaking, problems with non positivity constraints in spaces of continuous functions do not fit into these frameworks, where no-gap second-order conditions were obtained. The expression of the curvature term in this case was obtained by Kawasaki [22,21] in the one dimensional case, and generalized in Cominetti and Penot [13]. Necessary conditions for variational problems with state constraints taking into account the curvature term can be found in Kawasaki and Zeidan [23]. However, only sufficient conditions without curvature terms were known. Two exceptions are a quite specific situation studied in [6] (with applications to some eigenvalue problems), and the case of finitely many contact points, when the problem can be reduced locally to finitely many inequality constraints in semi-infinite programming, see e.g. Hettich and Jongen [16].

Our main result is the following. By a localization argument, we split the curvature term into a finite number of contributions of boundary arcs and touch points. Using the theory of junction conditions in Jacobson et al. [19] and Maurer [27], we are able to prove that, under quite weak assumptions, the contribution of boundary arcs to the curvature term is zero. For touch points, we use a reduction argument for those that are essential (i.e. that belong to the support of the multiplier) and we make no hypotheses for the non essential ones. The only delicate point is to compute the expansion of the minimum value of a function in $W^{2,\infty}$. Since it is not difficult to state sufficient conditions taking into account essential reducible touch points, we obtain in this way no-gap conditions, that in addition characterize quadratic growth in a convenient two-norms setting.

The paper is organized as follows. In section 2, we recall the material needed, in both points of view of abstract optimization and junction conditions analysis. The main contributions of the paper are in sections 3-5 where the no-gap second-order condition is established. Section 3 states the second-order necessary condition (computation of the curvature term). Section 4 handles the second-order sufficient condition. In section 5, a reduction approach is presented in order to deal with the non-zero part of the curvature term.

2 Framework

We consider the following optimal control problem with a scalar state constraint and a scalar control:

$$(\mathcal{P}) \qquad \min_{u,y} \int_0^T \ell(u(t),y(t))\mathrm{d}t + \phi(y(T)) \tag{1}$$

$$\text{s.t.} \qquad \dot{y}(t) = f(u(t),y(t)) \quad \text{a.e. } t \in [0,T] \quad ; \quad y(0) = y_0 \tag{2}$$

$$g(y(t)) \leq 0 \qquad \forall t \in [0,T]. \tag{3}$$

The data of the problem are the distributed cost $\ell : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$, the final cost $\phi : \mathbb{R}^n \to \mathbb{R}$, the dynamics $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$, the state constraint $g : \mathbb{R}^n \to \mathbb{R}$, the final time $T > 0$, and the initial condition $y_0 \in \mathbb{R}^n$. We make the following assumptions on the data:

(A0) The mappings ℓ, ϕ, f and g are k-times continuously differentiable (C^k) with $k \geq 2$ and have locally Lipschitz continuous second-order derivatives, and the dynamics f is Lipschitz continuous.
(A1) The initial condition satisfies $g(y_0) < 0$.

Throughout the paper, it is assumed that assumption (A0) holds.

2.1 Abstract Optimization

For $1 \leq p \leq \infty$, $L^p(0,T)$ denotes the Banach space of functions such that

$$\|u\|_p := \left(\int_0^T |u(t)|^p \mathrm{d}t\right)^{1/p} < \infty \quad \text{for } p < \infty; \quad \|u\|_\infty := \operatorname{supess} |u(t)| < \infty,$$

and $W^{1,p}(0,T)$ denotes the Sobolev space of functions having a weak derivative in L^p. The space of continuous functions over $[0,T]$ is denoted by $C[0,T]$, with the norm $\|x\|_\infty = \sup |x(t)|$.

Denote by $\mathcal{U} := L^\infty(0,T;\mathbb{R})$ (resp. $\mathcal{Y} := W^{1,\infty}(0,T;\mathbb{R}^n)$) the control (resp. state) space. A *trajectory* is an element $(u,y) \in \mathcal{U} \times \mathcal{Y}$ satisfying the state equation (2). Given $u \in \mathcal{U}$, denote by $y_u \in \mathcal{Y}$ the (unique) solution of (2). Under assumption (A0), by the Cauchy-Lipschitz Theorem, this mapping is well-defined and of class C^k. We may write problem $(\mathcal{P})$ as:

$$\min_{u \in \mathcal{U}} J(u) \quad ; \quad G(u) \in K \tag{4}$$

where $J : \mathcal{U} \to \mathbb{R}$ and $G : \mathcal{U} \to C[0,T]$ are defined, respectively, by $J(u) = \int_0^T \ell(u(t),y_u(t))\mathrm{d}t + \phi(y_u(T))$ and $G(u) = g(y_u)$. These mappings are C^k. Here $K = C_-[0,T]$ is the set of continuous functions over $[0,T]$, with values in $\mathbb{R}_-$.

We say that $u \in \mathcal{U}$ is a (weak) local optimal solution of (4) that satisfies the *quadratic growth condition*, if there exists $\alpha > 0$ and $\rho > 0$ such that:

$$J(\tilde{u}) \geq J(u) + \alpha \|\tilde{u} - u\|_2^2 \qquad \text{for all } \ \tilde{u} \in B_\infty(u,\rho),\ G(\tilde{u}) \in K \tag{5}$$

where $B_\infty(u,\rho)$ denotes the open ball in $L^\infty(0,T)$ with center u and radius ρ. This condition involves two norms, $L^\infty(0,T)$ for the neighborhood, and $L^2(0,T)$ for the growth condition.

The space of row vectors is denoted by $\mathbb{R}^{n*}$. The space of Radon measures, the dual space to $C[0,T]$, is denoted by $\mathcal{M}[0,T]$ and identified with functions of bounded variation vanishing at zero. The cone of nonnegative measures is denoted by $\mathcal{M}_+[0,T]$ and is equal to K^-, the polar cone of K. The duality product over $\mathcal{M}[0,T]\times C[0,T]$ is denoted by $\langle \eta, x\rangle = \int_0^T x(t)\mathrm{d}\eta(t)$. Adjoint operators (and transpose in $\mathbb{R}^n$) are denoted by a star *. Fréchet derivatives of f, etc. w.r.t. arguments $u\in\mathbb{R}$, $y\in\mathbb{R}^n$, are denoted by a subscript, for instance $f_u(u,y)=D_uf(u,y)$, $f_{uu}(u,y)=D^2_{uu}f(u,y)$, etc.

Define the classical *Hamiltonian* and *Lagrangian* functions of problem $(\mathcal{P})$, respectively $H:\mathbb{R}\times\mathbb{R}^n\times\mathbb{R}^{n*}\to\mathbb{R}$ and $L:\mathcal{U}\times\mathcal{M}[0,T]\to\mathbb{R}$ by:

$$H(u,y,p):=\ell(u,y)+pf(u,y) \quad ; \quad L(u,\eta):=J(u)+\langle \eta, G(u)\rangle\,. \tag{6}$$

Denote by $BV(0,T)$ the space of functions of bounded variation. Given $u\in\mathcal{U}$ and $\eta\in\mathcal{M}_+[0,T]$, let the costate $p_{u,\eta}$ be the unique solution in $BV(0,T;\mathbb{R}^{n*})$ of:

$$-\mathrm{d}p_{u,\eta}=(\ell_y(u,y_u)+p_{u,\eta}f_y(u,y_u))\mathrm{d}t+g_y(y_u)\mathrm{d}\eta\ ;\quad p_{u,\eta}(T)=\phi_y(y_u(T)). \tag{7}$$

Given $v\in\mathcal{U}$, let the linearized state $z_{u,v}\in\mathcal{Y}$ be solution of:

$$\dot z_{u,v}=f_y(u,y_u)z_{u,v}+f_u(u,y_u)v \qquad ; \qquad z_{u,v}(0)=0. \tag{8}$$

The mapping $\mathcal{U}\to\mathcal{Y}$, $v\mapsto z_{u,v}$ is the Fréchet derivative of the mapping $u\mapsto y_u$ at point u.

The next lemma gives the expressions of derivatives of Lagrangian, with respect to the control. For simplicity of notation, we write in the sequel $D^2H_{(u,y)^2}(u,y,p)(v,z)^2$ instead of $D^2_{(u,y),(u,y)}H(u,y,p)((v,z),(v,z))$.

Lemma 1 *Let $\eta\in\mathcal{M}_+[0,T]$. Then $u\mapsto L(u,\eta)$ is of class C^2 over $\mathcal{U}$, with first and second derivatives given by, for all $v\in\mathcal{U}$ (omitting time argument):*

$$D_uL(u,\eta)v=\int_0^T H_u(u,y_u,p_{u,\eta})v\mathrm{d}t, \tag{9}$$

$$\begin{aligned} D^2_{uu}L(u,\eta)(v,v) \ =\ & \int_0^T D^2H_{(u,y)^2}(u,y_u,p_{u,\eta})(v,z_{u,v})^2\mathrm{d}t \\ & + z_{u,v}(T)^*\phi_{yy}(y_u(T))z_{u,v}(T) \ + \int_0^T z^*_{u,v}g_{yy}(y_u)z_{u,v}\mathrm{d}\eta, \end{aligned} \tag{10}$$

where H is given by (6), $z_{u,v}$ and $p_{u,\eta}$ are the solutions, respectively, to (8) and (7).

Proof Since $u\mapsto y_u$ is C^2, the Cauchy-Lipschitz Theorem ensures the existence of the second-order expansion of the state

$$y_{u+v}=y_u+z_{u,v}+\frac{1}{2}z_{u,vv}+o\left(\|v\|^2_\infty\right). \tag{11}$$

It is easily seen, substituting (11) into the state equation and keeping the terms of second-order, that $z_{u,vv}$ is solution of:

$$\dot{z}_{u,vv} = f_y(u, y_u)z_{u,vv} + D^2 f_{(u,y)^2}(u, y_u)(v, z_{u,v})^2 \quad ; \quad z_{u,vv}(0) = 0. \qquad (12)$$

Using costate equation (7) and linearized state equations (8) and (12), we get easily (omitting arguments):

$$\begin{aligned} D_u L(u, \eta)v = \quad & - \int_0^T (\mathrm{d}p_{u,\eta} z_{u,v} + p_{u,\eta} \dot{z}_{u,v} \mathrm{d}t) + \phi_y(y_u(T))z_{u,v}(T) \\ & \qquad + \int_0^T H_u v \mathrm{d}t; \\ D^2_{uu} L(u, \eta)(v, v) = & \int_0^T D^2 H_{(u,y)^2}(v, z_{u,v})^2 \mathrm{d}t + z_{u,v}(T)^* \phi_{yy}(y_u(T))z_{u,v}(T) \\ & \qquad + \int_0^T z^*_{u,v} g_{yy}(y_u) z_{u,v} \mathrm{d}\eta \\ & - \int_0^T (\mathrm{d}p_{u,\eta} z_{u,vv} + p_{u,\eta} \dot{z}_{u,vv} \mathrm{d}t) + \phi_y(y_u(T))z_{u,vv}(T). \end{aligned}$$

To obtain (9) and (10) it suffices, in view of Lemma 33 in the Appendix, to integrate by parts in the above expressions $p_{u,\eta}$ with $z_{u,v}$ and with $z_{u,vv}$, respectively.

First Order Necessary Condition. For $x \in K = C_-(0, T)$, define the *first order contact set* $I(x) := \{t \in [0, T] \ ; \ x(t) = 0\}$. The expression of the *tangent* and *normal* cones (in the sense of convex analysis) to K at point x, respectively $T_K(x)$ and $N_K(x)$, are well-known (see e.g. [9]) and given, for $x \in K$ (these sets being empty if $x \notin K$), by:

$$\begin{aligned} T_K(x) &= \{h \in C[0, T] \ ; \ h(t) \leq 0 \text{ on } I(x)\}, \\ N_K(x) &= \{\eta \in \mathcal{M}_+[0, T] \ ; \ \mathrm{supp}(\mathrm{d}\eta) \subset I(x)\}. \end{aligned}$$

Here by $\mathrm{supp}(\mathrm{d}\eta)$ we denote the *support* of the measure $\eta \in \mathcal{M}[0, T]$, i.e. the complement in $[0, T]$ of the largest open set $W \subset [0, T]$ that satisfies: $\int_0^T x(t)\mathrm{d}\eta(t) = 0$, for all functions $x \in C[0, T]$ vanishing on $[0, T] \setminus W$.

Let $u \in \mathcal{U}$. We say that $\eta \in \mathcal{M}_+[0, T]$ is a *Lagrange multiplier* associated with u if the following first order necessary optimality condition holds:

$$D_u L(u, \eta) = DJ(u) + DG(u)^* \eta = 0 \quad ; \quad \eta \in N_K(G(u)). \qquad (13)$$

The set of Lagrange multipliers associated with u is denoted by $\Lambda(u)$.

Robinson's constraint qualification (see [35,36]) for problem (4) is as follows:

$$\exists\, \varepsilon > 0, \qquad \varepsilon B_C \subset G(u) + DG(u)\mathcal{U} - K. \qquad (14)$$

Here B_C denotes the unit (open) ball of $C[0, T]$.

The next theorem is well-known (see e.g. [9], Lemma 2.98 and Theorem 3.9). Note that for $v \in \mathcal{U}$, we have $DG(u)v = g_y(y_u)z_{u,v}$, i.e., $(DG(u)v)(t) = g_y(y_u(t))z_{u,v}(t)$, for all $t \in [0, T]$.

Theorem 2 (i) *A characterization of (14) is:*

$$\textit{There exists } v \in \mathcal{U}; \quad g_y(y_u(t))z_{u,v}(t) < 0, \textit{ for all } t \in I(g(y_u)). \tag{15}$$

(ii) *Let u be a local solution of (4), satisfying (15). Then with u is associated a non empty and bounded set of Lagrange multipliers.*

Second Order Analysis. Let the *critical cone* be defined by:

$$C(u) = \{v \in \mathcal{U} \ ; \ DG(u)v \in T_K(G(u)) \ ; \ DJ(u)v \leq 0\}. \tag{16}$$

For $h \in T_K(x)$, the *second-order contact set* is defined by:

$$I^2(x,h) = \{t \in I(x) \ ; \ h(t) = 0\}. \tag{17}$$

If (13) holds, then $DJ(u)v \geq 0$ for all v such that $DG(u)v \in T_K(G(u))$ and $DJ(u)v = 0$ iff $\eta \perp DG(u)v$. Since $\eta \geq 0$ has support in $I(G(u))$, and $DG(u)v \leq 0$ on $I(G(u))$, we obtain the following (classical) statement:

Lemma 3 *Let (u,η) satisfy the first order necessary condition (13). Then:*

$$C(u) = \{v \in \mathcal{U}; \ DG(u)v \in T_K(G(u)); \ \mathrm{supp}(\mathrm{d}\eta) \subset I^2(G(u), DG(u)v)\}. \tag{18}$$

The *inner* and *outer second-order tangent sets*, respectively $T_K^{2,i}(x,h)$ and $T_K^2(x,h)$, are defined by:

$$\begin{aligned} T_K^{2,i}(x,h) &:= \{w \in C[0,T]; \ \mathrm{dist}(x + \varepsilon h + \tfrac{1}{2}\varepsilon^2 w, K) = o(\varepsilon^2), \ \varepsilon \geq 0\}, \\ T_K^2(x,h) &:= \{w \in C[0,T]; \ \exists \varepsilon_n \downarrow 0, \mathrm{dist}(x + \varepsilon_n h + \tfrac{1}{2}\varepsilon_n^2 w, K) = o(\varepsilon_n^2)\}. \end{aligned}$$

We recall the characterization of the inner second-order tangent set $T_K^{2,i}(x,h)$ due to Kawasaki [22,21] (see also Cominetti [13]): if $x \in K$ and $h \in T_K(x)$, then

$$T_K^{2,i}(x,h) = \{w \in C[0,T] \ ; \ w(t) \leq \varsigma_{x,h}(t) \text{ on } [0,T]\}, \tag{19}$$

where $\varsigma_{x,h} : [0,T] \to \overline{\mathbb{R}}$ is given by:

$$\varsigma_{x,h}(t) = \begin{cases} 0 & \text{if } t \in (\mathrm{int}\, I(x)) \cap I^2(x,h) \\ \liminf\limits_{t' \to t \,;\, x(t') < 0} \dfrac{(h(t')_+)^2}{2x(t')} & \text{if } t \in \partial I(x) \cap I^2(x,h) \\ +\infty & \text{otherwise.} \end{cases} \tag{20}$$

Here $h(t)_+ := \max\{h(t), 0\}$, and $\mathrm{int}\, S$ and ∂S denote respectively the interior and boundary of set S. Set $\mathcal{T}(x,h) := \partial I(x) \cap I^2(x,h)$. We have $\varsigma_{x,h}(\tau) \leq 0$ for $\tau \in \mathcal{T}(x,h)$ and it is not difficult to check that $t \mapsto \varsigma_{x,h}(t)$ is lower semi-continuous. Consequently, $T_K^{2,i}(x,h) \neq \emptyset$ iff $\varsigma_{x,h}(t) > -\infty$ for all t. In that case, $\varsigma_{x,h}$ is the upper limit of a increasing sequence of continuous functions (ς_n). Given $\eta \in \mathcal{M}_+[0,T]$, we may define (see e.g. [22]):

$$\int_0^T \varsigma_{x,h}(t)\mathrm{d}\eta(t) := \sup \left\{ \int_0^T \varsigma(t)\mathrm{d}\eta(t); \ \varsigma \leq \varsigma_{x,h} \right\} \ \in \ \mathbb{R} \cup \{+\infty\}.$$

Then:

$$\sigma(\eta, T_K^{2,i}(x,h)) = \int_0^T \varsigma_{x,h}(t) \mathrm{d}\eta(t), \tag{21}$$

where $\sigma(\eta, S) = \sup_{w \in S} \langle \eta, w \rangle$ denotes the support function of the set S. If the support of η satisfies $\mathrm{supp}(\mathrm{d}\eta) \subset I^2(x,h)$, then

$$\sigma(\eta, T_K^{2,i}(x,h)) \leq 0. \tag{22}$$

A second-order necessary condition due to Kawasaki [20] is:

Theorem 4 *Let u be a local optimal solution of (4) satisfying (14). Then, for all $v \in C(u)$, the following holds:*

$$\sup_{\eta \in \Lambda(u)} \left\{ D^2_{uu} L(u,\eta)(v,v) - \sigma(\eta, T_K^{2,i}(G(u), DG(u)v)) \right\} \geq 0. \tag{23}$$

Remark 5 The above second-order necessary condition was improved by Cominetti in [12], by stating that for all convex set $\mathcal{S}_{u,v} \subset T_K^2(G(u), DG(u)v)$,

$$\sup_{\eta \in \Lambda(u)} \left\{ D^2_{uu} L(u,\eta)(v,v) - \sigma(\eta, \mathcal{S}_{u,v}) \right\} \geq 0. \tag{24}$$

Th. 4 is obtained for the particular choice of $\mathcal{S}_{u,v} = T_K^{2,i}(G(u), DG(u)v)$. For the problem considered in the present paper, we gain sufficient information from (23) (see Proposition 14).

2.2 Junction Condition Analysis

We first recall some classical definitions. A *boundary* (resp. *interior*) *arc* is a maximal interval of positive measure $\mathcal{I} \subset [0,T]$ such that $g(y(t)) = 0$ (resp. $g(y(t)) < 0$) for all $t \in \mathcal{I}$. If $[\tau_{en}, \tau_{ex}]$ is a boundary arc, τ_{en} and τ_{ex} are called *entry* and *exit* point, respectively. Entry and exit points are said to be *regular* if they are endpoints of an interior arc. A *touch* point τ in $(0,T)$ is an isolated contact point (endpoint of two interior arcs). Entry, exit and touch points are called *junction points* (or *times*). We say that the junctions are regular, when the junction points are regular. In this paper, only the case of finitely many regular junctions is dealt with.

The first-order time derivative of the state constraint when y satisfies the state equation (2), i.e., $g^{(1)}(u,y) = \frac{\mathrm{d}}{\mathrm{d}t} g(y(t)) = g_y(y) f(u,y)$, is denoted by $g^{(1)}(y)$ if the function $\mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$; $(u,y) \mapsto g_y(y) f(u,y)$ does not depend on u (that is, the function $(u,y) \mapsto g^{(1)}_u(u,y)$ is identically zero). We may define similarly $g^{(2)}, \ldots, g^{(q)}$ if g, f are C^q and if $g^{(j)}_u \equiv 0$, for all $j = 1, \ldots, q-1$, and we have $g^{(j)}(u,y) = g^{(j-1)}_y(y) f(u,y)$, for $j = 1, \ldots, q$.

Let $q \geq 1$ be the smallest number of times derivations of the state constraint, so that a dependence w.r.t. u appears, i.e. $g^{(q)}_u \not\equiv 0$. If q is finite, we say that q is the *order* of the state constraint (see e.g. Bryson et al. [11]).

Let $u \in \mathcal{U}$ be a solution of the first order necessary condition (13), with Lagrange multiplier η and costate $p_{u,\eta}$ solution of (7). Since η and $p_{u,\eta}$ are of bounded variation, they have at most countably many discontinuity times, and are everywhere on $[0,T]$ left and right continuous. We denote by $[\eta(\tau)] = \eta(\tau^+) - \eta(\tau^-)$ where $\eta(\tau^\pm) = \lim_{t\to\tau^\pm} \eta(t)$ the jump discontinuity of η at time $\tau \in [0,T]$. We make the following assumptions:

(A2) The Hamiltonian is strongly convex w.r.t. the control variable, uniformly w.r.t. $t \in [0,T]$:

$$\exists\, \gamma > 0, \qquad H_{uu}(\hat{u}, y_u(t), p_{u,\eta}(t^\pm)) \geq \gamma \qquad \forall \hat{u} \in \mathbb{R},\ \forall t \in [0,T]. \tag{25}$$

(A3) (Constraint regularity) The data of the problem are C^{2q}, i.e. $k \geq 2q$ in (A0), the state constraint is of order q and the condition below holds:

$$\exists\, \beta > 0, \qquad |g_u^{(q)}(\hat{u}, y_u(t))| \geq \beta, \qquad \forall \hat{u} \in \mathbb{R},\ \forall t \in [0,T]. \tag{26}$$

(A4) The trajectory (u, y_u) has a *finite set of junction times*, that will be denoted by $\mathcal{T} =: \mathcal{T}_{en} \cup \mathcal{T}_{ex} \cup \mathcal{T}_{to}$, with $\mathcal{T}_{en}$, $\mathcal{T}_{ex}$ and $\mathcal{T}_{to}$ the *disjoint* (and possibly empty) subsets of respectively regular entry, exit and touch points, and we suppose that $g(y_u(T)) < 0$.

Remark 6 1) An assumption weaker than (A2), that is enough for the sufficient conditions in section 4 and 5, is

(A2') (Strengthened Legendre-Clebsch condition)

$$\exists\, \gamma > 0, \qquad H_{uu}(u(t), y_u(t), p_{u,\eta}(t)) \geq \gamma \qquad \text{a.e. } t \in [0,T]. \tag{27}$$

Condition (27) does not imply the continuity of the control.
2) In assumption (A3), it is in fact sufficient to assume that (26) holds for t in the neighborhood of the contact set $I(g(y_u))$. In the definition of the order of the constraint q, it is sufficient as well to restrict the variable y to a neighborhood in $\mathbb{R}^n$ of $\{y_u(t)\ ;\ t \in I(g(y_u))\}$.

A touch point $\tau \in \mathcal{T}_{to}$ is said to be *essential*, if the Lagrange multiplier η satisfies $[\eta(\tau)] > 0$. The set of essential touch points of the trajectory (u, y_u) will be denoted by $\mathcal{T}_{to}^{ess}$.

The above hypotheses imply the continuity of the control variable and of some of its derivatives at junction points. The next proposition is due to Jacobson et al. [19]. Its proof was later clarified in Maurer [27], see also the survey by Hartl et al. [15].

Proposition 7 *Let $u \in \mathcal{U}$ satisfying (13) with Lagrange multiplier η and assume that (A2)-(A4) hold. Then:*

(i) *The control u is continuous over* $[0,T]$ *(in particular at junction points $\tau \in \mathcal{T}$) and C^q on* $[0,T]\backslash\mathcal{T}$. *The multiplier η is continuously differentiable on* $[0,T] \setminus \mathcal{T}$.
(ii) *If $\tau \in \mathcal{T}_{en} \cup \mathcal{T}_{ex}$ is a regular entry or exit point, then:* (a) *if q is* odd, *η and the $q-1$ first time derivatives of u are continuous at τ;* (b) *if q is* even, *the $q-2$ first time derivatives of u are continuous at τ.*
(iii) *If $\tau \in \mathcal{T}_{to}$ is a touch point, then:* (a) *the $q-2$ first derivatives of u are continuous at τ;* (b) *if $q = 1$, then η and $\dot{u}$ are also continuous at τ (that is, if $q = 1$, then (u, y_u) does not have essential touch point).*

Remark 8 Under the assumptions of Prop. 7, we have the following decomposition: $\mathrm{d}\eta(t) = \eta_0(t)\mathrm{d}t + \sum_{\tau \in \mathcal{T}} \nu_\tau \delta_\tau(t)$ where δ_τ denotes the Dirac measure at time τ, the density $\eta_0 \in L^1(0,T)$ is equal to $\frac{\mathrm{d}\eta}{\mathrm{d}t}$ almost everywhere and $\nu_\tau := [\eta(\tau)] \geq 0$. We have $\nu_\tau = 0$ if q is odd and τ is a regular entry/exit point, and if $q = 1$ and τ is a touch point.

We end this section by a result on constraint qualification and uniqueness of the multiplier. For this we need the expression of the time derivatives of $DG(u)v$.

Lemma 9 *Assume that f, g are C^q and that $g_u^{(j)} \equiv 0$, for $j = 1, \dots, q-1$. Then: (i) For all $v \in \mathcal{U}$, the following relations hold:*

$$\frac{\mathrm{d}^j}{\mathrm{d}t^j} g_y(y_u) z_{u,v} = g_y^{(j)}(u, y_u) z_{u,v}, \qquad j = 1, \dots, q-1, \tag{28}$$

$$\frac{\mathrm{d}^q}{\mathrm{d}t^q} g_y(y_u) z_{u,v} = g_y^{(q)}(u, y_u) z_{u,v} + g_u^{(q)}(u, y_u) v. \tag{29}$$

(ii) If in addition, (26) is satisfied, then $DG(u)$ is an isomorphism between $L^\infty(0,T)$ and the space $\mathcal{W}$ defined by:

$$\mathcal{W} := \{\varphi \in W^{q,\infty}(0,T) \; ; \; \varphi^{(j)}(0) = 0 \; ; \; j = 0, \dots, q-1\}. \tag{30}$$

Proof (i) By (8), we have:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} g_y(y_u) z_{u,v} &= g_{yy}(y_u) f(u, y_u) z_{u,v} + g_y(y_u) f_y(u, y_u) z_{u,v} + g_y(y_u) f_u(u, y_u) v \\ &= g_y^{(1)}(u, y_u) z_{u,v} + g_u^{(1)}(u, y_u) v. \end{aligned}$$

Since $g_u^{(j)} \equiv 0$ for $j = 1$ to $q-1$, we obtain by induction that $\frac{\mathrm{d}^j}{\mathrm{d}t^j} g_y(y_u) z_{u,v} = g_y^{(j)}(u, y_u) z_{u,v}$ is independent on v, and that the derivative of order q has the expression in (29).
(ii) If in addition (26) is satisfied, it is easily seen by (29) that for all $\varphi \in \mathcal{W}$, there exists a unique $v \in \mathcal{U}$ such that $g_y(y_u) z_{u,v} = \varphi$. The conclusion follows from the open mapping theorem.

Proposition 10 *Assume that (A1) holds, and let $u \in \mathcal{U}$ satisfy (A3). Then: (i) Robinson's constraint qualification (14) holds; (ii) if $\Lambda(u) \neq \emptyset$, the Lagrange multiplier η associated with u is unique.*

Proof It is obvious by Lemma 9(ii) and Th. 2(i) that (14) holds iff (A1) does. This proves (i). Assume that $\eta_1, \eta_2 \in \Lambda(u)$ and set $\mu := \eta_2 - \eta_1 \in \mathcal{M}[0,T]$. Since $DG(u)^* \mu = 0$, it follows that $\int_0^T \varphi(t) \mathrm{d}\mu(t) = 0$, for all $\varphi \in \mathcal{W}$. Since $g(y_0) < 0$, we have $\mathrm{supp}(\mathrm{d}\mu) \subset [2\varepsilon, T]$ for some $\varepsilon > 0$. Taking the restriction to $[\varepsilon, T]$ of functions in $DG(u)\mathcal{U}$, we obtain the whole space $W^{q,\infty}(\varepsilon, T)$. By density of the latter in $C[\varepsilon, T]$ we deduce that for all $\varphi \in C[0,T]$, $\int_0^T \varphi(t) \mathrm{d}\mu(t) = \int_\varepsilon^T \varphi(t) \mathrm{d}\mu(t) = 0$. Hence $\mathrm{d}\mu \equiv 0$, which achieves the proof of (ii).

3 Second-order Necessary Conditions

3.1 Basic Second-order Necessary Conditions

Let $u \in \mathcal{U}$ satisfy assumptions (A2)-(A4) and $\eta \in \Lambda(u)$. We make the following assumptions. Let $\hat{q} := 2q-1$ if q is *odd* and $\hat{q} := 2q-2$ if q is *even*.

(A5) (Non Tangentiality Condition)
(i) For all entry times $\tau_{en} \in \mathcal{T}_{en}$ and all exit times $\tau_{ex} \in \mathcal{T}_{ex}$:

$$(-1)^{\hat{q}+1}\frac{\mathrm{d}^{\hat{q}+1}}{\mathrm{d}t^{\hat{q}+1}}g(y_u(t))|_{t=\tau_{en}^-} < 0 \ ; \quad \frac{\mathrm{d}^{\hat{q}+1}}{\mathrm{d}t^{\hat{q}+1}}g(y_u(t))|_{t=\tau_{ex}^+} < 0. \quad (31)$$

(ii) For all *essential* touch points $\tau_{to} \in \mathcal{T}_{to}^{ess}$:

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}g(y_u(t))|_{t=\tau_{to}} < 0. \quad (32)$$

(A6) (Strict Complementarity *on boundary arcs*): $\operatorname{int} I(G(u)) \subset \operatorname{supp}(\mathrm{d}\eta)$.

Remark 11 1) By Proposition 7, the expressions appearing in assumption (A5)(i)-(ii) are well-defined, and $\hat{q}+1$ is the smallest possible order for which the corresponding derivative of $g(y_u)$ may be nonzero at an entry or exit point. Therefore assumption (A5) does not contradict the junction conditions in Prop. 7. Note that $\hat{q} = q$ for $q = 1, 2$.
2) Only the assumption (A6') below, weaker than (A6), is used in necessary condition of Theorem 12, in order to ensure that the second-order tangent set $T_K^{2,i}(G(u), DG(u)v)$ is not empty, for all $v \in C(u)$:

(A6') (Strict Complementarity *near entry/exit* of boundary arcs): For all entry points $\tau_{en} \in \mathcal{T}_{en}$ and exit points $\tau_{ex} \in \mathcal{T}_{ex}$, there exists $\varepsilon > 0$ such that:

$$(\tau_{en}, \tau_{en}+\varepsilon) \ \subset \ \operatorname{supp}(\mathrm{d}\eta) \quad ; \quad (\tau_{ex}-\varepsilon, \tau_{ex}) \ \subset \ \operatorname{supp}(\mathrm{d}\eta). \quad (33)$$

Note that we do not assume strict complementarity at touch points.

Theorem 12 *Assume that (A1) holds. Let $u \in \mathcal{U}$ be an optimal solution of (4), with its Lagrange multiplier η, satisfying (A2)-(A5) and (A6'). Let $\mathcal{T}_{to}^{ess}$ denote the (finite) set of essential touch points of the trajectory (u, y_u) and $\nu_\tau = [\eta(\tau)] > 0$, for $\tau \in \mathcal{T}_{to}^{ess}$. Then, for all $v \in C(u)$:*

$$D_{uu}^2 L(u,\eta)(v,v) - \sum_{\tau \in \mathcal{T}_{to}^{ess}} \nu_\tau \frac{(g_y^{(1)}(y_u(\tau))z_{u,v}(\tau))^2}{\frac{\mathrm{d}^2}{\mathrm{d}t^2}g(y_u(t))|_{t=\tau}} \geq 0. \quad (34)$$

Corollary 13 *Under the assumptions of Theorem 12, if the trajectory (u, y_u) has no essential touch point (in particular, if the state constraint is of* first order *$q = 1$), then $D_{uu}^2 L(u,\eta)(v,v) \geq 0$, for all $v \in C(u)$.*

In the sequel, we denote $I^2(G(u), DG(u)v)$ by $I^2_{u,v}$. For all $v \in C(u)$, by (18), we have $\mathcal{T}_{to}^{ess} \subset (\mathcal{T}_{to} \cap I^2_{u,v})$. Let us denote the subset of critical directions that "avoid" non essential touch point (i.e., such that $g(y_u(\tau))z_{u,v}(\tau) < 0$, for all $\tau \in \mathcal{T}_{to} \setminus \mathcal{T}_{to}^{ess}$) by:

$$C_0(u) := \{v \in C(u) \ ; \ \mathcal{T}_{to} \cap I^2_{u,v} = \mathcal{T}_{to}^{ess}\}.$$

The first step of the proof of Theorem 12 consists in computing the sigma-term for the critical directions in $C_0(u)$.

Proposition 14 *Let $v \in C_0(u)$. Under the assumptions of Theorem 12, we have that*

$$\sigma(\eta, T_K^{2,i}(G(u), DG(u)v)) = \sum_{\tau \in \mathcal{T}_{to}^{ess}} \nu_\tau \frac{(g_y^{(1)}(y_u(\tau))z_{u,v}(\tau))^2}{\frac{\mathrm{d}^2}{\mathrm{d}t^2} g(y_u(t))|_{t=\tau}}. \tag{35}$$

Proof The proof is divided into 3 steps. We first analyse the contribution of entry/exit points, then the one of touch points, and finally conclude.

Remind that by (20), only the points in $\partial I(G(u)) \cap I^2_{u,v}$ have a contribution to the sigma term. Note that $\partial I(G(u)) = \mathcal{T}$. Set $\varsigma_{u,v} := \varsigma_{g(y_u), g_y(y_u)z_{u,v}} = \varsigma_{G(u),DG(u)v}$ and let $\tau \in \mathcal{T} \cap I^2_{u,v}$. By (20), we have:

$$\varsigma_{u,v}(\tau) \ = \ \liminf_{t \to \tau; \ g(y_u(t))<0} \frac{(\{g_y(y_u(t))z_{u,v}(t)\}_+)^2}{2g(y_u(t))}. \tag{36}$$

1) (Entry/exit point). Assume that $\tau \in \mathcal{T}_{en} \cup \mathcal{T}_{ex}$. According to Prop. 7(ii), time derivatives of the control at regular entry/exit points are continuous until order $q-2$ if q is even, and $q-1$ if q is odd. Consequently, by definition of the order of the state constraint, the time derivatives of $g(y_u)$ are continuous at τ until order $2q-2$ is q is even, and $2q-1$ if q is odd. Hence they all vanish at entry/exit time τ of a boundary arc. It follows that for t in a neighborhood of τ on the interior arc side, a Taylor expansion gives, by definition of $\hat{q}$:

$$g(y_u(t)) = \frac{\mathrm{d}^{\hat{q}+1}}{\mathrm{d}t^{\hat{q}+1}} g(y_u)|_{t=\tau^\pm} \frac{(t-\tau)^{\hat{q}+1}}{(\hat{q}+1)!} + o((t-\tau)^{\hat{q}+1}), \tag{37}$$

where, for the sake of simplicity, we denote by $\tau^\pm$ either τ^- if $\tau \in \mathcal{T}_{en}$ or τ^+ if $\tau \in \mathcal{T}_{ex}$.

Combining Lemma 3 and (A6'), we see that for all $v \in C(u)$, the function (of time) $g_y(y_u)z_{u,v}$ vanishes just after entering or before leaving a boundary arc on a small interval $[\tau, \tau \pm \varepsilon]$, and so do its first $q-1$ time derivatives since the latter are continuous by Lemma 9(i). The q-th derivative of $g_y(y_u)z_{u,v}$ being a bounded function by (29), we have, on the interior arc side:

$$|g_y(y_u(t))z_{u,v}(t)| \leq C|t-\tau|^q. \tag{38}$$

If q is *odd*, combining (37) with $\hat{q} = 2q-1$ and (38) and by tangentiality assumption (A5)(i), we deduce from (36) that:

$$\varsigma_{u,v}(\tau) \geq \lim_{t \to \tau^\pm} \frac{C^2(t-\tau)^{2q}}{\frac{\mathrm{d}^{2q}}{\mathrm{d}t^{2q}} g(y_u)|_{t=\tau^\pm} \frac{(t-\tau)^{2q}}{(2q)!} + o((t-\tau)^{2q})} \ > \ -\infty.$$

If q is *even*, (37) with $\hat{q} = 2q-2$, (38) and (A5)(i) in (36) give:

$$\varsigma_{u,v}(\tau) \geq \lim_{t\to\tau^{\pm}} \frac{C^2(t-\tau)^{2q}}{\frac{\mathrm{d}^{2q-1}}{\mathrm{d}t^{2q-1}} g(y_u)|_{t=\tau^{\pm}} \frac{(t-\tau)^{2q-1}}{(2q-1)!} + o((t-\tau)^{2q-1})} \;=\; 0.$$

Since $\varsigma_{u,v}(\tau) \leq 0$ by (20) at an entry or exit point, it follows that (when q is even) $\varsigma_{u,v}(\tau) = 0$.

2) (Touch point). Assume now that $\tau \in \mathcal{T}_{to} \cap I^2_{u,v}$. If that case happens, since $v \in C_0(u)$, our hypotheses imply that τ is an essential touch point satisfying (32), and hence, that $q \geq 2$. Since $g(y_u)$ has a isolated local maximum at τ, $g(y_u)$ and $g^{(1)}(y_u)$ vanish at τ while $\frac{\mathrm{d}}{\mathrm{d}t} g^{(1)} = g^{(2)}(u, y_u)$ is nonpositive and continuous at τ since u is continuous by Prop. 7(i). We thus have:

$$g(y_u(t)) = \frac{\mathrm{d}}{\mathrm{d}t} g^{(1)}(y_u)|_{t=\tau} \frac{(t-\tau)^2}{2} + o((t-\tau)^2). \tag{39}$$

Since $\tau \in I^2_{u,v}$, we also have $g_y(y_u(\tau))z_{u,v}(\tau) = 0$. The function $g_y(y_u)z_{u,v}$ being C^1 (since $q \geq 2$) with almost everywhere a bounded second derivative, we get by (28), taking the nonnegative part:

$$(g_y(y_u(t))z_{u,v}(t))_+ = (g^{(1)}_y(y_u(\tau))z_{u,v}(\tau)(t-\tau))_+ + o(t-\tau). \tag{40}$$

From (39), (40) and (A5)(ii), $(g_y(y_u)z_{u,v})^2_+/g(y_u)$ is left-and right continuous when $t \to \tau$. Therefore, taking the $\liminf$ when $t \to \tau$ comes to take the min of both limits when $t \to \tau^+$ and $t \to \tau^-$, thus we obtain:

$$\varsigma_{u,v}(\tau) = \min\left\{ \frac{(g^{(1)}_y(y_u(\tau))z_{u,v}(\tau))^2}{g^{(2)}(u(\tau), y_u(\tau))} \;;\; 0 \right\} \;=\; \frac{(g^{(1)}_y(y_u(\tau))z_{u,v}(\tau))^2}{g^{(2)}(u(\tau), y_u(\tau))} > -\infty. \tag{41}$$

3) (Conclusion). For all $\tau \in \mathcal{T} \cap I^2_{u,v}$, we showed that $\varsigma_{u,v}(\tau) > -\infty$. Therefore we may apply (21). Set $I_0 := \operatorname{int} I(G(u))$. By (18), we have $\operatorname{supp}(\mathrm{d}\eta) \subset I^2_{u,v}$ and in view of remark 8 we may write that:

$$\sigma(\eta, T^{2,i}_K(G(u), DG(u)v)) = \int_{I_0} \varsigma_{u,v}(t)\eta_0(t)\mathrm{d}t + \sum_{\tau \in \mathcal{T} \cap I^2_{u,v}} \nu_\tau \varsigma_{u,v}(\tau) \tag{42}$$

where $\eta_0 \in L^1(I_0)$ and $\nu_\tau = [\eta(\tau)]$. By (20), $\varsigma_{u,v}$ vanishes on $I_0 \cap I^2_{u,v}$ and thus on $I_0 \cap \operatorname{supp}(\eta_0)$. Hence, $\int_{I_0} \varsigma_{u,v}(t)\eta_0(t)\mathrm{d}t = 0$. If $\tau \in \mathcal{T}_{en} \cup \mathcal{T}_{ex}$, we have, if q is *odd*, $\nu_\tau = 0$ by Prop. 7(ii)(a) and we showed that $\varsigma_{u,v}(\tau) > -\infty$. If q is *even*, we showed in point 1) that $\varsigma_{u,v}(\tau) = 0$ (and we have $\nu_\tau < +\infty$). In both cases, we deduce that $\nu_\tau \varsigma_{u,v}(\tau) = 0$.

It remains only in (42), when $q \geq 2$, the contribution of finitely many touch points τ in $\mathcal{T}_{to} \cap I^2_{u,v} = \mathcal{T}^{ess}_{to}$ with $\varsigma_{u,v}(\tau)$ given by (41). Hence (35) follows.

Proof (Proof of Theorem 12) Combining Theorem 4 and Propositions 10 and 14, we obtain that (34) holds, for all $v \in C_0(u)$. Since the left-hand-side of (34) is a continuous quadratic form, it remains nonnegative on the closure of $C_0(u)$. We end the proof by checking that the latter is equal to $C(u)$, the cone of critical directions.

Since $C(u)$ is closed and contains $C_0(u)$, we have of course $\overline{C_0(u)} \subset C(u)$. We prove the converse relation. Let $v_0 \in C(u)$. We remind that $v \in C(u)$ iff $g_y(y_u)z_{u,v} \leq 0$ on $I(g(y_u))$ and $g_y(y_u)z_{u,v} = 0$ on the support of the Lagrange multiplier η. Let $\rho : \mathbb{R} \to \mathbb{R}$ be a function of class C^∞ having support on $[-1, 1]$ which is positive on $(-1, 1)$. For $\varepsilon > 0$, set $\rho_\varepsilon(t) := \varepsilon^{q+1}\rho(t/\varepsilon)$, thus we have $\rho_\varepsilon \to 0$ in $W^{q,\infty}$. By Lemma 9(ii), for $\varepsilon > 0$ small enough, there exists a unique $v_\varepsilon \in L^\infty(0,T)$ such that $g(y_u)z_{u,v_\varepsilon} = g(y_u)z_{u,v_0} - \sum_{t\in\mathcal{T}_{to}\setminus\mathcal{T}_{to}^{ess}} \rho_\varepsilon(t-\tau) \in W^{q,\infty}(0,T)$. Then we have $g_y(y_u)z_{u,v_\varepsilon} = g_y(y_u)z_{u,v_0}$ outside $(\tau-\varepsilon, \tau+\varepsilon)$, for all non essential touch point τ, $g_y(y_u(\tau))z_{u,v_\varepsilon}(\tau) < 0$ for such τ, and hence, the touch points being isolated, for $\varepsilon > 0$ small enough, $v_\varepsilon \in C_0(u)$. Since $DG(u)v_\varepsilon \to DG(u)v_0$ in $\mathcal{W}$, where $\mathcal{W}$ was defined in (30), and $DG(u)$ has a bounded inverse by Lemma 9(ii), we have $v_\varepsilon \to v_0$ in $L^\infty(0,T)$ when $\varepsilon \downarrow 0$. The conclusion follows.

3.2 Extended Second-order Necessary Conditions

The solution $z_{u,v}$ of the linearized state equation (8) when $v \in L^2(0,T)$, is well-defined and belongs to $H^1(0,T) \subset C[0,T]$. Thus we may extend continuously $DJ(u)$ and $DG(u)$ over $L^2(0,T)$ (we keep the same notations for the extensions). Since $DG(u) : L^2(0,T) \to C[0,T]$, it makes sense to extend the critical cone $C(u)$ defined in (16) to critical directions in L^2, as follows:

$$C_{L^2}(u) = \{v \in L^2(0,T) \setminus DG(u)v \in T_K(G(u)) \ ; \ DJ(u)v \leq 0\}. \tag{43}$$

Note that when (u, η) satisfies (13), relation (18) remains true with $C_{L^2}(u)$ and $L^2(0,T)$ instead of $C(u)$ and $\mathcal{U}$, respectively.

The necessary and sufficient second-order conditions involve respectively $C(u)$ and $C_{L^2}(u)$ (see sections 4 and 5). Therefore, to obtain the no-gap second-order conditions, we need the following variant of Theorem 12.

Corollary 15 *The statements of Theorem 12 and Corollary 13 still hold replacing assumption (A6') and $C(u)$ respectively by (A6) and $C_{L^2}(u)$.*

Corollary 15 is obtained as a consequence of Th. 12, the continuity of the left-hand side of (34) w.r.t. $v \in L^2$, and the density of $C(u)$ in $C_{L^2}(u)$ (Lemma 17). To prove the latter, we first need a general result.

Lemma 16 *Let $q \geq 1$ and $a < b \in \mathbb{R}$. Then for all $\hat{x} \in H^q(a,b) = W^{q,2}(a,b)$, there exists a sequence (x_n) of $W^{q,\infty}(a,b)$ such that $x_n^{(j)}(a) = \hat{x}^{(j)}(a)$, $x_n^{(j)}(b) = \hat{x}^{(j)}(b)$ for all $j = 0, \ldots, q-1$, $n \in \mathbb{N}$ and $\|x_n - \hat{x}\|_{q,2} \to 0$.*

Proof Set $\hat{x}_a := (\hat{x}(a), \ldots, \hat{x}^{(q-1)}(a))^*$, $\hat{x}_b := (\hat{x}(b), \ldots, \hat{x}^{(q-1)}(b))^* \in \mathbb{R}^q$ and $\hat{u} := \hat{x}^{(q)} \in L^2(a,b)$. For $u \in L^2(a,b)$, let $x_u \in H^q(a,b)$ be the solution of:

$$x_u^{(q)}(t) = u(t) \quad \text{a.e. on } [a,b] \ ; \ (x_u(a), \ldots, x_u^{(q-1)}(a)) = \hat{x}_a^*. \tag{44}$$

For $n \in \mathbb{N}$, consider the following problem:

$$(\mathcal{P}_n) \qquad \min \tfrac{1}{2}\|u-\hat{u}\|_2^2 \quad ; \quad \mathcal{A}u = \hat{x}_b \quad ; \quad u \in \mathcal{U}_n, \tag{45}$$

where $\mathcal{U}_n := \{u \in L^2(0,T) \; ; \; |u(t)| \leq n \text{ a.e.}\}$ and $\mathcal{A} : L^2 \to \mathbb{R}^q \; ; \; u \mapsto (x_u(b), \dots, x_u^{(q-1)}(b))^*$. By construction, $\mathcal{A}\hat{u} = \hat{x}_b$. It is readily seen that the mapping $L^2(a,b) \to H^q(a,b)$; $u \mapsto x_u$ solution of (44) is continuous. Since $H^q(a,b)$ has a continuous inclusion into $C^{q-1}[a,b]$, it follows that the linear mapping $\mathcal{A}$ is also continuous.

Let us first show that for n large enough, the problems $(\mathcal{P}_n)$ are feasible and uniformly qualified, that is there exist $n_0 \in \mathbb{N}$ and $\delta_0 > 0$ such that

$$\hat{x}_b + \delta_0 B_{\mathbb{R}^q} \subset \mathcal{A}\mathcal{U}_{n_0} \subset \mathcal{A}\mathcal{U}_n \qquad \forall n \geq n_0, \tag{46}$$

with $B_{\mathbb{R}^q}$ the unit ball in $\mathbb{R}^q$. Indeed, consider e.g. for $\delta \in \mathbb{R}^q$ the (unique) polynomial function x_δ of degree $2q-1$ that takes with its $q-1$ first derivatives the values $\hat{x}_a$ and $\hat{x}_b$ at a and b. It is easily seen that its coefficients are solution of a full-rank linear system with $\hat{x}_b - \hat{x}_a + \delta$ as right-hand side, hence, taking the sup over $(t,\delta) \in [a,b] \times B_{\mathbb{R}^q}(0,\delta_0)$ of the functions $u_\delta(t) = x_\delta^{(q)}(t)$ that are C^∞ w.r.t. t and δ provides an uniform bound n_0 such that (46) holds.

Since Robinson's constraint qualification holds for n large enough, there exists a (unique) optimal solution u_n of $(\mathcal{P}_n)$ and a normal Lagrange multiplier $\lambda_n \in \mathbb{R}^{q*}$, such that (throughout the proof, $\langle \cdot,\cdot \rangle$ denotes the scalar product over L^2):

$$0 \leq \langle u_n - \hat{u} + \mathcal{A}^* \lambda_n, v - u_n \rangle \qquad \forall v \in \mathcal{U}_n. \tag{47}$$

Since the feasible set of problem $(\mathcal{P}_n)$ is increasing for inclusion when $n \to +\infty$, the cost function is decreasing, thus $\|u_n - \hat{u}\|_2$ is bounded. Hence the sequence (u_n) converges weakly to some $\bar{u} \in L^2$. We may rewrite (47) as:

$$\|u_n - \hat{u}\|_2^2 + \lambda_n(\hat{x}_b - \mathcal{A}v) \leq \langle u_n - \hat{u}, v - \hat{u} \rangle \qquad \forall v \in \mathcal{U}_n. \tag{48}$$

Qualification property (46) implies that $\delta_0|\lambda_n| \leq \sup_{v \in \mathcal{U}_{n_0}} \lambda_n(x_b - \mathcal{A}v)$, hence, taking the sup for $v \in \mathcal{U}_{n_0}$ successively in the right and left hand side of (48), we deduce that for some constant $K(n_0) > 0$ that depends on n_0, we have $\delta_0|\lambda_n| \leq K(n_0)$, for all $n \geq n_0$. Therefore the sequence (λ_n) is uniformly bounded. Define now $v_n \in \mathcal{U}_n$ as $v_n(t) = \max\{-n; \min\{n, \hat{u}(t)\}\}$ a.e. By the Lebesgue dominated convergence Theorem, $v_n \to \hat{u}$ in L^2 and by (48):

$$\|u_n - \hat{u}\|_2^2 \;\leq\; \langle u_n - \hat{u}, v_n - \hat{u} \rangle \;+\; \lambda_n(\mathcal{A}v_n - \hat{x}_b) \;\longrightarrow\; 0,$$

since $u_n - \hat{u} \rightharpoonup \bar{u} - \hat{u}$ weakly in L^2, $v_n - \hat{u} \to 0$ strongly in L^2, λ_n is bounded and $\mathcal{A}v_n \to \mathcal{A}\hat{u} = \hat{x}_b$ by continuity of $\mathcal{A}$. It follows that $\|u_n - \hat{u}\|_2 \to 0$ and the sequence $x_n := x_{u_n}$ satisfies all the required properties, so the proof is completed.

Lemma 17 *Let $u \in \mathcal{U}$ and $\eta \in \Lambda(u)$ such that (A3), (A4) and (A6) are satisfied. Then $C(u)$ is a dense subset of $C_{L^2}(u)$.*

Proof Since (A4) holds, denote by $0 < \tau_1 < \ldots < \tau_N < T$ the junction times of the trajectory (u, y_u), and set $\tau_0 := 0$, $\tau_{N+1} := T$. Let $v \in C_{L^2}(u)$ and set $x := DG(u)v$. By Lemma 16 applied on intervals $[\tau_k, \tau_{k+1}]$ that are not boundary arcs, there exists a sequence $x_n \in W^{q,\infty}(0,T)$ such that $x_n = 0 = x$ by (A6) on boundary arcs, $x_n^{(j)}(\tau_k) = x^{(j)}(\tau_k)$ for all $j = 0, \ldots, q-1$ and $k = 0, \ldots, N+1$, and $x_n \to x$ in H^q. By (A3), we may define $v_n \in L^\infty(0,T)$ such that $DG(u)v_n = x_n$ for all n. It is readily seen that $v_n \in C(u)$ for all n and $v_n \to v$ in L^2, which achieves the proof.

4 Second-order Sufficient Conditions

The second-order sufficient conditions theory classically involves two norms, namely L^2 and L^∞, see Ioffe [17, Part III] and Maurer [28].

Assume that X, Z are Banach spaces endowed with the norms $\|\cdot\|_X$ and $\|\cdot\|_Z$, respectively, such that $Z \subset X$ with continuous embedding. Let $k \in \mathbb{N}$. We say that $r(x) = \mathcal{O}_Z(\|x\|_X^k)$ if $|r(x)| \leq C\|x\|_X^k$ for some $C > 0$ when $\|x\|_Z$ is small enough. We say that $r(x) = o_Z(\|x\|_X^k)$ if $|r(v)|/\|x\|_X^k$ goes to zero when $\|x\|_Z$ goes to zero. In the sequel, $\|\cdot\|_p$ (resp. $\|\cdot\|_{r,p}$) denotes the norm of the space $L^p(0,T)$ (resp. the Sobolev space $W^{r,p}(0,T)$), for $1 \leq p \leq \infty$ and $r = 1, \ldots < +\infty$. We write $\mathcal{O}_p$ and $\mathcal{O}_{r,p}$ for respectively $\mathcal{O}_{\|\cdot\|_{L^p}}$ and $\mathcal{O}_{\|\cdot\|_{W^{r,p}}}$, and we use the same convention for o_p and $o_{r,p}$. Similarly, B_p and $B_{r,p}$ denote open balls in L^p and $W^{r,p}$, respectively.

We remind that a quadratic form $Q(v)$ on a Hilbert space is a *Legendre form* (Ioffe and Tihomirov [18]), if it is weakly lower semi-continuous (w.l.s.c.) and if $v_n \rightharpoonup v$ weakly and $Q(v_n) \to Q(v)$ imply that $v_n \to v$ strongly.

The next theorem gives the second-order sufficient condition in its well-known form (i.e. without the curvature term).

Theorem 18 *Let $u \in \mathcal{U}$ satisfy (13) with Lagrange multiplier η and assume that (A2') holds. If the following second-order sufficient condition is satisfied:*

$$D^2_{uu}L(u,\eta)(v,v) > 0 \qquad \forall\, v \in C_{L^2}(u) \setminus \{0\} \tag{49}$$

then u is a local solution of (4) satisfying the quadratic growth condition (5).

Conversely, if (A1)-(A6) hold and if (u, y_u) has no essential touch point (in particular, if the state constraint is of first order *$q = 1$), then the second-order sufficient condition (49) is satisfied* iff *the quadratic growth condition (5) is satisfied.*

The proof of Theorem 18 will be given after a sequence of short lemmas.

Lemma 19 *Let $(u, \eta) \in \mathcal{U} \times \mathcal{M}_+[0,T]$ and $v \in \mathcal{U}$. The following holds, for all $\sigma \in [0,1]$:*

$$\|y_{u+\sigma v} - y_u\|_\infty = \mathcal{O}_\infty(\|v\|_1) \tag{50}$$

$$\|p_{u+\sigma v,\eta} - p_{u,\eta}\|_\infty = \mathcal{O}_\infty(\|v\|_1) \tag{51}$$

$$\|z_{u+\sigma v,v}\|_\infty = \mathcal{O}_\infty(\|v\|_1) \tag{52}$$

$$\|z_{u+\sigma v,v} - z_{u,v}\|_\infty = \mathcal{O}_\infty(\|v\|_2^2). \tag{53}$$

Proof Set $u_\sigma := u + \sigma v$, and let C denote a positive constant. Since f is Lipschitz continuous by (A0), (50) is an easy consequence of Lemma 32. Thus, u and v being essentially bounded, u_σ and y_{u_σ} take values a.e. in a compact set of type

$$V_\delta = \{(\hat{u}, \hat{y}) \in \mathbb{R} \times \mathbb{R}^n \ ; \ |\hat{u}| + |\hat{y}| \leq \delta\}, \tag{54}$$

for some $\delta > 0$. The mappings f, ℓ and g as well as their first order derivatives are C^1, and hence Lipschitz continuous over the compact set V_δ. Lemma 32, applied to the costate equation (7), ensures that $p_{u_\sigma,\eta}$ also remains uniformly bounded. The derivation of (51) and (52) being similar to the one of (53), we detail only the latter. We have (omitting time argument):

$$\begin{aligned}|\dot{z}_{u_\sigma,v}(t) - \dot{z}_{u,v}(t)| \leq \|f_y\|_\infty |z_{u_\sigma,v} - z_{u,v}| \\ + \left(|Df(u_\sigma, y_{u_\sigma}) - Df(u, y_u)|\right)\left(|z_{u,v}| + |v(t)|\right).\end{aligned}$$

Since Df is Lipschitz on V, we have by (50) $|Df(u_\sigma, y_{u_\sigma}) - Df(u, y_u)| \leq C(\|v\|_1 + |v|)$. Combining with (52) and the inequality $ab \leq \frac{1}{2}(a^2 + b^2)$, we deduce from the above display that

$$|\dot{z}_{u_\sigma,v}(t) - \dot{z}_{u,v}(t)| \leq \|f_y\|_\infty |z_{u_\sigma,v} - z_{u,v}| + C\left(\|v\|_1^2 + |v(t)|^2\right).$$

We conclude with Lemma 32 and the inequality $\|v\|_1 \leq \sqrt{T}\|v\|_2$.

Lemma 20 *Let $(u, \eta) \in \mathcal{U} \times \mathcal{M}_+[0, T]$ and $v \in \mathcal{U}$. Then:*

$$L(u + v, \eta) = L(u, \eta) + D_u L(u, \eta)v + \frac{1}{2} D^2_{uu} L(u, \eta)(v, v) + r(v) \tag{55}$$

with $r(v) = \mathcal{O}_\infty(\|v\|_3^3)$. In particular, $r(v) = o_\infty(\|v\|_2^2)$.

Proof For $\sigma \in [0, 1]$, set again $u_\sigma := u + \sigma v$ and $p_{u_\sigma} := p_{u_\sigma,\eta}$. By Lemma 1:

$$\begin{aligned} r(v) &= \left[\int_0^1 (1 - \sigma)\left(D^2_{uu} L(u + \sigma v, \eta) - D^2_{uu} L(u, \eta)\right) \mathrm{d}\sigma\right](v, v) \qquad (56) \\ &= \int_0^1 \int_0^T \varDelta_1(t)\mathrm{d}t\mathrm{d}\sigma + \int_0^1 \int_0^T \varDelta_2(t)\mathrm{d}\eta(t)\mathrm{d}\sigma + \int_0^1 \varDelta_3 \mathrm{d}\sigma,\end{aligned}$$

with (omitting time argument)

$$\begin{aligned}\varDelta_1(t) &= D^2 H_{(u,y)^2}(u_\sigma, y_{u_\sigma}, p_{u_\sigma})(v, z_{u_\sigma,v})^2 - \ D^2 H_{(u,y)^2}(u, y_u, p_u)(v, z_{u,v})^2 \\ \varDelta_2(t) &= z^*_{u_\sigma,v} g_{yy}(y_{u_\sigma}) z_{u_\sigma,v} - z^*_{u,v} g_{yy}(y_u) z_{u,v} \\ \varDelta_3 &= z_{u_\sigma,v}(T)^* \phi_{yy}(y_{u_\sigma}(T)) z_{u_\sigma,v}(T) - z_{u,v}(T)^* \phi_{yy}(y_u(T)) z_{u,v}(T).\end{aligned}$$

Under assumption (A0), second-order derivatives g_{yy}, etc. are Lipschitz continuous over a compact set V_δ defined in (54) for some $\delta > 0$. By Lemma 19 we get, for some constant $C > 0$:

$$\begin{aligned}\varDelta_2(t) &\leq C\left(|y_{u_\sigma} - y_u||z_{u_\sigma,v}|^2 + (|z_{u_\sigma,v}| + |z_{u,v}|)|z_{u_\sigma,v} - z_{u,v}|\right) \\ &\leq \mathcal{O}_\infty(\|v\|_1^3 + \|v\|_1 \|v\|_2^2) \leq \mathcal{O}_\infty(\|v\|_3^3),\end{aligned}$$

since by the Cauchy-Schwarz and Hölder inequalities, that give respectively $\|\cdot\|_2^2 \leq \|\cdot\|_3^{3/2}\|\cdot\|_1^{1/2}$ and $\|\cdot\|_1 \leq T^{2/3}\|\cdot\|_3$, we have $\|\cdot\|_2^2\|\cdot\|_1 \leq T\|\cdot\|_3^3$. Since the measure $\mathrm{d}\eta$ is bounded and the $\mathcal{O}_\infty$ are uniform w.r.t. time, we obtain $\int_0^T \Delta_2(t)\mathrm{d}\eta(t) = \mathcal{O}_\infty(\|v\|_3^3)$. The same upper bound holds for $\Delta_3(T)$. As for $\Delta_1(t)$, we have in the same way, by Lemma 19:

$$\begin{aligned}\Delta_1(t) &\leq C(|y_{u_\sigma} - y_u| + |p_{u_\sigma} - p_u| + \sigma|v|)(|z_{u_\sigma,v}|^2 + |v|^2) \\ &\quad + C(|z_{u_\sigma,v}| + |z_{u,v}| + |v|)|z_{u_\sigma,v} - z_{u,v}| \\ &\leq C(\|v\|_1^3 + \|v\|_1^2|v(t)| + \|v\|_1|v(t)|^2 + |v(t)|^3).\end{aligned}$$

Hence, $\int_0^T \Delta_1(t)\mathrm{d}t = \mathcal{O}_\infty(\|v\|_3^3)$. Finally, since the $\mathcal{O}_\infty$ do not depend on $\sigma \in [0,1]$, we obtain after integration over $[0,1]$ that $r(v) = \mathcal{O}_\infty(\|v\|_3^3)$. Since $\|\cdot\|_3^3 \leq \|\cdot\|_2^2\|\cdot\|_\infty$, it follows that $r(v) = o_\infty(\|v\|_2^2)$.

Lemma 21 *Let $(u,\eta) \in \mathcal{U} \times \mathcal{M}_+[0,T]$ satisfy (A2'). Then the quadratic form $\mathcal{U} \to \mathbb{R}$, $v \mapsto D^2_{uu}L(u,\eta)(v,v)$ has a unique extension to a continuous quadratic form over $L^2(0,T)$, and the latter is a Legendre form.*

Proof Since L^∞ is a dense subset of L^2 and $v \mapsto D^2_{uu}L(u,\eta)(v,v)$ is continuous for the norm of L^2, it has a unique continuous extension Q over L^2. Set $p := p_{u,\eta}$. By (10), we can write $Q(v) = Q_0(v) + Q_1(v) + Q_2(v)$ with:

$$\begin{aligned}Q_2(v) &= \textstyle\int_0^T H_{yy}(u,y_u,p)(z_{u,v},z_{u,v})\mathrm{d}t \\ &\qquad + z_{u,v}(T)^*\phi_{yy}(y_u(T))z_{u,v}(T) + \textstyle\int_0^T z_{u,v}^* g_{yy}(y_u)z_{u,v}\mathrm{d}\eta \\ Q_1(v) &= 2\textstyle\int_0^T H_{yu}(u,y_u,p)(z_{u,v},v)\mathrm{d}t \\ Q_0(v) &= \textstyle\int_0^T H_{uu}(u,y_u,p)(v,v)\mathrm{d}t\end{aligned}$$

Let $v_n \rightharpoonup \bar{v} \in L^2(0,T)$. The mapping $L^2(0,T) \to H^1(0,T)$; $v \mapsto z_{u,v}$ being linear continuous, $z_n := z_{u,v_n}$ converges weakly to $\bar{z} := z_{u,\bar{v}}$. Since (z_n) is bounded in $H^1(0,T)$ and the inclusion of the latter in $C[0,T]$ is compact, (z_n) is strongly convergent to $\bar{z}$, and thus $Q_2(v_n)$ converges strongly to $Q_2(\bar{v})$. The term $Q_1(v_n)$, bilinear in (z_n, v_n), also converges strongly to $Q_1(\bar{v})$ when z_n converges strongly and v_n weakly. Therefore, Q is a Legendre form iff Q_0 is.

Since $H_{uu}(u(t), y_u(t), p(t))$ is essentially bounded and, by (27), is uniformly invertible for almost all $t \in [0,T]$, $v \mapsto \sqrt{Q_0(v)}$ is a norm equivalent to the one of $L^2(0,T)$. Hence by [9, Prop. 3.76(i)], Q_0 is a Legendre form, and therefore so is Q.

Proof (Proof of Theorem 18) Assume that (49) holds but that the quadratic growth condition (5) is not satisfied. Then there exist a sequence $u_n \to u$ in L^∞, $u_n \neq u$, such that $G(u_n) \in K$ for all n and

$$J(u_n) \leq J(u) + o(\|u_n - u\|_2^2). \tag{57}$$

Since $G(u_n) \in K$ and $\eta \in N_K(G(u))$, we have:

$$J(u_n) - J(u) = L(u_n,\eta) - L(u,\eta) - \langle \eta, G(u_n) - G(u)\rangle \geq L(u_n,\eta) - L(u,\eta).$$

Since $u_n - u \to 0$ in L^∞, Lemma 20 yields $r(u_n - u) = o(\|u_n - u\|_2^2)$. As $D_u L(u,\eta) = 0$, we have:

$$o(\|u_n - u\|_2^2) \geq J(u_n) - J(u) \geq \frac{1}{2} D^2_{uu} L(u,\eta)(u_n - u, u_n - u) + o(\|u_n - u\|_2^2).$$

Let (v_n, ϵ_n) be such that $u_n - u = \epsilon_n v_n$ with $\|v_n\|_2 = 1$ and $\epsilon_n = \|u_n - u\|_2 \to 0$. Dividing by $\epsilon_n^2 > 0$ the above inequality, we get:

$$D^2_{uu} L(u,\eta)(v_n, v_n) + o(1) \leq o(1). \tag{58}$$

The sequence (v_n) being bounded in $L^2(0,T)$, taking if necessary a subsequence, we may assume that (v_n) converges weakly to some $\bar{v} \in L^2(0,T)$. Since $D^2_{uu} L(u,\eta)$ is weakly l.s.c., we get passing to the limit:

$$D^2_{uu} L(u,\eta)(\bar{v}, \bar{v}) \leq 0. \tag{59}$$

From (57), we derive that $J(u + \epsilon_n v_n) - J(u) = \epsilon_n DJ(u) v_n + r_n \leq o(\epsilon_n^2)$, where $r_n = \mathcal{O}(\epsilon_n^2)$ (by the same arguments as in the proof of Lemma 20). Thus $DJ(u)v_n + \mathcal{O}(\epsilon_n) \leq o(\epsilon_n)$, and passing to the limit, since the mapping $v \mapsto DJ(u)v = \int_0^T (\ell_y(u, y_u) z_{u,v} + \ell_u(u, y_u) v) \mathrm{d}t + \phi_y(y_u(T)) z_{u,v}(T)$ is weakly continuous, we obtain:

$$DJ(u)\bar{v} \leq 0. \tag{60}$$

Since $u + \epsilon_n v_n \in G^{-1}(K)$ (the inverse image of the set K), $\bar{v}$ is a weak limit of $v_n = (u_n - u)/\epsilon_n \in DG(u)^{-1} T_K(G(u))$. Since K is closed and convex, so is $T_K(G(u))$. The mapping $DG(u)$ being continuous, $DG(u)^{-1} T_K(G(u))$ is also a closed convex set, and hence weakly closed. It follows that $\bar{v} \in DG(u)^{-1} T_K(G(u))$. Therefore, with (60), $\bar{v} \in C_{L^2}(u)$. Thus (49) and (59) imply that $\bar{v} = 0$. On the other hand, (58) gives (with $Q := D^2_{uu} L(u,\eta)$):

$$0 = Q(\bar{v}) \leq \underline{\lim}\, Q(v_n) \leq \overline{\lim}\, Q(v_n) \leq 0$$

therefore $Q(v_n) \to Q(v)$. But Q is a Legendre form by Lemma 21 and $v_n \rightharpoonup \bar{v}$, which implies that $v_n \to \bar{v}$ in $L^2(0,T)$, hence $\|v_n\|_2 \to \|\bar{v}\|_2$. The expected contradiction arises since $\|v_n\|_2 = 1$ for all n whereas $\|\bar{v}\|_2 = 0$.

The converse, that holds under stronger assumptions, is a consequence of Corollaries 13-15. For convenience, we prove it later with Theorem 27.

5 Reduction Approach

There is still a gap between statements of Corollary 15 of Theorem 12 and Theorem 18, whenever essential touch points occur. We show in this section how to deal with this case, using a reduction approach in order to reformulate the constraint.

The idea of reduction methods (see e.g. [16] and [9, section 3.4.4]) is, when the constraint has finitely many contact points, to replace it by finitely many inequality constraints. The Hessian of Lagrangian of the corresponding reduced problem has an additional term that matches the curvature term. We obtain thus a no-gap second-order condition.

5.1 General results on reduction

It is known that the Sobolev spaces $W^{1,\infty}(0,T)$ and $W^{2,\infty}(0,T)$, endowed with the norms $\|x\|_{1,\infty} = \|x\|_\infty + \|\dot{x}\|_\infty$ and $\|x\|_{2,\infty} = \|x\|_{1,\infty} + \|\ddot{x}\|_\infty$, coincide with the spaces of Lipschitz continuous functions and the one of functions having a Lipschitz continuous derivative, respectively. For all $t, t_0 \in [0,T]$, $h \in W^{1,\infty}(0,T)$ and $x \in W^{2,\infty}(0,T)$, we have:

$$|h(t) - h(t_0)| \leq |t - t_0| \|\dot{h}\|_\infty, \tag{61}$$

$$|x(t) - x(t_0) - \dot{x}(t_0)(t - t_0)| \leq \tfrac{1}{2}|t - t_0|^2 \|\ddot{x}\|_\infty. \tag{62}$$

We now give some general results about zeros of functions of $W^{1,\infty}(0,T)$, and local minima/maxima of functions of $W^{2,\infty}(0,T)$.

Lemma 22 *Let $h_0 \in W^{1,\infty}(0,T)$ and $\tau_0 \in (0,T)$ satisfy the three following conditions: $h_0(\tau_0) = 0$; $\dot{h}_0$ is continuous at τ_0 ; $\dot{h}_0(\tau_0) \neq 0$. Then for some $\delta,\ \varepsilon > 0$, the mapping:*

$$\varXi \ : \ B_{1,\infty}(h_0, \delta) \mapsto (\tau_0 - \varepsilon, \tau_0 + \varepsilon) \quad ; \quad h \mapsto \tau_h \ \textit{ such that } \ h(\tau_h) = 0, \tag{63}$$

is well-defined and Lipschitz continuous on $B_{1,\infty}(h_0, \delta)$, and Fréchet differentiable at h_0, with derivative given by:

$$D\varXi(h_0)d = -d(\tau_0)/\dot{h}_0(\tau_0), \qquad \textit{for all } d \in W^{1,\infty}. \tag{64}$$

More precisely, we have for all $h, h_i \in B_{1,\infty}(h_0, \delta)$, $i = 1, 2$ and $\tau_i = \tau_{h_i}$:

$$\tau_2 - \tau_1 = \mathcal{O}_{1,\infty}(\|h_2 - h_1\|_\infty), \tag{65}$$

$$\dot{h}_0(\tau_0)(\tau_h - \tau_0) + h(\tau_0) = o_{1,\infty}\left(\|h - h_0\|_\infty\right). \tag{66}$$

Proof Assume w.l.o.g that $\beta := \dot{h}_0(\tau_0) > 0$, and denote by $c(\cdot)$ the modulus of continuity of $\dot{h}_0$ at τ_0. Fix $\varepsilon > 0$ such that $c(\varepsilon) < \frac{1}{4}\beta$. Thus, $\dot{h}_0 \geq \frac{3}{4}\beta$ on $(\tau_0 - \varepsilon, \tau_0 + \varepsilon)$ and it follows that $h_0(\tau_0 - \varepsilon) < -\frac{3}{4}\beta\varepsilon$ and $h_0(\tau_0 + \varepsilon) > \frac{3}{4}\beta\varepsilon$. Set $\delta := \min\{\frac{1}{4}\beta\varepsilon; \frac{1}{4}\beta\}$ and let $h \in B_{1,\infty}(h_0, \delta)$. Thus, $h(\tau_0 - \varepsilon) < 0 < h(\tau_0 + \varepsilon)$ and h is continuous, so h has at least one zero τ_h in $(\tau_0 - \varepsilon, \tau_0 + \varepsilon)$. Let $(h_1, h_2) \in B_{1,\infty}(h_0, \delta)$ and τ_i such that $h_i(\tau_i) = 0$, $i = 1, 2$. By the definition of δ, we have $\dot{h}_1 \geq \frac{1}{2}\beta$ a.e. on $(\tau_0 - \varepsilon, \tau_0 + \varepsilon)$, and, in consequence,

$$\frac{\beta}{2}|\tau_2 - \tau_1| \leq |h_1(\tau_2)| = |h_1(\tau_2) - h_2(\tau_2)| \leq \|h_2 - h_1\|_\infty. \tag{67}$$

Hence $|\tau_2 - \tau_1| \leq \frac{2}{\beta}\|h_2 - h_1\|_\infty$, which shows the uniqueness of the zero (take $h_1 = h_2$), Lipschitz continuity and (65).

By continuity of $\varXi$ and h_0, and (61) applied to $h - h_0$, we have:

$$h_0(\tau_h) - \dot{h}_0(\tau_0)(\tau_h - \tau_0) = o(|\tau_h - \tau_0|)$$

$$(h - h_0)(\tau_h) - (h - h_0)(\tau_0) \ = \ h_0(\tau_h) - h(\tau_0) = \mathcal{O}(\|\dot{h} - \dot{h}_0\|_\infty |\tau_h - \tau_0|).$$

Since $\tau_h - \tau_0 = \mathcal{O}_{1,\infty}(\|h - h_0\|_\infty)$ by (67), summing the above expansions yields (66), from which (64) follows.

Lemma 23 *Let $x_0 \in W^{2,\infty}(0,T)$ and $\tau_0 \in (0,T)$ be such that $\dot{x}_0(\tau_0) = 0$, $\ddot{x}_0$ is continuous at τ_0 and $\ddot{x}_0(\tau_0) < 0$. Thus x_0 has a local maximum at τ_0, and for $\varepsilon > 0$ and $\delta > 0$ small enough, $x \in B_{2,\infty}(x_0,\delta)$ attains its maximum over $(\tau_0 - \varepsilon, \tau_0 + \varepsilon)$ at a unique point τ_x. The mapping $\Theta : B_{2,\infty}(x_0,\delta) \to (\tau_0 - \varepsilon, \tau_0 + \varepsilon)$; $x \mapsto \tau_x$ is Lipschitz continuous over $B_{2,\infty}(x_0,\delta)$, Fréchet differentiable at x_0, with derivative given by:*

$$D\Theta(x_0)w = -\dot{w}(\tau_0)/\ddot{x}_0(\tau_0) \qquad \forall\, w \in W^{2,\infty}. \tag{68}$$

Furthermore, the mapping

$$\Phi : B_{2,\infty}(x_0,\delta) \to \mathbb{R} \ ; \ x \mapsto x(\tau_x), \tag{69}$$

that associates with x the value of its maximum on $(\tau_0 - \varepsilon, \tau_0 + \varepsilon)$, is C^1 over $B_{2,\infty}(x_0,\delta)$ and twice Fréchet differentiable at x_0 with first and second derivatives given by, for all $x \in B_{2,\infty}(x_0,\delta)$ and $d \in W^{2,\infty}$:

$$D\Phi(x)d = d(\tau_x) \quad ; \quad D^2\Phi(x_0)(d,d) = -\frac{\dot{d}(\tau_0)^2}{\ddot{x}_0(\tau_0)}. \tag{70}$$

More precisely, for all $x, x_i \in B_{2,\infty}(x_0,\delta)$, $i = 1,2$ and $\tau_i = \tau_{x_i}$, we have:

$$x_2(\tau_2) = x_2(\tau_1) + \mathcal{O}_{2,\infty}(\|x_2 - x_1\|^2_{1,\infty}), \tag{71}$$

$$x(\tau_x) = x(\tau_0) - \frac{\dot{x}(\tau_0)^2}{2\ddot{x}_0(\tau_0)} + o_{2,\infty}(\|x - x_0\|^2_{1,\infty}). \tag{72}$$

Proof Define δ as in the proof of Lemma 22, with h_0 replaced by $-\dot{x}_0$. It follows that for all $x \in B_{2,\infty}(x_0,\delta)$, there exists a unique τ_x satisfying $\dot{x}(\tau_x) = 0$, and we have $\ddot{x}(t) \le \ddot{x}_0(\tau_0)/2 < 0$ a.e. on $(\tau_0 - \varepsilon, \tau_0 + \varepsilon)$. Hence $\dot{x}$ is decreasing on $(\tau_0 - \varepsilon, \tau_0 + \varepsilon)$, and x has unique maximum over $[\tau_0 - \varepsilon, \tau_0 + \varepsilon]$ attained at time τ_x. By composition of the mapping Ξ of Lemma 22 by the mapping $x \mapsto h = \dot{x} \in W^{1,\infty}$, Θ is well-defined, continuous over $B_{2,\infty}(x_0,\delta)$ and Fréchet differentiable at x_0, and (68) follows from (64).

By (62) applied to $x_2 - x_1$, as $\dot{x}_1(\tau_1) = 0$ and $\tau_2 - \tau_1 = \mathcal{O}_{2,\infty}(\|x_2 - x_1\|_{1,\infty})$ by (65), we get:

$$\begin{aligned} x_2(\tau_2) &= x_2(\tau_1) + (\dot{x}_2(\tau_1) - \dot{x}_1(\tau_1))(\tau_2 - \tau_1) + \mathcal{O}(|\tau_2 - \tau_1|^2) \\ &= x_2(\tau_1) + \mathcal{O}_{2,\infty}(\|x_2 - x_1\|^2_{1,\infty}) \end{aligned}$$

which shows (71) and proves that Φ is C^1 with first order derivative given by (70). By continuity of $\ddot{x}_0$ and (62) applied to $x - x_0$, we have, as $\dot{x}_0(\tau_0) = 0$:

$$\begin{aligned} x_0(\tau_x) &= x_0(\tau_0) + \ddot{x}_0(\tau_0)\tfrac{(\tau_x - \tau_0)^2}{2} + o(|\tau_x - \tau_0|^2), \\ (x - x_0)(\tau_x) &= (x - x_0)(\tau_0) + \dot{x}(\tau_0)(\tau_x - \tau_0) + \mathcal{O}(\|\ddot{x} - \ddot{x}_0\|_\infty |\tau_x - \tau_0|^2). \end{aligned}$$

Summing the above expansions, and since by (66),

$$\tau_x - \tau_0 = -\frac{\dot{x}(\tau_0)}{\ddot{x}_0(\tau_0)} + o_{2,\infty}\left(\|x - x_0\|_{1,\infty}\right),$$

we obtain (72). Hence Φ is twice Fréchet differentiable at x_0 with second-order derivative given by (70).

5.2 Application to optimal control problems.

If the state constraint is of first order $q = 1$, then Theorem 18 gives a no-gap second-order condition, that characterizes the quadratic growth. We show in this section how to extend this no-gap condition to the case when the trajectory has essential touch points (see Theorem 27).

Therefore, we assume in this section that the state constraint *is not of first order*, that is, the function $g^{(1)}(u, y) = g_y(y)f(u, y)$ does not depend on u (which means $g_u^{(1)}(u, y) \equiv 0$). Note that this implies that $G(u) = g(y_u) \in W^{2,\infty}$, for all $u \in \mathcal{U}$.

Definition 24 *Assume that $g_u^{(1)} \equiv 0$ (the state constraint is not of order one). Let $u \in G^{-1}(K)$. We say that a touch point τ of the trajectory (u, y_u) is* reducible*, if the following conditions are satified:* (i) *the function $t \mapsto g^{(2)}(u(t), y_u(t))$ is continuous at τ;* (ii) *non-tangentiality condition (32) is satisfied at τ.*

Remark 25 1) Point (i) in the above definition is always satisfied if the state constraint is of order $q > 2$, since in that case $g^{(2)}(u, y_u) = g^{(2)}(y_u)$.
2) If $q = 2$ and $\eta \in \Lambda(u) \neq \emptyset$, sufficient conditions for point (i) are assumptions (A2)-(A4), since by Prop. 7(i) they imply the continuity of u.

Let $u \in G^{-1}(\mathcal{U})$, and let $\mathcal{T}_{red}$ be a *finite* subset of reducible touch points of the trajectory (u, y_u). By definition of touch points, there exists $\varepsilon > 0$ such that $(\tau - 2\varepsilon, \tau + 2\varepsilon) \subset (0, T)$ and $(\tau - 2\varepsilon, \tau + 2\varepsilon) \cap I(g(y_u)) = \{\tau\}$, for all $\tau \in \mathcal{T}_{red}$. Set $I_a = \cup_{\tau\in\mathcal{T}_{red}}(\tau - \varepsilon, \tau + \varepsilon)$ and $I_b = [0, T] \setminus I_a$. Note that I_b is closed. Let N be the cardinal of $\mathcal{T}_{red}$ and denote by $\tau_u^1, \dots, \tau_u^N$ the elements of $\mathcal{T}_{red}$. By definition of reducible touch points and continuity of the mapping $\mathcal{U} \mapsto W^{2,\infty}$, $u \mapsto g(y_u)$, we may apply Lemma 23. Reducing ε if necessary, there exists $\delta > 0$, such that for all $i = 1, \dots, N$, the mappings

$$\mathcal{R}^i : B_\infty(u, \delta) \to \mathbb{R} \quad ; \quad \tilde{u} \mapsto g(y_{\tilde{u}}(\tau_{\tilde{u}}^i)),$$

such that $g(y_{\tilde{u}})$ attains its (unique) maximum over $[\tau_u^i - \varepsilon, \tau_u^i + \varepsilon]$ at time $\tau_{\tilde{u}}^i$, are well-defined. It follows that for all $\tilde{u} \in B_\infty(u, \delta)$,

$$G(\tilde{u}) \in K \quad \text{iff} \quad g(y_{\tilde{u}}(t)) \leq 0 \ \forall t \in I_b \quad \text{and} \quad \mathcal{R}^i(\tilde{u}) \leq 0 \ \forall i = 1, \dots, N. \tag{73}$$

Denote by $g(y_{\tilde{u}})|_b$ the restriction of $g(y_{\tilde{u}})$ to I_b and $\mathcal{R} : \tilde{u} \mapsto (\mathcal{R}^i(\tilde{u}))_{1\leq i\leq N}$. The *reduced problem* is defined as follows:

$$\min_{\tilde{u}\in\mathcal{B}_\infty(u,\delta)} J(\tilde{u}) \quad ; \quad \mathcal{G}(\tilde{u}) = \begin{pmatrix} g(y_{\tilde{u}})|_b \\ \mathcal{R}(\tilde{u}) \end{pmatrix} \in \mathcal{K} := C_-[I_b] \times \mathbb{R}^N_-. \tag{74}$$

From (73), it follows that (74) is locally equivalent to problem (4) in a L^∞ neighborhood of u. The Lagrangian $\mathcal{L}$ of the reduced problem (74) is given, for $\tilde{u} \in B_\infty(u, \delta)$ and $\lambda = (\eta_b, \nu) \in \mathcal{M}_+[I_b] \times \mathbb{R}^N_+$, by:

$$\mathcal{L}(\tilde{u}, \lambda) = J(\tilde{u}) + \int_{I_b} g(y_{\tilde{u}}(t))\mathrm{d}\eta_b(t) + \sum_{i=1}^N \nu_i \mathcal{R}^i(\tilde{u}). \tag{75}$$

The next lemma shows how the Lagrangian, multipliers and critical cone of the reduced problem (74) are related to the ones of problem (4).

Lemma 26 *Assume that $g_u^{(1)} \equiv 0$, and let $u \in G^{-1}(K)$ and $\mathcal{T}_{red}$, I_a, I_b, $\mathcal{R}$, $\mathcal{G}$ and $\mathcal{L}$ be defined as above. Let $\lambda = (\eta_b, \nu) \in \mathcal{M}_+[I_b] \times \mathbb{R}_+^N$. For $\delta > 0$ small enough, the function $\tilde{u} \mapsto \mathcal{L}(\tilde{u}, \lambda)$ is C^1 on $B_\infty(u, \delta)$ and twice Fréchet differentiable at u. Define $\eta \in \mathcal{M}_+[0,T]$ by:*

$$\mathrm{d}\eta(t) = \mathrm{d}\eta_b(t) \ \text{ on } I_b \qquad ; \qquad \mathrm{d}\eta(t) = \sum_{i=1}^{N} \nu_i \delta_{\tau_u^i}(t) \ \text{ on } I_a. \tag{76}$$

Then we have: $\mathcal{L}(u, \lambda) = L(u, \eta)$, $D_u\mathcal{L}(u, \lambda) = D_u L(u, \eta)$,

$$\begin{gathered} D\mathcal{G}(u)^{-1} T_{\mathcal{K}}(D\mathcal{G}(u)) = DG(u)^{-1} T_K(G(u)), \\ \lambda \in N_{\mathcal{K}}(\mathcal{G}(u)) \text{ iff } \eta \in N_K(G(u)), \end{gathered} \tag{77}$$

$$D^2_{uu}\mathcal{L}(u, \lambda) \ = \ D^2_{uu}L(u, \eta) - \sum_{i=1}^{N} \nu_i \frac{(g_y^{(1)}(y_u(\tau_u^i)) z_{u,v}(\tau_u^i))^2}{g^{(2)}(u(\tau_u^i), y_u(\tau_u^i))}. \tag{78}$$

Proof Note that $\mathcal{R}^i = \Phi^i \circ G$, $i = 1, \ldots, N$, where the mappings Φ^i are defined by (69) in Lemma 23 applied to $(x_0, \tau_0) = (g(y_u), \tau_u^i)$. It follows from Lemma 23 that $\mathcal{R}$ is C^1 over a small ball $B_\infty(u, \delta)$. By (70), the second-order expansion of the state (11) and (28) (since $g_u^{(1)} \equiv 0$), that gives $\frac{\mathrm{d}}{\mathrm{d}t} DG(u)v = g_y^{(1)}(y_u) z_{u,v}$, we see that, for all $v \in \mathcal{U}$:

$$D\mathcal{R}^i(u)v = D\Phi^i(G(u))DG(u)v = g_y(y_u(\tau_u^i)) z_{u,v}(\tau_u^i), \tag{79}$$

$$\begin{aligned} D^2\mathcal{R}^i(u)(v,v) &= D\Phi^i(G(u))D^2G(u)(v,v) + D^2\Phi^i(G(u))(DG(u)v, DG(u)v) \\ &= z_{u,v}(\tau_u^i)^* g_{yy}(y_u(\tau_u^i)) z_{u,v}(\tau_u^i) + g_y(y_u(\tau_u^i)) z_{u,vv}(\tau_u^i) \\ &\quad - \frac{(g_y^{(1)}(y_u(\tau_u^i)) z_{u,v}(\tau_u^i))^2}{g^{(2)}(u(\tau_u^i), y_u(\tau_u^i))}. \end{aligned}$$

The conclusion follows easily from the above expressions (see the proof of Lemma 1), (77) is obtained as a consequence of (79).

It follows that if $u \in \mathcal{U}$ and $\Lambda(u) \neq \emptyset$, the Lagrange multipliers λ and η associated with u in problems (74) and (4) respectively, are related by (76). By (77), it follows also that the critical cone $\mathcal{C}(u)$ for problem (74) is equal to $C(u)$. We shall show that the statement of Th. 18 remains true by replacing $L(u, \eta)$ by $\mathcal{L}(u, \lambda)$. That is, the main result of this paper, with Th. 12 (and Th. 18 for first-order state constraint), is the next theorem.

Theorem 27 *Assume that $g_u^{(1)} \equiv 0$ (the state constraint is not of first order). Let $u \in \mathcal{U}$ satisfy (13) with Lagrange multiplier η, and assume that (A2') holds. Let $\mathcal{T}_{red}$ be a finite set of reducible touch points of u, and $\nu_\tau := [\eta(\tau)]$. If the following second-order sufficient condition is satisfied:*

$$D^2_{uu}L(u, \eta)(v,v) - \sum_{\tau \in \mathcal{T}_{red}} \nu_\tau \frac{(g_y^{(1)}(y_u(\tau)) z_{u,v}(\tau))^2}{\frac{\mathrm{d}^2}{\mathrm{d}t^2} g(y_u(t))|_{t=\tau}} > 0 \quad \forall v \in C_{L^2}(u) \setminus \{0\} \tag{80}$$

then u is a local solution of (4) satisfying the quadratic growth condition (5).

Conversely, if (A1)-(A6) hold, then the finitely many essential touch points of the trajectory (u, y_u) are all reducible, and the second-order sufficient condition (80) is satisfied with $\mathcal{T}_{red} = \mathcal{T}_{to}^{ess}$ iff *the quadratic growth condition (5) is satisfied.*

Remark 28 Note that if $\mathcal{T}_{red} = \emptyset$, (80) coincides with (49). If $\mathcal{T}_{red}$ contains essential touch points, then by (32) the contribution in (80) of points in $\mathcal{T}_{red}$ is such that the sum is nonpositive, and therefore the sufficient condition (80) is in general weaker than (49).

We first need to extend Lemma 20 to the Lagrangian $\mathcal{L}$. Note that $\mathcal{L}$ is not C^2 in a L^∞ neighborhood of u, thus (56) does not hold with $\mathcal{L}$.

Lemma 29 *For $\delta > 0$ small enough and all $v \in B_\infty(0, \delta)$,*

$$\mathcal{L}(u+v, \lambda) = \mathcal{L}(u, \lambda) + D_u\mathcal{L}(u, \lambda)v + \frac{1}{2}D^2_{uu}\mathcal{L}(u, \lambda)(v, v) + \tilde{r}(v), \tag{81}$$

with $\tilde{r}(v) = o_\infty(\|v\|_2^2)$.

Proof It is easily seen from (75) and (76) that

$$\mathcal{L}(u+v, \lambda) = L(u+v, \eta) + \sum_{i=1}^{N} \nu_i (g(y_{u+v}(\tau^i_{u+v})) - g(y_{u+v}(\tau^i_u))).$$

We may write $\tilde{r}(v) = r(v) + \hat{r}(v)$, where $r(v)$ is given by (55) and satisfies $r(v) = \mathcal{O}(\|v\|_3^3)$ by Lemma 20, and by (78) we have $\hat{r}(v) = \sum_{i=1}^N \nu_i \hat{r}_i(v)$ with, for $i = 1, \ldots, N$:

$$\hat{r}_i(v) := g(y_{u+v}(\tau^i_{u+v})) - g(y_{u+v}(\tau^i_u)) + \frac{(g_y^{(1)}(y_u(\tau^i_u))z_{u,v}(\tau^i_u))^2}{2g^{(2)}(u(\tau^i_u), y_u(\tau^i_u))}. \tag{82}$$

Fix $i = 1, \ldots, N$, and set $x_0 := g(y_u)$ and $\tau_0 := \tau^i_u$. By definition of reducible touch points, (x_0, τ_0) satisfy the assumptions of Lemma 23. Set $x := g(y_{u+v}) \in W^{2,\infty}$, then $\tau_x = \tau^i_{u+v}$, and since the state constraint is not of first order, we have $\dot{x} = g^{(1)}(y_{u+v})$, $\ddot{x} = g^{(2)}(u+v, y_{u+v})$ and hence, by (50):

$$\|x - x_0\|_{1,\infty} = \mathcal{O}_\infty(\|v\|_1) \quad ; \qquad \|\ddot{x} - \ddot{x}_0\|_\infty = \mathcal{O}_\infty(\|v\|_\infty). \tag{83}$$

Since $g^{(1)}(y_{u+v}) - g^{(1)}(y_u) - g_y^{(1)}(y_u)z_{u,v} = \int_0^1 (g(y_{u+\sigma v})z_{u+\sigma v} - g(y_u)z_u)\mathrm{d}\sigma$, we also have by (50) and (52)-(53), setting $h := g_y^{(1)}(y_u)z_{u,v}$, that

$$\|\dot{x} - \dot{x}_0 - h\|_\infty = \mathcal{O}_\infty(\|v\|_2^2). \tag{84}$$

We may now write $\hat{r}_i(v) = \hat{r}_{i,1}(v) + \hat{r}_{i,2}(v)$ with:

$$\hat{r}_{i,1}(v) = x(\tau_x) - x(\tau_0) + \frac{\dot{x}(\tau_0)^2}{2\ddot{x}_0(\tau_0)} \quad ; \quad \hat{r}_{i,2}(v) = \frac{h(\tau_0)^2 - \dot{x}(\tau_0)^2}{2\ddot{x}_0(\tau_0)}.$$

By (72) and (83), we have $\hat{r}_{i,1}(v) = o_\infty(\|v\|_1^2)$. From $|a^2 - b^2| \leq (2|a| + |a-b|)|a-b|$, $\|h\|_\infty = \mathcal{O}_\infty(\|v\|_1)$ by (52), (84) with $\dot{x}_0(\tau_0) = 0$, and $\|\cdot\|_2 \leq \|\cdot\|_1 \|\cdot\|_\infty$, we see that $\hat{r}_{i,2}(v) = \mathcal{O}_\infty(\|v\|_1 \|v\|_2^2) \leq \mathcal{O}_\infty(\|v\|_1^2 \|v\|_\infty)$. It follows that $\hat{r}_i(v) = o_\infty(\|v\|_1^2)$ for all i and finally that $\tilde{r}(v) = o_\infty(\|v\|_2^2)$, which achieves the proof.

Proof (Proof of Theorem 27) Since the sum of a Legendre form and of a weakly continuous quadratic form remains a Legendre form, we deduce easily from (75) and Lemma 21, since the additional terms

$$v \ \mapsto \ z_{u,v}(\tau_u^i)^* \frac{g_y^{(1)}(y_u(\tau_u^i))^* g_y^{(1)}(y_u(\tau_u^i))}{g^{(2)}(u(\tau_u^i), y_u(\tau_u^i))} z_{u,v}(\tau_u^i)$$

are weakly continuous quadratic forms, that the unique continuous extension of $D_{uu}\mathcal{L}(u,\lambda)$ over L^2 is a Legendre form. In addition, since $\tilde{r}(v) = o_\infty(\|v\|_2^2)$ by Lemma 29, the proof of Theorem 18 still applies, replacing $L(u,\eta)$ by $\mathcal{L}(u,\lambda)$. It follows that (80) implies the quadratic growth condition (5).

Conversely, if (A1)-(A6) hold, there are finitely many essential touch points of (u, y_u), all being reducible. Assume that (5) holds. Then for sufficiently small $\varepsilon > 0$, u is solution of the following problem:

$$(\mathcal{P}_\varepsilon) \qquad \min_{\tilde{u} \in L^2} \{ J^\varepsilon(\tilde{u}) := J(\tilde{u}) - \varepsilon \|\tilde{u} - u\|_2^2 \} \quad ; \quad G(\tilde{u}) \in K, \tag{85}$$

with the same (unique) Lagrange multiplier η, since $D_u J^\varepsilon(u) = D_u J(u)$. Since in addition $(\mathcal{P}_\varepsilon)$ and (4) have the same constraints, they have the same critical cone. Denote the Lagrangian of $(\mathcal{P}_\varepsilon)$ by $L^\varepsilon(u,\eta)$. Note that if assumptions (A0)-(A6) are satisfied for problem (4), so are they for problem (85), so that, for ε small enough, $H_{uu}^\varepsilon = H_{uu} - \varepsilon$ is positive, uniformly over t. Since $D_{uu}^2 L^\varepsilon(u,\eta)(v,v) = D_{uu}^2 L(u,\eta)(v,v) - \varepsilon \|v\|_2^2$, the extended second-order necessary condition in Corollary 15 for $(\mathcal{P}_\varepsilon)$ yields:

$$D_{uu}^2 L(u,\eta)(v,v) - \sum_{\tau \in \mathcal{T}_{to}^{ess}} \nu_\tau \frac{(g_y^{(1)}(y_u(\tau)) z_{u,v}(\tau))^2}{\frac{\mathrm{d}^2}{\mathrm{d}t^2} g(y_u(t))|_{t=\tau}} \geq \varepsilon \|v\|_2^2, \quad \forall v \in C_{L^2}(u). \tag{86}$$

Hence (80) is satisfied with $\mathcal{T}_{red} = \mathcal{T}_{to}^{ess}$.

Note that taking $\mathcal{T}_{red} = \emptyset = \mathcal{T}_{to}^{ess}$ proves the converse in Th. 18, when (u, y_u) has no essential touch point (including the case $q = 1$).

Remark 30 The second-order sufficient condition in (80) remains in quite an abstract form, of little help to check the optimality of a trajectory in application to real life problems. Some *verifiable* second-order sufficient conditions exist in the literature that are based on Riccati equations, see e.g. Maurer [28]. They may be too strong, however, since they ensure in general the coercivity of the Hessian of the Lagrangian over a space that is larger than the critical cone $C_{L^2}(u)$. See also Malanowski et al. [25,26] for first order state constraints.

Remark 31 Handling an infinite number of junction points remains an open problem. It was shown indeed by Robbins in [34], on an example involving a *third order* state constraint, and though satisfying all regularity assumptions (A0)-(A3), that the optimal trajectory has a boundary arc, but except for a nowhere dense subset of initial conditions y_0, the latter is not regular, its entry point being the cluster point of an infinite sequence of touch points.

It happens that boundary arcs with regular junctions may occur for any order of the state constraint q, see for instance the example given in [7, Rem. 4.10]. However, when q is greater than or equal to three, it seems that regular boundary arcs occur only in degenerate (i.e., non generic) situations, and that generically, as Robbins' example suggests, the junctions at boundary arcs are irregular with an infinite sequence of touch points.

6 Conclusion

Our main result is a no-gap condition for an optimal control problem with a single state constraint of any order and only one control. The main hypotheses are that there are finitely many junction points, the essential touch points being reducible, the entry/exit points being regular, and strict complementarity on boundary arcs. The extension of the result to the case when $g(y_u(T)) = 0$ should present no difficulty.

In our recent work [7], we relate these second-order conditions to the study of the well-posedness of the shooting algorithm, and to the characterization of strong regularity in the sense of Robinson [37] (see also related results [9, Section 5.1] and Malanowski [24]). We study in [8] first-order state constraints with non essential touch points, and describe transitions between touch points and boundary arcs.

We hope in the future to extend some of the results of these papers to the case of several state constraints and control variables.

Acknowledgements The authors thank two anonymous referees for their useful suggestions.

A Appendix

Lemma 32 (Extension of Gronwall Lemma) *Let $p \in BV([0,T];\mathbb{R}^n)$ be such that:*

$$|\mathrm{d}p(t)| \leq \kappa |p(t)| \mathrm{d}t + \mathrm{d}\mu(t), \quad \forall t \in [0,T], \tag{87}$$

for some positive constant κ, and a nonnegative bounded measure μ. Then:

$$\|p\|_\infty \leq e^{\kappa T}|p(0)| + \int_0^T e^{\kappa(T-t)} \mathrm{d}\mu(t).$$

Proof Set $\rho(t) = |p(t)|$. Then ρ is a nonnegative bounded measure, and for all $t \in [0,T)$ and $s \to 0^+$, we have:

$$\int_t^{t+s} \mathrm{d}\rho(\sigma) = \rho(t+s) - \rho(t) = |p(t+s)| - |p(t)|$$

$$\leq |p(t+s) - p(t)| = |\int_t^{t+s} \mathrm{d}p(\sigma)| \leq \int_t^{t+s} |\mathrm{d}p(\sigma)|.$$

From (87) it follows that $\rho(t) \leq \varphi(t)$ for all $t \in [0,T]$, where φ is solution of

$$\varphi(t) = |p(0)| + \kappa \int_0^t \varphi(s)\mathrm{d}s + \int_0^t \mathrm{d}\mu(s), \qquad \text{for all } \ t \in [0,T].$$

Then

$$\mathrm{d}(e^{-\kappa t}\varphi(t)) = e^{-\kappa t}\mathrm{d}\varphi(t) - \kappa e^{-\kappa t}\varphi(t)\mathrm{d}t = e^{-\kappa t}\mathrm{d}\mu(t).$$

Therefore, $e^{-\kappa t}\rho(t) = |p(0)| + \int_0^t e^{-\kappa s}\mathrm{d}\mu(s)$. The result follows.

Lemma 33 (Integration by parts) *The following relation holds, for any* $p \in BV([0,T],\mathbb{R}^{n*})$ *and* $z \in W^{1,1}(0,T;\mathbb{R}^n)$*:*

$$\int_0^T \mathrm{d}p(t)z(t) = -\int_0^T p(t)\dot{z}(t)\mathrm{d}t + p(T)z(T) - p(0)z(0). \tag{88}$$

Proof It is sufficient to give the proof when $n = 1$. Since p is of bounded variation, it has at most countably many discontinuity times. Assume first that p is continuous (but possibly not absolutely continuous). Then z and p are uniformly continuous over $[0,T]$. Let $\varepsilon > 0$, and $N > 0$ such that $|z(t) - z(t')|, |p(t) - p(t')| \leq \varepsilon$ for all $|t - t'| \leq 1/N$. For $i = 0, \dots, N$, set $t_i := iT/N$. Let $M > 0$ be larger than $|p(T) - p(0)|$ and $|z(T) - z(0)|$. We have:

$$\begin{aligned}\int_0^T \mathrm{d}p(t)z(t) &\leq \sum_{i=0}^{N-1}(p(t_{i+1}) - p(t_i))(z(t_i) + \varepsilon)\\ &\leq p(T)z(T) - p(0)z(0) + \sum_{i=1}^{N} p(t_i)(z(t_{i-1}) - z(t_i)) + M\varepsilon\\ &\leq p(T)z(T) - p(0)z(0) - \int_0^T p(t)\dot{z}(t)\mathrm{d}t + 2M\varepsilon.\end{aligned}$$

The converse inequality is obtained in the same way. Letting $\varepsilon \to 0$, we obtain the result.

Assume now that p has finitely many discontinuity times distinct of 0 and T (to simplify) $0 \leq \tau_1 < \ldots < \tau_N \leq T$. Set $\tau_0 := 0$ and $\tau_{N+1} := T$. We have:

$$\int_0^T \mathrm{d}p(t)z(t) = \sum_{n=1}^{N+1}\int_{\tau_{n-1}}^{\tau_n} \dot{p}(t)z(t)\mathrm{d}t + \sum_{n=1}^{N}[p(\tau_n)]z(\tau_n).$$

Since on (τ_{n-1}, τ_n), p is continuous and of bounded variation and z is absolutely continuous, we may integrate by parts using the above result and obtain

$$\begin{aligned}\int_0^T \mathrm{d}p(t)z(t) = &\sum_{n=1}^{N+1}\int_{\tau_{n-1}}^{\tau_n} -p(t)\dot{z}(t)\mathrm{d}t - \sum_{n=1}^{N+1} p(\tau_{n-1}^+)z(\tau_{n-1})\\ &+ \sum_{n=1}^{N+1} p(\tau_n^-)z(\tau_n) \ + \ \sum_{n=1}^{N}[p(\tau_n)]z(\tau_n),\end{aligned}$$

which immediately gives (88). Assume now that p has countably many discontinuity times, (τ_n), $n \in \mathbb{N}$, (the latter being not necessarily in increasing order) with jumps discontinuities $\nu_n = [p(\tau_n)]$, such that $\sum_{n\in\mathbb{N}} |\nu_n| \leq \|\mathrm{d}p\| < +\infty$. Denote by

p_c the continuous part of p, such that $\mathrm{d}p = \mathrm{d}p_c + \sum_{n\in\mathbb{N}} \nu_n \delta_{\tau_n}$. For $N > 0$, set $\mathrm{d}p_N := \mathrm{d}p_c + \sum_{n=0}^{N} \nu_n \delta_{\tau_n}$. It follows that, for all N,

$$\int_0^T \mathrm{d}p_N(t) z(t) = -\int_0^T p_N(t)\dot{z}(t)\mathrm{d}t + p_N(T)z(T) - p_N(0)z(0). \tag{89}$$

Since we have

$$|\mathrm{d}(p - p_N)(t)| \leq \sum_{n=N+1}^{\infty} |\nu_n| \delta_{\tau_n}(t),$$

by Lemma 32 we deduce that $\|p - p_N\|_\infty \leq \sum_{n=N+1}^{\infty} |\nu_n| \to 0$ when $N \to +\infty$. Hence we can pass to the limit in (89), which gives the result.

References

1. Bérend, N., Bonnans, J., Haddou, M., Laurent-Varin, J., Talbot, C.: An interior-point approach to trajectory optimization (2006). To appear in J. Guidance, Control and Dynamics
2. Berkmann, P., Pesch, H.: Abort landing in windshear: optimal control problem with third-order state constraint and varied switching structure. J. of Optimization Theory and Applications **85** (1995)
3. Betts, J.: Practical methods for optimal control using nonlinear programming. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (2001)
4. Bonnans, J.: Second order analysis for control constrained optimal control problems of semilinear elliptic systems. Journal of Applied Mathematics & Optimization **38**, 303–325 (1998)
5. Bonnans, J., Cominetti, R., Shapiro, A.: Sensitivity analysis of optimization problems under abstract constraints. Mathematics of Operations Research **23**, 806–831 (1998)
6. Bonnans, J., Cominetti, R., Shapiro, A.: Second order optimality conditions based on parabolic second order tangent sets. SIAM Journal on Optimization **9**, 466–492 (1999)
7. Bonnans, J., Hermant, A.: Well-posedness of the shooting algorithm for state constrained optimal control problems with a single constraint and control (April 2006). URL http://www.inria.fr/rrrt/rr-5889.html. INRIA Research Report 5889
8. Bonnans, J., Hermant, A.: Stability and sensitivity analysis for optimal control problems with a first-order state constraint (July 2006). URL http://hal.inria.fr/inria-00087573. INRIA Research Report
9. Bonnans, J., Shapiro, A.: Perturbation analysis of optimization problems. Springer-Verlag, New York (2000)
10. Bonnard, B., Faubourg, L., Trelat, E.: Optimal control of the atmospheric arc of a space shuttle and numerical simulations with multiple-shooting method. Mathematical Models & Methods in Applied Sciences **15**(1), 109–140 (2005)
11. Bryson, A., Denham, W., Dreyfus, S.: Optimal programming problems with inequality constraints I: necessary conditions for extremal solutions. AIAA Journal **1**, 2544–2550 (1963)
12. Cominetti, R.: Metric regularity, tangent sets and second order optimality conditions. Journal of Applied Mathematics & Optimization **21**, 265–287 (1990)
13. Cominetti, R., Penot, J.: Tangent sets to unilateral convex sets. Comptes Rendus de l'Académie des Sciences de Paris, Série I **321**, 1631–1636 (1995)
14. Haraux, A.: How to differentiate the projection on a convex set in Hilbert space. Some applications to variational inequalities. Journal Mathematical Society of Japan **29**, 615–631 (1977)

15. Hartl, R., Sethi, S., Vickson, R.: A survey of the maximum principles for optimal control problems with state constraints. SIAM Review **37**, 181–218 (1995)
16. Hettich, R., Jongen, H.: Semi-infinite programming: conditions of optimality and applications. In: J. Stoer (ed.) Optimization Techniques. Proc. 8th IFIP Conf. on Optimization Techniques, Würzburg. Part 2, Springer-Verlag, New York (1978)
17. Ioffe, A.: Necessary and sufficient conditions for a local minimum I: A reduction theorem and first order conditions, II: Conditions of Levitin-Miljutin-Osmolovskii type, III: Second order conditions and augmented duality. SIAM Journal on Control Optimization **17**, 245–250, 251–265 and 266–288 (1979)
18. Ioffe, A., Tihomirov, V.: Theory of Extremal Problems. North-Holland Publishing Company, Amsterdam (1979). Russian Edition: Nauka, Moscow, 1974
19. Jacobson, D., Lele, M., Speyer, J.: New necessary conditions of optimality for control problems with state-variable inequality contraints. J. of Mathematical Analysis and Applications **35**, 255–284 (1971)
20. Kawasaki, H.: An envelope-like effect of infinitely many inequality constraints on second order necessary conditions for minimization problems. Mathematical Programming **41**, 73–96 (1988)
21. Kawasaki, H.: The upper and lower second order directional derivatives of a sup-type function. Mathematical Programming **41**, 327–339 (1988)
22. Kawasaki, H.: Second order necessary optimality conditions for minimizing a sup-type function. Mathematical Programming (Ser. A) **49**, 213–229 (1990/91)
23. Kawasaki, H., Zeidan, V.: Conjugate points for variational problems with equality and inequality state constraints. SIAM Journal on Control and Optimization **39**(2), 433–456 (electronic) (2000)
24. Malanowski, K.: Second order conditions and constraint qualifications in stability and sensitivity analysis of solutions to optimization problems in Hilbert spaces. Journal of Applied Mathematics & Optimization **25**, 51–79 (1992)
25. Malanowski, K.: Sufficient optimality conditions for optimal control subject to state constraints. SIAM J. on Control and Optimization **35**, 205–227 (1997)
26. Malanowski, K., Maurer, H., Pickenhain, S.: Second-order sufficient conditions for state-constrained optimal control problems. J. of Optimization Theory and Applications **123**, 595–617 (2004)
27. Maurer, H.: On the minimum principle for optimal control problems with state constraints. Schriftenreihe des Rechenzentrum 41, Universität Münster (1979)
28. Maurer, H.: First and second order sufficient optimality conditions in mathematical programming and optimal control. Mathematical Programming Study (14), 163–177 (1981)
29. Maurer, H., Zowe, J.: First and second-order necessary and sufficient optimality conditions for infinite-dimensional programming problems. Mathematical Programming **16**, 98–110 (1979)
30. Mignot, F.: Contrôle dans les inéquations variationnelles elliptiques. Journal of Functional Analysis **22**, 130–185 (1976)
31. Milyutin, A., Osmolovskii, N.: Calculus of Variations and Optimal Control. American Mathematical Society, Providence (1998)
32. Osmolovskiĭ, N.: Higher-order necessary and sufficient conditions for Pontryagin and restricted-strong minima in an optimal control problem. Doklady Akademii Nauk SSSR **303**(5), 1052–1056 (1988). Translation in Soviet Phys. Dokl. 33 (1988), no. 12, 883–885 (1989)
33. Osmolovskiĭ, N.: Quadratic conditions for nonsingular extremals in optimal control (a theoretical treatment). Russian Journal of Mathematical Physics **2**(4), 487–516 (1995)
34. Robbins, H.: Junction phenomena for optimal control with state-variable inequality constraints of third order. J. of Optimization Theory and Applications **31**, 85–99 (1980)
35. Robinson, S.: First order conditions for general nonlinear optimization. SIAM Journal on Applied Mathematics **30**, 597–607 (1976)
36. Robinson, S.: Stability theorems for systems of inequalities, part II: Differentiable nonlinear systems. SIAM Journal on Numerical Analysis **13**, 497–513 (1976)

37. Robinson, S.: Strongly regular generalized equations. Mathematics of Operations Research **5**, 43–62 (1980)
38. Sokolowski, J.: Sensitivity analysis of control constrained optimal control problems for distributed parameter systems. SIAM Journal on Control and Optimization **25**, 1542–1556 (1987)
39. Zeidan, V.: The Riccati equation for optimal control problems with mixed state-control constraints: necessity and sufficiency. SIAM Journal on Control and Optimization **32**, 1297–1321 (1994)

Numer. Math. (2006) 103: 1–10
DOI 10.1007/s00211-005-0661-y

Numerische
Mathematik

J. Frédéric Bonnans · Julien Laurent-Varin

Computation of order conditions for symplectic partitioned Runge-Kutta schemes with application to optimal control

Order conditions for symplectic partitioned Runge-Kutta schemes (second revision)

Received: 3 December 2004 / Revised: 28 September 2005 / Published online: 16 December 2005

Abstract We derive order conditions for the discretization of (unconstrained) optimal control problems, when the scheme for the state equation is of Runge-Kutta type. This problem appears to be essentially the one of checking order conditions for symplectic partitioned Runge-Kutta schemes. We show that the computations using bi-coloured trees are naturally expressed in this case in terms of *oriented free tree*. This gives a way to compute them by an appropriate computer program.

Mathematics Subject Classification (2000): 49M25 · 34M99 · 65L06

1 Introduction

The motivation of this work comes from an analysis by Hager [8] of order conditions for optimal control problems (of an ordinary differential equation). The idea is to discretize the state equation by a RK (Runge-Kutta) scheme, with a different value of control associated with each “inner state”. Hager observes that the resulting optimality system, after some change of variable, is a PRK (partitioned RK) scheme. Then he computes (by hand, i.e., without a computer code) the order conditions for order up to 4. See also the results of [6] and [7] on constrained optimal

This study is supported by CNES and ONERA, in the framework of the CNES fellowship of the second author.

J.F. Bonnans (✉)
Projet Sydoco, Inria-Rocquencourt, Domaine de Voluceau, BP 105, 78153 Le Chesnay, France
E-mail: Frederic.Bonnans@inria.fr

J. Laurent-Varin
Long-Term Design & System Integration Department, ONERA, and Projet Sydoco,
INRIA Rocquencourt, Domaine de Voluceau, BP 105, 78153 Le Chesnay, France
E-mail: Julien.Laurent-Varin@inria.fr

control problems (a first order state constrained problem, discretized by Euler's scheme, and a control constrained problem with a RK discretization).

There are essentially two hypotheses in the analysis of [8], one on the original problem and the other being a restriction on the scheme. One has to assume that the Hamiltonian is strongly convex w.r.t. the control, or more generally that the second derivative of Hamiltonian w.r.t. the control is invertible. This allows to eliminate the control thanks to the implicit theorem, so that we have an equivalent scheme for the reduced (state, adjoint state) system. The second hypothesis is that none of the coefficients b_i's (of the particular RK scheme) is zero.

The main result of Hager [8] is that, if the original RK scheme is of (global) order p (i.e., when applied to an uncontrolled differential equation) then the resulting scheme has order $q \leq p$, with equality if $p \leq 2$ but strict inequality in some cases whenever $p \geq 3$. In addition, $q = p$ if the scheme is explicit of order at most 4.

For an order greater than four, one cannot do computations by hand. It is then useful to rely on the theory of order conditions for PRK schemes. This theory, based on bi-coloured rooted trees with which certain numbers are associated, is an extension of the original work by Butcher for (non partitioned) RK schemes, see Butcher [5, p. 88].

It appears that the class of PRK schemes coming from the discretization of optimal control problems is in fact partitioned symplectic RK schemes, characterized by relation (4) below. So the question boils down to the one of expressing order conditions for this class. The main result of this paper is that we can obtain the desired expressions using a "calculus" on oriented free trees. To be specific, some weights are associated with each oriented free tree, and the main operation is to "split" any rooted tree into a sum (with coefficients ±1) of such oriented free trees.

We use the correspondance between bi-coloured rooted trees and oriented free trees due to Murua [13]. Our contribution is to show how to construct an explicit expression of the order conditions, whereas [13] gives only a means to compute the number of conditions. Let us also mention the work by Sofroniou and Oevel [14], where order conditions for symplectic, non partitioned RK schemes are obtained using a specific parametrizatic of coefficients a and b.

The paper is organized as follows. In the next section we detail the discretization of optimal control problems by RK schemes, and show the relation with partitioned symplectic RK schemes. satisfying (4). Then in Section 3 we review the theory of order conditions for PRK schemes. Section 3.1 introduces oriented free trees, and shows how the order conditions can be expressed in terms of the latter. Finally section 4 discusses the implementation, and displays the results for order up to 5 and the number of conditions for order up to 7.

2 Discretization of unconstrained optimal control problems

Let $f : I\!R^m \times I\!R^n \rightarrow I\!R^n$ and $\Phi : I\!R^n \rightarrow I\!R$ be C^∞ mappings, and consider the following unconstrained optimal control problem:

$$\text{Min } \Phi(y(T)); \quad \dot{y}(t) = f(u(t), y(t)), \quad t \in [0, T]; \quad y(0) = y^0. \qquad (P)$$

We restrict the analysis to continuous control functions. Let us denote by $H(u, y, p) := p \cdot f(u, y)$ the *pseudo-Hamiltonian* of the problem. The first order necessary optimality conditions of this problem are:

$$\left\{\begin{array}{r@{\,}l} \dot{y}(t) &= f(u(t), y(t)), \\ \dot{p}(t) &= -H_y(u(t), y(t), p(t)), \\ 0 &= H_u(u(t), y(t), p(t)), \\ p(T) &= \Phi'(y(T)), \quad y(0) = y^0. \end{array}\right\} t \in [0, T], \qquad (OC)$$

We say that $(\bar{u}, \bar{y}, \bar{p})$ is an extremal if it satisfies (OC) ($\bar{u}$ being a continuous function). Let $(\bar{u}, \bar{y}, \bar{p})$ be an extremal. If

$$u \mapsto H_{uu}(u, y, p) \quad \text{is invertible along the trajectory,} \qquad (1)$$

then by the implicit functions theorem, in a small L^∞ neighbourhood of this trajectory, we have that $H_u(u(t), y(t), p(t)) = 0$ iff $u = \phi(y(t), p(t))$, where ϕ is a C^∞ mapping. Define the *true Hamiltonian* as $\mathcal{H}(y, p) := H(\phi(y, p), y, p)$. Using $H_u(\phi(y(t), p(t)), y(t), p(t)) = 0$, obtain

$$\mathcal{H}_y(y, p) = H_y(\phi(y, p), y, p); \quad \mathcal{H}_p(y, p) = H_p(\phi(y, p), y, p). \qquad (2)$$

Consequently, under hypothesis (1), (OC) is locally equivalent to the *reduced Hamiltonian system*

$$\begin{array}{ll} \dot{y}(t) = \mathcal{H}_p(y(t), p(t)), & -\dot{p}(t) = \mathcal{H}_y(y(t), p(t)), \quad t \in [0, T], \\ p(T) = \Phi'(y(T)), & y(0) = y^0. \end{array} \qquad (3)$$

Let us turn now to the discussion of the discretization of the optimal control problem (P). The RK discretization considered in [8] is

$$\left\{\begin{array}{l} \text{Min } \Phi(y_N); \\ y_{k+1} = y_k + h_k \sum_{i=1}^s b_i f(u_{ki}, y_{ki}), \\ \quad y_{ki} = y_k + h_k \sum_{j=1}^s a_{ij} f(u_{kj}, y_{kj}), \\ \quad y_0 = y^0, \end{array}\right. \qquad (DP_1)$$

for $k = 0, \ldots, N-1$, $i = 1, \ldots, s$, where $h_k > 0$ is the k step size, and (a, b) is the set of RK coefficients. Choosing different values of controls u_{kj} associated with inner states y_{kj} contrasts with other approaches, in which the discretization of controls is coarser than the one of the state (e.g. [1–3]). Let us rewrite (DP_1) under the equivalent form

$$\text{Min } \Phi(y_N); \quad \left\{\begin{array}{l} 0 = h_k \displaystyle\sum_{i=1}^s b_i K_{ki} + y_k - y_{k+1}, \\ 0 = f(u_{ki}, y_k + h_k \sum_{j=1}^s a_{ij} K_{kj}) - K_{ki}, \\ 0 = y^0 - y_0, \end{array}\right. \qquad (DP_2)$$

for $k = 0, \ldots, N-1$, $i = 1, \ldots, s$. Contract $y_k + h_k \sum_{j=1}^s a_{ij} K_{kj}$ into y_{ki}. The Lagrangian function associated with (DP_2) is:

$$\begin{array}{l} \Phi(y_N) + p^0 \cdot (y^0 - y_0) \\ \quad + \displaystyle\sum_{k=0}^{N-1} \left\{ p_{k+1} \cdot \left(h_k \sum_{i=1}^s b_i K_{ki} + y_k - y_{k+1} \right) + \sum_{i=1}^s \xi_{ki} \cdot (f(u_{ki}, y_{ki}) - K_{ki}) \right\}. \end{array}$$

Here p_{k+1}, ξ_{ki}, and p^0 are Lagrange multipliers associated with constraints of (DP_2). Variables p_k will be interpreted as the discretization of co-state of continuous formulation. We obtain the optimality conditions :

$$
\begin{aligned}
p_N &= \Phi'(y_N), \quad p_1 = p^0, \\
p_k - p_{k+1} &= \sum_{i=1}^{s} f_y(u_{ki}, y_{ki})^\top \xi_{ki}, \\
0 &= h_k b_i p_{k+1} + h_k \sum_{j=1}^{s} a_{ji} f_y(u_{kj}, y_{kj})^\top \xi_{kj} - \xi_{ki}, \\
0 &= f_u(u_{ki}, y_{kj})^\top \xi_{ki}, \quad k = 0 \dots N-1, \quad i = 1 \dots s.
\end{aligned}
$$

Using now the hypothesis that $b_i \neq 0$, set $p_{ki} := \xi_{ki}/(h_k b_i)$ for all $k = 0$ to $N-1$, and $i = 1$ to s. Eliminating the ξ_{ki}'s, get

$$
\left\{
\begin{aligned}
y_{k+1} &= y_k + h_k \textstyle\sum_{i=1}^{s} b_i f(u_{ki}, y_{ki}), \\
y_{ki} &= y_k + h_k \textstyle\sum_{j=1}^{s} a_{ij} f(u_{kj}, y_{kj}), \\
p_{k+1} &= p_k - h_k \textstyle\sum_{i=1}^{s} \hat{b}_i H_y(u_{ki}, y_{ki}, p_{ki}), \\
p_{ki} &= p_k - h_k \textstyle\sum_{j=1}^{s} \hat{a}_{ij} H_y(u_{kj}, y_{kj}, p_{kj}), \\
0 &= H_u(u_{ki}, y_{ki}, p_{ki}), \\
y_0 &= y^0, \qquad p_N = \Phi'(y_N),
\end{aligned}
\right. \qquad (DOC)
$$

where coefficients $\hat{b}$ and $\hat{a}$ are defined by the following relations:

$$
\hat{b}_i := b_i, \quad \hat{a}_{ij} := b_j - \frac{b_j}{b_i} a_{ji}, \quad i = 1, \dots, s \;\; j = 1, \dots, s. \tag{4}
$$

If the algebraic constraints $H_u(u_{ki}, y_{ki}, p_{ki}) = 0$ are locally equivalent to $u_{ki} = \phi(y_{ki}, p_{ki})$, then (DOC) is equivalent to the same PRK scheme applied to the reduced system (3). This approach based on formulation (DP_2) is slightly simpler, but equivalent to the one of Hager [8].

It is said that a PRK scheme (or more generally any one step scheme) is symplectic if the corresponding flow is symplectic. It is known that PRK schemes satisfying (4) are symplectic, see [11, Theorem 4.6]; we denote this class by SPRK. We obtain that the scheme obtained by discretization of problem (P) belongs to SPRK. In particular the following diagram commutes, when we use the above discretization:

$$
\begin{tikzcd}
(P) \arrow[r, "\text{discretization}"] \arrow[d, "\text{optimality conditions}"'] & (DP) \arrow[d, "\text{optimality conditions}"'] \\
(OC) \arrow[r, "\text{discretization}"] & (DOC)
\end{tikzcd} \qquad (D)
$$

For a detailed presentation of PRK and symplectic methods we refer to the books [9, 10].

3 Order conditions for symplectic PRK schemes

A basic tool in the study of order conditions is the theory of B-series and associated calculus on rooted trees, Butcher [5]. For PRK schemes an extension of this theory is the one of expansion in P-series, and the associated calculus on bi-coloured rooted trees, see [9]. The latter allows to state the order conditions in terms of coefficients a, b, $\hat{a}$, and $\hat{b}$, of the following type: certain polynomials (which are in fact the sum of monomials with unit coefficients) in these variables have to be equal to certain fractional numbers. The substitution of $\hat{a}$, and $\hat{b}$ (using (4)) would give complicated expressions, since instead of each monomial we would have a sum of rational fractions. We will show that among all these terms it is sufficient to express a condition on a "principal term" since the other terms of the sum are already determined by previous conditions. The resulting simplification allows us to display these conditions up to order 5.

We assume that the reader has some basic knowledge of B and P-series, see [9]. As in [13] we use H-trees, i.e., oriented free trees. These are connected graphs defined by a pair (V, E), the sets of vertices and edges resp., such that $\#E = \#V - 1$.

Bi-coloured graphs are graphs together with a mapping that to each vertex v associates a colour $c(v)$, of value B or W (black and white).

We denote by $V_B = c^{-1}(\{B\})$, $V_W = c^{-1}(\{W\})$ the set of black, white vertices, and by E_B, E_W the set of edges ending on black, white vertices.

For RK (PRK) schemes, we have order condition of the form : $\phi(t) = 1/\gamma(t)$ where t is a (bi-coloured) rooted tree, ϕ the elementary weight function of coefficient of the scheme, and γ the density associated to the tree. We will extend the formalism for symplectic PRK, and get order condition of the form $\Phi(h) = \delta(h)$, where h is now an H-tree, $\Phi(h)$ its elementary weight and $\delta(h)$ is a rational number, defined respectively in definition 1 and theorem 2. The advantage of the new condition $\Phi(h) = \delta(h)$ is that its expression is much simpler than the one of $\phi(t) = 1/\gamma(t)$, t being a bi-coloured rooted tree, for PRK schemes, expressed in terms of a and b.

3.1 Calculus on oriented free trees

We may assume that the set of vertices V is $\{1, \dots, \#V\}$. We start by stating alternative expressions of the elementary weight $\phi(t)$, where t is a rooted bi-coloured tree, for a general partitioned scheme. For a given bi-coloured graph, whose vertices are associated with letters i, j, ... it is convenient to denote

$$\begin{array}{rl} \tilde{b}_{i_k} = b_{i_k} & \text{if vertex } k \text{ is white, } \hat{b}_{i_k} \text{ otherwise.} \\ \tilde{a}_{i_k i_\ell} = a_{ij} & \text{if vertex } \ell \text{ is white, } \hat{a}_{ij} \text{ otherwise.} \end{array} \tag{5}$$

Here i_k associates with each vertex $k \in \{1, \dots, \#V\}$ a number varying from 1 to s.

Definition 1 *For a given* oriented bi-coloured graph $g = (V, E, c)$ *we define the* elementary weight *as follows :*

$$\Phi(g) := \sum_{i_v=1, v \in V}^{s} \prod_{k \in V} \tilde{b}_{i_k} \prod_{(k,\ell) \in E} \tilde{a}_{i_k i_\ell} / \tilde{b}_{i_\ell}, \tag{6}$$

where $\sum_{i_v=1, v\in V}^{s} = \sum_{i_1=1}^{s} \cdots \sum_{i_{\#V}=1}^{s}$.

This definition is consistent with the one of weights ϕ for bi-coloured rooted trees, stated in the theory of PRK schemes (e.g. [10, ch. II 15]). Indeed, the root induces an orientation on edges from root to leaves. For the corresponding oriented bi-coloured graph g, all the $\tilde{b}_{i_\ell}$ in (6) cancel except $\tilde{b}_{i_1}$ so that $\Phi(g) = \phi(t)$. However we can observe, that the above formula makes sense for (non necessarily connected) bi-coloured oriented graphs. Any such graph g is a finite union of connected graphs with empty intersections of vertices, called *connected components* of g, and denoted $\{g^q, q \in Q\}$.

Lemma 1 *Let the bi-coloured oriented graph $g = (V, E, c)$ have connected components $\{g^q, q \in Q\}$. Then $\Phi(g) = \prod_{q \in Q} \Phi(g^q)$.*

Proof The product term in $\Phi(g)$ may be factored on terms depending on each connected component:

$$\Phi(g) = \sum_{i_v=1, v\in V}^{s} \prod_{q\in Q} \left(\prod_{k\in V_q} \tilde{b}_{i_k} \prod_{(k,\ell)\in E_q} \frac{\tilde{a}_{i_k i_\ell}}{\tilde{b}_{i_\ell}} \right).$$

Denote by V_q the set of vertices in the qth connected component. We may then rewrite this sum of products as products of sums:

$$\Phi(g) = \prod_{q\in Q} \left(\sum_{i_v=1, v\in V}^{s} \left(\prod_{k\in V_q} \tilde{b}_{i_k} \prod_{(x,y)\in E_q} \frac{\tilde{a}_{i_k i_\ell}}{\tilde{b}_{i_\ell}} \right) \right).$$

The conclusion follows. □

Given an oriented graph $g = (V, E)$, and $F \subset E$, the set of arcs in opposite direction to those of F is denoted as

$$F^\top := \{(x, y) \in V \times V; \quad (y, x) \in F\}. \tag{7}$$

Theorem 1 *The elementary weight of a bi-coloured oriented graph $g = (V, E, c)$, when* (4) *holds, satisfies*

$$\Phi(g) = \textstyle\sum_{\hat{E}_B \in \mathcal{P}(E_B)} (-1)^{\#\hat{E}_B} \Phi(V, E_W \cup \hat{E}_B^\top). \tag{8}$$

where all vertices of oriented graph $(V, E_W \cup \hat{E}_B^\top)$ are white, and $\mathcal{P}(E_B)$ denotes the set of all subsets of E_B.

Proof Substituting the expressions of $\hat{a}$ and $\hat{b}$ in (4), we may write the elementary weight (6) as follows:

$$\Phi(g) = \sum_{i_v=1, v\in V}^{s} \prod_{k\in V} b_{i_k} \prod_{(k,\ell)\in E_W} \frac{a_{i_k i_\ell}}{b_{i_\ell}} \prod_{(k,\ell)\in E_B} \left(1 - \frac{a_{i_\ell i_k}}{b_{i_k}} \right). \tag{9}$$

Expanding the last term, we get

$$\Phi(g) = \sum_{\hat{E}_B \in P(E_B)} (-1)^{\#\hat{E}_B} \sum_{i_v=1, v\in V}^{s} \prod_{k\in V} b_{i_k} \prod_{(k,\ell)\in E_W} \frac{a_{i_k i_\ell}}{b_{i_\ell}} \prod_{(k,\ell)\in \hat{E}_B} \frac{a_{i_\ell i_k}}{b_{i_k}}.$$

The conclusion follows. □

In the expression (8), the graphs $(V, E_W \cup \hat{E}_B^\top)$ on the right hand side, are mono-coloured oriented graph. Let h be a bi-coloured H-tree. Then the only connected graph in sum of the right-hand-side of (8) is the one with $\hat{E}_B = E_B$. Observe that given an (mono-coloured) H-tree h, we can reconstruct a bi-coloured rooted tree t^h of the same order having h in its expansion in (8), as follows: take an arbitrary vertex of h as the root, of say white colour (since $b = \hat{b}$, a black root would give the same elementary weight); then for each path from root to leaves, let the next vertex be white if the edge is oriented towards the leaves, and black otherwise. In view of the expression of weights for bi-coloured rooted trees [10, ch. II 15] we can rewrite (8), separating the principal term for the others, as

$$\phi(t^h) = \sigma_0 \Phi(h) + \textstyle\sum_{i \in I} \sigma_i \Phi(g_i^h), \qquad (10)$$

where

$$I = \mathcal{P}(E_B) \setminus E_B, \;\; g_i^h = (V, E_W \cup i^\top), \;\; \sigma_i = (-1)^{\#i}, \;\; \sigma_0 = (-1)^{\#E_B}. \qquad (11)$$

Theorem 2 *For an SPRK scheme, the conditions for global order n are given by :*

$$\Phi(h) = \delta(h) \qquad (12)$$

for all H-trees h of order not more than n, with $(I, g_i^h, \sigma_i, \sigma_0)$ *defined by* (11)*, and* $\delta(h)$ *inductively defined as*

$$\delta(h) = \sigma_0 \left(\tfrac{1}{\gamma(t^h)} - \textstyle\sum_{i \in I} \sigma_i \prod_{j \in J_i} \delta(h_i^j) \right). \qquad (13)$$

Here h_i^j*,* $j \in J_i$*, are the connected components of* g_i^h*.*

Proof An SPRK scheme being a PRK scheme, we have to check the order conditions for PRK schemes, whose expression for order n is $\Phi(t) = 1/\gamma(t)$, for all bicoloured rooted trees t of order not more than n. Let us now proceed by induction over this order n of a SPRK scheme. For order 1, the statement is obvious. Assume it to hold for $n - 1$. Lemma 1 combined with our induction hypothesis implies $\Phi(g_i^h) = \prod_{j \in J_i} \Phi(h_i^j) = \prod_{j \in J_i} \delta(h_i^j)$, where g_i^h is defined in (11) and h_i^j are its connected components. We conclude with (10). □

Remark 1 We recover the result of Murua [13]: there are as many order conditions as there exist H-trees of order n. As mentioned in the introduction, the derivation of the H-tree from a bi-coloured rooted tree is already in [13]. Our "calculus" on graphs has the property of generating additional non connected graphs. They allow to take advantage of the order conditions for smaller n, to simplify the expression of order conditions.

3.2 Computation procedure

Theorem 2 provides an algorithm for computing order conditions of SPRK. We generate all H-trees h of order up to n, the corresponding bi-coloured tree t^h and its density $\gamma(t^h)$. Our induction over n allows to compute δ by (13).

Since we used a code for tree generation of Li and Ruskey [12], we generated all bi-coloured trees and computed the conditions for all related H-trees (checking the compatibility of results in the case of already generated H-trees).

Table 1 Number of order conditions

Order	1	2	3	4	5	6	7
Simple	1	1	2	4	9	20	48
Symplectic	1	1	3	8	27	91	350
Partitioned	2	4	14	52	214	916	4116

4 Computational results

Table 1 states the number of conditions. Let us observe the rapid increase of these numbers with p for symplectic schemes, and even more with general partitioned schemes.

The next tables use $c_i = \sum_j a_{ij}$. Indexes in sum vary from 1 to s, is the number of stages in the RK method. All (latex source latex) the tables are automatically generated by the computer code. Conditions for order 1 to 4 were already obtained by Hager [8]. We display the order conditions up to order 5 (order 6 conditions are displayed in [4]).

Table 2 Ordre 1

Graph	Condition
•	$\sum b_i = 1$

Table 3 Ordre 2

Graph	Condition
	$\sum d_j = \frac{1}{2}$

Table 4 Ordre 3

Graph	Condition	Graph	Condition
	$\sum c_j d_j = \frac{1}{6}$		$\sum b_i c_i^2 = \frac{1}{3}$
	$\sum \frac{1}{b_k} d_k^2 = \frac{1}{3}$		

Table 5 Ordre 4

Graph	Condition	Graph	Condition
	$\sum \frac{1}{b_k} a_{lk} d_k d_l = \frac{1}{8}$		$\sum a_{jk} d_j c_k = \frac{1}{24}$
	$\sum \frac{b_i}{b_k} a_{ik} c_i d_k = \frac{5}{24}$		$\sum b_i a_{ij} c_i c_j = \frac{1}{8}$
	$\sum c_j^2 d_j = \frac{1}{12}$		$\sum b_i c_i^3 = \frac{1}{4}$
	$\sum \frac{1}{b_k} c_k d_k^2 = \frac{1}{12}$		$\sum \frac{1}{b_l^2} d_l^3 = \frac{1}{4}$

Table 6 Ordre 5

Graph	Condition	Graph	Condition
	$\sum \frac{b_i}{b_k} a_{ik} a_{il} d_k c_l = \frac{3}{40}$		$\sum b_i a_{ik} a_{ij} c_j c_k = \frac{1}{20}$
	$\sum a_{lk} a_{kj} c_j d_l = \frac{1}{120}$		$\sum b_i a_{ik} a_{kj} c_i c_j = \frac{1}{30}$
	$\sum \frac{b_i}{b_k} a_{lk} a_{il} c_i d_k = \frac{11}{120}$		$\sum \frac{b_i}{b_k} a_{lk} a_{ik} c_i d_l = \frac{3}{40}$
	$\sum \frac{b_i b_j}{b_k} a_{jk} a_{ik} c_i c_j = \frac{2}{15}$		$\sum \frac{b_i}{b_l b_m} a_{im} a_{il} d_l d_m = \frac{2}{15}$
	$\sum \frac{1}{b_k} a_{ml} a_{lk} d_k d_m = \frac{1}{30}$		$\sum \frac{1}{b_k} a_{mk} a_{lk} d_l d_m = \frac{1}{20}$
	$\sum \frac{1}{b_l b_m} a_{lm} d_l^2 d_m = \frac{1}{15}$		$\sum \frac{1}{b_k} a_{kl} d_k^2 c_l = \frac{1}{60}$
	$\sum \frac{1}{b_l^2} a_{ml} d_l^2 d_m = \frac{1}{10}$		$\sum \frac{b_i}{b_l^2} a_{il} c_i d_l^2 = \frac{3}{20}$
	$\sum \frac{1}{b_k} a_{lk} d_k c_l d_l = \frac{7}{120}$		$\sum a_{jk} c_j d_j c_k = \frac{1}{40}$
	$\sum \frac{1}{b_k} a_{lk} c_k d_k d_l = \frac{1}{40}$		$\sum \frac{b_i}{b_k} a_{ik} c_i c_k d_k = \frac{7}{120}$
	$\sum \frac{b_i}{b_k} a_{ik} c_i^2 d_k = \frac{3}{20}$		$\sum b_i a_{ij} c_i^2 c_j = \frac{1}{10}$
	$\sum a_{kj} c_j^2 d_k = \frac{1}{60}$		$\sum b_i a_{ij} c_i c_j^2 = \frac{1}{15}$
	$\sum c_j^3 d_j = \frac{1}{20}$		$\sum b_i c_i^4 = \frac{1}{5}$
	$\sum \frac{1}{b_k} c_k^2 d_k^2 = \frac{1}{30}$		$\sum \frac{1}{b_l^2} c_l d_l^3 = \frac{1}{20}$
	$\sum \frac{1}{b_m^3} d_m^4 = \frac{1}{5}$		

Acknowledgements We thank the associated editor and two referees for their useful remarks.

References

1. Betts, J.T.: Survey of numerical methods for trajectory optimization. AIAA J. Guidance, Control and Dynamics **21**, 193–207 (1998)
2. Betts, J.T.: Practical methods for optimal control using nonlinear programming. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2001

3. Bonnans, J.F., Launay, G.: Large scale direct optimal control applied to a re-entry problem. AIAA J. of Guidance, Control and Dynamics **21**, 996–1000 (1998)
4. Bonnans, J.F., Laurent-Varin, J.: Computation of order conditions for symplectic partitioned runge-kutta schemes with application to optimal control. Rapport de recherche RR-5398, INRIA, 2004. http://www.inria.fr/rrrt/rr-5398.html
5. Butcher, J.C.: The numerical analysis of ordinary differential equations. A Wiley-Interscience Publication. John Wiley & Sons Ltd., Chichester, 2003
6. Dontchev, A.L., Hager, W.W.: The Euler approximation in state constrained optimal control. Math. Comput. **70**, 173–203 (2001)
7. Dontchev, A.L., Hager, W.W., Veliov, V.M.: Second-order Runge-Kutta approximations in control constrained optimal control. SIAM J. Numer. Anal. **38**, 202–226 (2000) (electronic)
8. Hager, W.: Runge-Kutta methods in optimal control and the transformed adjoint system. Numerische Mathematik **87**(2), 247–282 (2000)
9. Hairer, E., Lubich, Ch., Wanner, G.: Geometric numerical integration. Springer Series in Computational Mathematics, vol 31, Springer-Verlag, Berlin, 2002
10. Hairer, E., Nørsett, S.P., Wanner, G.: Solving ordinary differential equations. I, volume 8 of Springer Series in Computational Mathematics. Springer-Verlag, Berlin, second edition, 1993
11. Hairer, E., Wanner, G.: Solving ordinary differential equations. II. Springer-Verlag, Berlin, second edition, 1996
12. Li, G., Ruskey, F.: The advantages of forward thinking in generating rooted and free trees. In: 10th Annual ACM-SIAM symposium on discrete algorithms (SODA), 1999. http://www.theory.csc.UVic.CA/ fruskey/ ou http://www.theory.csc.UVic.CA/ cos
13. Murua, A.: On order conditions for partitioned symplectic methods. SIAM J. Numer. Anal. **34**, 2204–2211 (1997)
14. Sofroniou, M., Oevel, W.: Symplectic Runge-Kutta schemes. I. Order conditions. SIAM J. Numer. Anal. **34**, 2063–2086 (1997)