Optimal control of ordinary differential equations[1]

J. Frédéric Bonnans[2]

June 20, 2008

[1]Lecture notes, CIMPA School on Optimization and Control, Castro Urdiales, August 28 - September 8, 2006. Revised version, June 18, 2006.

[2]INRIA-Saclay and Centre de Mathématiques Appliquées (CMAP), Ecole Polytechnique, 91128 Palaiseau, 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.

We end with some additional references that may be useful. Chapter 3 uses some material from the paper [12], coauthored with J. Laurent-Varin. The time limitation did not allow to study state constrained optimal control problems. See on this subject the recent papers [8, 9, 10], coauthored with A. Hermant. The discussion of optimization algorithms to be used is an important related subject. Classical references are Betts [4, 5]. Applications to aerospace problems may be found in [3, 14].

These notes are in some sense a continuation of the book [13] written with A. Shapiro, devoted to the sensitivity analysis for general optimization problems. Nonsmooth analysis is presented in Rockafellar and Wets [37] and used in the study of optimal control problems in many places; let us quote Clarke and coworkers [16, 17], Frankowska [24]. A classical and still useful reference is Ioffe and Tihomirov [29]. A more recent book on optimal control is Milyutin and Osmolovskii [33].

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.

Note on the revision of June 2008 Following remarks and discussions with S. Aronna (U. Rosario) and P. Lotito (U. Tandil), various typos were corrected. I thank them for their careful reading of the notes.

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

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

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

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

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

$$\begin{array}{rcll} 0 & = & \dot{P}_t + P_t A_t + A_t^\top P_t + C_t - (P_t B_t + D_t^\top) R_t^{-1} (B_t^\top P_t + D_t) & t \in \mathcal{T}, \\ P_T & = & M. & \end{array} \tag{1.1.21}$$

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

$$fFx) = \tfrac{1}{2} x \cdot P_s x. \tag{1.1.25}$$

Consequently, the nonnegativity of 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 [13, Lemma 4.123].

Lemma 1.10 *Assume that $D^2 f(\bar{u}) \succeq 0$. 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 [13, 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 sequentially 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, $Q(w^k) \to 0$ implies that $w^k \to 0$ 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:*

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

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

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 $f_1(u) := \frac{1}{2}\int_0^T u_t \cdot R_t u_t \mathrm{d}t$ is the part that does not depend on the state 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 $I \subset (s, T)$ of nonzero measure such that

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

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

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

This implies that f is not w.l.s.c. So we have proved that 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) of example 1.13).

Otherwise, if R_t is not uniformly positive, 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

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

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.

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] or Dunford and Schwartz [22]. ■

Lemma 1.17 *We have that*

$$\limsup_{s \uparrow T} \frac{H_s(v, v)}{\|v\|^2_{\mathcal{U}_s}} = 0 \tag{1.1.35}$$

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

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

■

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

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

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*

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

In addition,

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

Proof. Relation (1.2.37) follows from the well-known first-order optimality condition

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

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 in K such that $x^k \to x$, $x^k \neq x$ for all k, and

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

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. Conversely, 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

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

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

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

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 $\bar{x}$)

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

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*

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

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

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

with the quadratic growth condition (1.2.49), we get

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

implying (1.2.50). ■

Remark 1.25 The proof extends to the sensitivity analysis for approximate solutions. In the above proof, take for x_u a $\varepsilon(u)$ solution, where $\varepsilon(u) \to 0$ when $\varepsilon \downarrow 0$. Using $f(x_u, u) - f(\bar{x}, u) \leq \varepsilon(u)$, we obtain $\alpha \|x_u - \bar{x}\|^2 \leq O\left(\|x_u - \bar{x}\| \, \|u - \bar{u}\| \right) + \varepsilon(u)$. Set $\delta_u := \alpha^{1/2} \|u - \bar{u}\|$ and $\chi_u := \|x_u - \bar{x}\|$. We have that $\chi_u^2 \leq c \chi_u \delta_u + \varepsilon(u)$, for some $c > 0$. Therefore

$$\chi_u = \tfrac{1}{2}(c\delta_u + \sqrt{c^2 \delta_u^2 + 4\varepsilon(u)}) = \tfrac{1}{2} c \delta_u (1 + \sqrt{1 + 4\varepsilon(u)/(c\delta_u)^2}). \tag{1.2.51}$$

In particular, if x_u is a $O(\|u - \bar{u}\|^2)$ solution, then $\|x_u - \bar{x}\| = O(\|u - \bar{u}\|)$. If x_u is a $O(\|u - \bar{u}\|^s)$ for $s < 2$, then $\|x_u - \bar{x}\| = O(\|u - \bar{u}\|^{s/2})$.

1.2.5 Sensitivity analysis

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

$$u(t) = \bar{u} + td + r(t); \quad \|r(t)\| = o(t). \tag{1.2.52}$$

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

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

Theorem 1.26 *Assume that K is polyhedric, that f is weakly l.s.c., that $D^2f(\bar{x})$ is a Legendre form, and that the second-order condition* (1.2.47) *is satisfied. Then the value function may be expanded as follows:*

$$v(t) = v(0) + D_u f(\bar{x},\bar{u})(u(t)-\bar{u}) + \tfrac{1}{2}t^2 \operatorname{val}(SP) + o(t^2). \tag{1.2.53}$$

In addition, let x_t be a $o(t^2)$-solution. Then $h_t := (x_t - \bar{x})/t$ is bounded, and any weak limit-point $\bar{h}$ is a strong limit-point, that satisfies $\bar{h} \in S(SP)$. If (SP) has the unique solution $\bar{h}$, then the following expansion of solutions holds

$$x_t = \bar{x} + t\bar{h} + o(t). \tag{1.2.54}$$

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

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

The following holds:

$$\begin{array}{rl} f(\bar{x}+th,u(t)) = & f(\bar{x},\bar{u}) + D_u f(\bar{x},\bar{u})(u(t)-\bar{u}) \\ & +\frac{1}{2}t^2 D^2 f(\bar{x},\bar{u})((h,d),(h,d)) + o(t^2). \end{array} \tag{1.2.55}$$

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

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

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

$$v(t) \le f(\bar{x},\bar{u}) + D_u f(\bar{x},\bar{u})(u(t)-\bar{u}) + \tfrac{1}{2}t^2 \operatorname{val}(SP) + o(t^2). \tag{1.2.57}$$

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

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

Comparing to (1.2.57), obtain after division by $\frac{1}{2}t^2$

$$2t^{-1} D_x f(\bar{x},\bar{u})h_t + D^2 f(\bar{x},\bar{u})((h_t,d),(h_t,d)) \le \operatorname{val}(SP) + o(1). \tag{1.2.58}$$

This implies $D_x f(\bar{x},\bar{u})h_t \le o(t)$, and hence, $D_x f(\bar{x},\bar{u})\bar{h} \le 0$. 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 $D_x f(\bar{x},\bar{u})h_t \ge 0$. Using the weak l.s.c. of $D^2f(\bar{x},\bar{u})$, get with (1.2.58)

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

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. We have proved (1.2.53). In particular, if (SP) has a unique solution, then $h_t \to \bar{h}$, implying (1.2.54). ■

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 1.27 (Dominated convergence) *Let x_n a sequence of elements of $L^2(\Omega)$. Suppose that there exists $g \in L^2(\Omega)$ such that $|x_n(\omega)| \leq g(\omega)$ 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 1.28 (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*

$$T_K(x) := \{h \in X;\ h \geq 0, \quad a.e. \quad sur\ \ I(x)\}, \tag{1.2.59}$$
$$N_K(x) := \{h \in X_-;\ h = 0, \quad a.e. \quad sur\ \ J(x)\}. \tag{1.2.60}$$

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

$$T_K(x) \cap q^\perp = \{h \geq 0, \quad a.e. \quad sur\ \ I(x);\ \ h(\omega)q(\omega) = 0 \quad a.e.\ \}. \tag{1.2.61}$$

(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. [13, Example 2.62]:

$$N_K(x) = K^- \cap x^\perp \tag{1.2.62}$$

The one of the tangent cone follows, using the relation $T_K(x) = N_K(x)^-$, and the latter implying (1.2.61).
(iv) Let $h \in T_K(x) \cap q^\perp$, where $q \in N_K(x)$. Set, for $\varepsilon > 0$, $h_\varepsilon := ((x + \varepsilon h)_+ - x)/\varepsilon$. Then $x + \varepsilon h_\varepsilon = (x + \varepsilon h)_+ \in K$, and hence, $h_\varepsilon \in \mathcal{R}_K(x)$. 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:

$$\left\{ \begin{array}{l} Df(\bar{x})(\omega) \geq 0, \quad Df(\bar{x})(\omega)x(\omega) = 0, \ \text{ a.e.} \\ D^2 f(\bar{x})(h,h) > 0, \ \text{ for all } \ h \geq 0 \ \ \text{over} \ \ I(\bar{x}), \quad h \neq 0, \\ Df(\bar{x})(\omega)h(\omega) = 0 \quad \text{a.e.} \end{array} \right. \tag{1.2.63}$$

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

$$K := \{u \in \mathcal{U};\ g(u(t)) \leq 0,\ a.e.\ t \in (0,T)\}. \tag{1.3.64}$$

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

$$g(0) = 0. \tag{1.3.65}$$

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

$$Df(\bar{u})(u - \bar{u}) \geq 0, \quad \text{for all } u \in K, \tag{1.3.66}$$

or equivalently

$$Df(\bar{u}) + N_K(\bar{u}) \ni 0. \tag{1.3.67}$$

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 1.29 *Assume that R_t is uniformly positive:*

$$\exists\ \alpha > 0;\ u \cdot R_t u \geq \alpha |u|^2, \quad \text{for almost all } t \in (0,T). \tag{1.3.68}$$

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 (1.3.66), we obtain that the following holds:

$$(B_t^\top p_t + R_t \bar{u}_t + D_t y_t) \cdot (v - \bar{u}_t) \geq 0, \quad \text{for all } v \in g^{-1}(]-\infty, 0]), \quad t \in [0,T]. \tag{1.3.69}$$

In view of (1.3.65), we may take $v = 0$, obtaining (using (1.3.68) and the fact that B_t, D_t, p, y are essentially bounded)

$$\alpha |\bar{u}_t|^2 \leq \bar{u}_t \cdot R_t \bar{u}_t \leq (B_t^\top p_t + D_t y_t) \cdot \bar{u}_t \leq c|\bar{u}_t| \quad t \in [0,T], \tag{1.3.70}$$

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 1.30 *Let $u \in K$. Then*

$$T_K(u) = \{v \in \mathcal{U};\ v_t \in T_{K_g}(u_t) \ \text{ for almost all } t \in (0,T)\}. \tag{1.3.71}$$

$$N_K(u) := \{\mu \in \mathcal{U};\ \mu_t \in N_{K_g}(u_t) \ \text{ for almost all } t \in (0,T)\}. \tag{1.3.72}$$

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

$$v_t^\varepsilon = \varepsilon^{-1}(P_{K_g}(u_t + \varepsilon v_t) - u_t), \quad \text{a.e. } t \in (0,T). \tag{1.3.73}$$

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 (1.3.71) follows, and (1.3.72) is an easy consequence of (1.3.71). ■

Our aim now is to relate the expression of the Lagrange multipliers to $g(u)$ and $Dg(u)$. Note that $g(u)$ is not necessarily integrable for given $u \in \mathcal{U}$ (even if $u \in K$). Therefore we cannot apply standard calculus rules for computing tangent and normal cones and we must do a specific construction. We need the following qualification condition:

$$\exists\ \beta > 0 \ \text{ and } \ u^0 \in \mathbb{R}^m; \quad g_i(u^0) < -\beta, \quad i = 1, \ldots, n_g. \tag{1.3.74}$$

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

$$T_{K_g}(u) = \{v \in \mathbb{R}^m;\ Dg_i(u)v \leq 0, \ \text{ for all } \ i;\ g_i(u) = 0\} \tag{1.3.75}$$

$$N_{K_g}(u) = \left\{ \sum_{i=1}^{n_g} \lambda_i Dg_i(u);\ \lambda \in \mathbb{R}^m_+;\ \lambda_i = 0, \ \text{ for all } \ i;\ g_i(u) < 0 \right\}. \tag{1.3.76}$$

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

$$I_t(u) := \{1 \le i \le n_g;\ g_i(u_t) = 0\}. \tag{1.3.77}$$

Lemma 1.31 *Let $u \in K$ be such that the qualification condition* (1.3.74) *holds. Then*

$$T_K(u) = \{v \in \mathcal{U};\ Dg_i(u_t)v_t \le 0, \quad \textit{for a.a. } \ t \in (0,T), \ \ i \in I_t(u_t)\}, \tag{1.3.78}$$

$$\begin{array}{rl} N_K(u) = & \{\, \mu \in \mathcal{U};\ \ \mu_t = \sum_{i=1}^{n_g} \lambda_{i,t} Dg_i(u_t);\ \lambda_{i,t} \in \mathbb{R}^m_+; \\ & \lambda_{i,t} = 0, \ \ \textit{for all } \ i;\ g_i(u_t) < 0 \ \ \textit{a.e. } \ t \in (0,T)\}. \end{array} \tag{1.3.79}$$

In addition we have that if λ satisfies (1.3.79)*, then*

$$\sum_i |\lambda_{i,t}| \le \beta^{-1} |\mu_t|\, |u^0 - u_t|. \tag{1.3.80}$$

Proof. Relations (1.3.78) and (1.3.79) are immediate consequences of the above relations. If λ satisfies (1.3.79), then since g is convex, then a.e., for all $i \in I_t(u)$:

$$-\beta \ge g(u^0) \ge Dg_i(u_t)(u^0 - u_t). \tag{1.3.81}$$

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

$$-\beta \sum_i |\lambda_{i,t}| \ge \mu_t \cdot (u^0 - u_t) \ge -|\mu_t|\, |u^0 - u_t|, \tag{1.3.82}$$

from which (1.3.80) follows. ■

Remark 1.32 The λ constructed above has no reason to be measurable, but if it is, it belongs to $L^1(0,T,\mathbb{R}^{n_g})$. Indeed, both μ and $(u^0 - u_t)$ belong to $\mathcal{U}$, and we conclude with (1.3.80). 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

$$\mathcal{T}_J := \{t \in (0,T);\ I_t(u_t) = J\}. \tag{1.3.83}$$

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

$$\operatorname*{Min}_{\lambda \in \mathbb{R}^{n_g}_+} |\lambda|; \quad \eta_t := \sum_{i \in J} \lambda_i \gamma_i; \quad \lambda_i = 0, \ \ i \notin J. \tag{1.3.84}$$

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

$$\lambda_t := \sum_{J \subset \{1,\dots,n_g\}} \varphi_J(\mathbf{1}_{t \in \mathcal{T}_J} \mu_t, Dg(u_t)). \tag{1.3.85}$$

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

Denote the set of *Lagrange multipliers* by

$$\Lambda(u) := \left\{ \lambda \in L^2(0,T;\mathbb{R}^{n_g});\ \lambda_t \in N_{\mathbb{R}^{n_g}_-}(g(u_t))\ a.e.;\ Df(u)_t + \sum_{i=1}^{n_g} \lambda_{i,t} Dg_i(u_t) = 0 \right\}. \tag{1.3.86}$$

Lemma 1.33 *The point $\bar{u}$ satisfies the first-order necessary optimality conditions* (1.3.66) *iff $\Lambda(\bar{u})$ is not empty. If in addition R_t is uniformly positive, then $\Lambda(\bar{u})$ 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. By lemma 1.29, $\bar{u}$ is esentially bounded; so is $Df(\bar{u}) = -\mu$. Combining with (1.3.80), we deduce that $\Lambda(\bar{u})$ is a bounded subset of $L^\infty(0,T,\mathbb{R}^{n_g})$. We next show that $\Lambda(u)$ is weakly$*$ closed. A half-space of the form

$$H_{\psi,\beta} := \{\gamma \in L^\infty(0,T,\mathbb{R}^{n_g});\ \int_0^T \gamma_t \cdot \psi_t \mathrm{d}t \le \beta\} \tag{1.3.87}$$

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

$$\int_0^T [Df(\bar{u})_t + \sum_{i=1}^{n_g} \lambda_{i,t} Dg_i(\bar{u}_t)\psi_{i,t}]\mathrm{d}t = 0, \quad \text{for all } \psi \in L^1(0,T,\mathbb{R}^m). \tag{1.3.88}$$

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

$$I_t^+(u) := \{1 \le i \le n_g;\ \lambda_{i,t} > 0, \ \text{ for some } \lambda \in \Lambda(u)\}. \tag{1.3.89}$$

The critical cone is as follows:

$$C(u) := \{v \in T_K(u);\ Dg_i(u_t)v_t = 0,\ i \in I_t^+(u), \ \text{ a.a. } t\}. \tag{1.3.90}$$

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

$$I_t^\varepsilon(u) := \{1 \le i \le n_g;\ -\varepsilon \le g_i(u_t) < 0\}. \tag{1.3.91}$$

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

$$C_\varepsilon(\bar{u}) := \{v \in C(\bar{u});\ \|v\|_\infty \le 1/\varepsilon; v_t = 0 \ \text{ if } I_t^\varepsilon(\bar{u}) \ne \emptyset, \text{ for a.a. } t\}. \tag{1.3.92}$$

Lemma 1.34 *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

$$v_t^{1,\varepsilon} := \max(-1/\varepsilon, \min(1/\varepsilon, v_t)), \quad \text{for all} \quad t \in (0,T), \tag{1.3.93}$$

and v^ε be defined by

$$v_t^\varepsilon = \begin{cases} 0 & \text{if } I_t^\varepsilon(\bar{u}) \neq \emptyset \\ v_t^{1,\varepsilon} & \text{if not} \end{cases} \tag{1.3.94}$$

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. ■

Define $J_t^\varepsilon(\bar{u}) := I_t(\bar{u}) \cup I_t^\varepsilon(\bar{u})$. 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 1.35 *Given $\varepsilon > 0$ and $v \in C_\varepsilon(\bar{u})$, let $w \in L^\infty(0,T;\mathbb{R}^m)$ be such that*

$$Dg_i(\bar{u}_t)w + D^2 g_i(\bar{u}_t)(v_t, v_t) \leq -\varepsilon, \ i \in J_t^\varepsilon(\bar{u}). \tag{1.3.95}$$

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

$$u^\theta := \bar{u} + \theta v + \tfrac{1}{2}\theta^2 w. \tag{1.3.96}$$

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

$$\Lambda_\varepsilon(u) := \left\{ \lambda \in L^\infty(0,T,\mathbb{R}^{n_g})_+; \ \lambda_{i,t} = 0, \, t \notin J_t^\varepsilon(\bar{u}); \ Df(u)_t + \sum_{i \in J_t^\varepsilon(\bar{u})} \lambda_{i,t} Dg_i(u_t) = 0 \right\}. \tag{1.3.97}$$

The qualification condition (1.3.74) 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 $L : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$,

$$L(u,\lambda) = f(u) + \int_0^T \sum_{i=1}^{n_g} \lambda_{i,t} g_i(u_t) \mathrm{d}t. \tag{1.3.98}$$

Theorem 1.36 *Let $\bar{u}$ be a local solution of (P). Then for any critical direction v, there exists a multiplier $\lambda \in \Lambda_\varepsilon(\bar{u})$ such that*

$$D^2_{uu} L(\bar{u},\lambda)(v,v) \geq 0. \tag{1.3.99}$$

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

$$\begin{array}{l} \underset{w \in \mathcal{U}}{\text{Min}} \, Df(\bar{u})w + D^2 f(\bar{u})(v,v); \\ Dg_i(\bar{u}_t)w + D^2 g_i(\bar{u}_t)(v_t,v_t) \leq -\varepsilon, \quad i \in J_t^\varepsilon(\bar{u}), \ \text{ a.e.} \end{array} \tag{SP_ε}$$

We choose $L^2(0,T,\mathbb{R}^{n_g})$ as constraint space. By lemma 1.35, for any feasible w in $F(SP_\varepsilon)\cap L^\infty(0,T;\mathbb{R}^m)$, the path u^θ defined in (1.3.96) 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

$$0\le \lim_{\theta\downarrow 0}\frac{f(u^\theta)-f(\bar u)}{\frac12\theta^2}=Df(\bar u)w+D^2f(\bar u)(v,v). \tag{1.3.100}$$

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

$$\begin{array}{l}\underset{w\in\mathcal{U}}{\mathrm{Min}}\displaystyle\int_0^T|w_t-w_t^\gamma|^2\mathrm{d}t;\quad \|w^\gamma\|_\infty\le 1/\gamma;\\ Dg_i(\bar u_t)w+D^2g_i(\bar u_t)(v_t,v_t)\le-\varepsilon,\ i\in J_t^\varepsilon(\bar u).\end{array} \tag{1.3.101}$$

Let us show that for $\varepsilon<\frac12\beta$ and small enough γ, this problem is feasible. Denote $\hat w:=(u^0-\bar u_t)$. Then $\hat w\in L^\infty(0,T,\mathbb{R}^n)$, and

$$Dg_i(\bar u_t)\hat w_t\le g_i(u^0)-g_i(\bar u_t)\le-\beta+\varepsilon<-\tfrac12\beta,\ \ i\in J_t^\varepsilon(\bar u),\quad \text{a.a.}\quad t\in(0,T). \tag{1.3.102}$$

Since $\bar u$ is essentially bounded, this proves that the linear constraints in (1.3.101) are satisfied by $w=c\hat w$, with $c=O(\|v\|_\infty^2+\varepsilon)$. Therefore if $1/\gamma\ge c\|\hat w\|_\infty$, (1.3.101) is feasible.

Now $w_t^\gamma=w_t$ if $|w_t|\le 1/\gamma$, and $|w_t^\gamma|\le|w_t|$ a.e.; it follows that when $\gamma\downarrow 0$, $w^\gamma\to w$ in $\mathcal{U}$. Passing to the limit in (1.3.100) (in which w is w^γ) we obtain that

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

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

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

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

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

where $\eta\in L^2(0,T,\mathbb{R}^{n_g})$, is feasible; indeed, using $\hat w$ satisfying (1.3.102), it suffices to take w of the form

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

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 $\mathrm{val}(SD_\varepsilon)\ge 0$.

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 1.33.

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 $D^2f(\bar u)(v^\varepsilon,v^\varepsilon)\to D^2f(\bar u)(v,v)$ and $D^2g(\bar u)(v^\varepsilon,v^\varepsilon)\to D^2g(\bar u)(v,v)$ in $L^1(0,T,\mathbb{R}^{n_g})$. For each $\varepsilon>0$ there exists $\lambda^\varepsilon\in\Lambda_\varepsilon(\bar u)$ such that $D^2_{uu}L(\bar u,\lambda^\varepsilon)(v^\varepsilon,v^\varepsilon)+\varepsilon\|\lambda^\varepsilon\|_{L^1}\ge 0$. Given

$\varepsilon_0 > 0$, λ^ε belongs to $\Lambda_{\varepsilon_0}(\bar{u})$ when $\varepsilon < \varepsilon_0$. Since $\Lambda_{\varepsilon_0}(\bar{u})$ is a weakly$*$ compact subset of $L^\infty(0,T,\mathbb{R}^{n_g})$, 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 $\bar{\lambda}$ in the weak$*$ topology of $L^\infty(0,T,\mathbb{R}^{n_g})$. Denoting by v^k the corresponding sequence extracted from v^ε, we obtain

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

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

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

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 [28] and Mignot [32]. Our presentation in section 1.2 follows [7]. Various extensions are presented in [13]. 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 [6].

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

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

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

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 [23]. 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

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

In other words, $H_u(u(t), y(t), p(t))$ 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*

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

Remark 2.3 The above relations are reminiscent of classical *Hamiltonian systems*, introduced by Hamilton in [27]. 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. In that case we recover an "ordinary" Hamiltonian system, whose new Hamiltonian is obtained by substituting the expression of the control (as a function of state and costate).

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 [38, 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:

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

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 u 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

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

(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:

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

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 Pontryagin's principle is satisfied by a trajectory, the strong Legendre-Clebsch condition holds, and in addition $H(., \bar{y}(t), \bar{p}(t))$ is quasi-convex (i.e., has convex sublevel 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

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

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 $((x))^2 \equiv (x,x)$:

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

Therefore:

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

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

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 $v \to \sqrt{D^2_{((u,y))^2}\mathcal{L}(u, y_u, p_u)((v,z))^2}$ is a norme equivalent 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 will 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

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

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 $\|\cdot\|_\infty \le C\|\cdot\|_2$ holds.

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

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

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 $O(\cdot) \le c|\cdot|$ uniformly whenever $\|u\|_\infty \le M$ and say $\|v\|_\infty \le 1$, 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

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

This proves that

$$y_{u+v} = y_u + z + z_{v,v} + r_{v,v} \tag{2.1.44}$$

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

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

and

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

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

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

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

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

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:*

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

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

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

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, with continuous and dense inclusion, 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

$$r(u,v) := J(u+v) - J(u) - DJ(u)v - \tfrac{1}{2}D^2J(u)(v,v).$$

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

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

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

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

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 [13, 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*

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

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 1.28 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 [18] (our formulation is slightly different, but equivalent):

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

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

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

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

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

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 $v^\varepsilon \in \mathcal{U}$

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

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:*

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

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 [19], and Malanowski [30]. Related results can be found in Pales and Zeidan [35, 36]. 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 [31].

Chapter 3

Discretization analysis

3.1 Setting

3.1.1 Framework

We consider the following unconstrained optimal control problem:

$$\text{Min } \Phi(y(T)); \quad \dot{y}(t) = f(u(t), y(t)), \;\; t \in [0,T]; \quad y(0) = y^0, \tag{P}$$

where for simplicity we assumed that there is no integral cost; we remind that it is always possible to reduce to this case by adding a scalar state variable. The mappings $f : \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R}^n$ and $\Phi : \mathbb{R}^n \to \mathbb{R}$ are assumed to be of class C^∞. The first order necessary optimality conditions of this problem are:

$$\left\{ \begin{array}{rcl} \dot{y}(t) & = & f(u(t), y(t)), \\ \dot{p}(t) & = & -H_y(u(t), y(t), p(t)), \\ 0 & = & H_u(u(t), y(t), p(t)), \\ p(T) = \Phi'(y(T)), & & y(0) = y^0. \end{array} \right\} \;\; t \in [0,T], \tag{OC}$$

In the sequel we restrict the analysis to the case of continuous control variables. So we say that $(\bar{u}, \bar{y}, \bar{p})$ is an extremal if $\bar{u}$ is a continuous function of time, and $(\bar{u}, \bar{y}, \bar{p})$ satisfies (OC) ($\bar{u}$ being a continuous function). Let $(\bar{u}, \bar{y}, \bar{p})$ be an extremal. If

$$u \mapsto H_{uu}(u, y, p) \quad \text{is invertible along the trajectory,} \tag{3.1.1}$$

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.

Remark 3.1 Actually the reduction of the control variable to a function of state and costate is of local nature, and hence, is valid only for sufficiently small intervals of time. Yet we keep the above notation for simplicity.

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 the link between partial derivatives of the "ordinary" and true Hamiltonian:

$$\mathcal{H}_y(y,p) = H_y(\phi(y,p), y, p); \quad \mathcal{H}_p(y,p) = H_p(\phi(y,p), y, p). \tag{3.1.2}$$

Consequently, under hypothesis (3.1.1), (OC) is locally equivalent to the *reduced Hamiltonian system*

$$\begin{array}{rclrcl} \dot{y}(t) & = & \mathcal{H}_p(y(t),p(t)), & -\dot{p}(t) & = & \mathcal{H}_y(y(t),p(t)), \quad t\in[0,T], \\ p(T) & = & \Phi'(y(T)), & y(0) & = & y^0. \end{array} \tag{3.1.3}$$

In the sequel we discuss the discretization of the above problem, starting by Euler's method.

3.1.2 The simplest possible discretization: Euler's method

$$\left\{ \begin{array}{l} \text{Min } \Phi(y_N); \\ \quad y_{k+1} = y_k + h_k f(u_k,y_k), \quad k=0,\dots,N-1 \\ \quad y_0 = y^0, \end{array} \right. \tag{ED_1}$$

Here $h_k>0$ is the k step size; the discretized times are

$$t_k := \sum_{i=0}^{k-1} h_i, \quad k=0,\dots,N-1, \tag{3.1.4}$$

with $t_0=0$. We may consider that the u_k represent a piecewise constant control, having value u_k on the time interval (t_k,t_{k+1}). Then y_k is an approximation of the associated state using Euler's discretization.

Let us write the Lagrangian function associated with problem (ED_1) under the form

$$\Phi(y_N) + p^0\cdot(y^0-y_0) + \sum_{k=0}^{N-1} p_{k+1}\cdot(y_k+h_k f(u_k,y_k)-y_{k+1})\,. \tag{3.1.5}$$

We obtain the optimality conditions :

$$\begin{array}{rcll} p_N & = & \Phi'(y_N), & \\ p_1 & = & p^0 - h_0 f_y(u_k,y_k)^\top p_1, & \\ p_{k+1} & = & p_k - h_k f_y(u_k,y_k)^\top p_{k+1}, & k=1,\dots,N-1, \\ 0 & = & f_u(u_k,y_k)^\top p_{k+1}, & k=0,\dots,N-1. \end{array}$$

If the algebraic constraints

$$H_u(u_k,y_k,p_{k+1}) = f_u(u_k,y_k)^\top p_{k+1} = 0 \tag{3.1.6}$$

are locally equivalent to $u_k=\phi(y_k,p_k)$, then we see that we obtain the equivalent system with two end conditions:

$$\begin{array}{rcll} y_{k+1} & = & y_k + h_k \mathcal{H}_p(y_k,p_{k+1}), & k=1,\dots,N-1, \\ p_{k+1} & = & p_k - h_k \mathcal{H}_y(y_k,p_{k+1}), & k=1,\dots,N-1, \\ y_0 & = & y^0, \quad p_N=\Phi'(y_N). & \end{array}$$

This of course can be interpreted as a one step discretization scheme for the reduced system (3.1.2), with a kind of composite Euler scheme that is explicit in y and implicit in p. It is easily seen that the local error is of local order two, as for Euler's method.

3.2 Global and local errors

Consider an ordinary differential equation

$$\dot{y}(t) = f(y(t)), \quad t \in (0, T), \tag{3.2.7}$$

to be solved by a one-step method:

$$y_{k+1} = y_k + h_k \Psi(y_k), \quad k = 1, \ldots, N-1, \tag{3.2.8}$$

where Ψ is a smooth mapping. The *variational equation* for the original and discretized system are

$$\dot{Z}(t) = Df(y(t))Z(t), \quad t \in (0, T), \quad Z(0) = I, \tag{3.2.9}$$

$$Z_{k+1} = Z_k + h_k D\Psi(y_k) Z_k, \quad k = 0, \ldots, N-1, \quad Z_0 = I. \tag{3.2.10}$$

Here I denotes the identity matrix of size n, and $Z(t)$, Z_k are square matrices of size n. Note that this framework includes all Rung-Kutta schemes, and in particular, implicit ones (even for explicit Rung-Kutta schemes, the mapping Ψ is not explicit). It is well known that, if the local error is $q+1$, then the global error is, for the Cauchy problem, of order q. We show next that the same result holds for the TPBVB, provided the latter is well-posed.

Lemma 3.2 *Consider the problem of solving* (3.2.7) *subject to the end-point conditions*

$$F(y(0), y(T)) = 0, \tag{3.2.11}$$

where F is smooth $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$. Let $\phi_f(t, y_0)$ denote the flow associated with the vector field f. If the shooting equation

$$\Psi(y_0) := F(y_0, \phi_f(T, y_0)) = 0 \tag{3.2.12}$$

has an invertible Jacobian, and the scheme as well as the variational equation have local error order $q+1$, then the global error order is q.

Proof. Let $\phi_f^h(T, y_0))$ denote the numerical flow. The solution y^h of the scheme with discretization h satisfies

$$\mathcal{F}(\phi^h, y_0) := F[y_0^h, \phi_f^h(T, y_0^h)] = 0. \tag{3.2.13}$$

We apply the implicit function theorem, in the form $\Omega = \overline{B(y_0, 1)}$, W the Banach space of C^1 functions $\Omega \to \mathbb{R}^n$, and $\mathcal{F} : W \times \mathbb{R}^n \to \mathbb{R}^n$. Obviously $\mathcal{F}$ is of class C^1. Since the shooting equation is well-posed, $\partial \mathcal{F}(\phi^h, y_0)/\partial y_0$ is invertible, so that the hypotheses of the implicit function theorem are satisfied. The conclusion follows. ■

3.3 Runge-Kutta schemes

We now consider a Runge-Kutta (RK) discretization scheme of the state equation, of the following form:

$$\left\{\begin{array}{rcl} \text{Min } \Phi(y_N); & & \\ y_{k+1} & = & y_k + h_k \sum_{i=1}^s b_i f(u_{ki}, y_{ki}), \\ y_{ki} & = & y_k + h_k \sum_{j=1}^s a_{ij} f(u_{kj}, y_{kj}), \\ y_0 & = & y^0, \end{array}\right. \tag{RKD}$$

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.

Let us rewrite (DP_1) under the equivalent form

$$\text{Min } \Phi(y_N); \quad \left\{\begin{array}{l} 0 = h_k \displaystyle\sum_{i=1}^s b_i K_{ki} + y_k - y_{k+1}, \\ 0 = f(u_{ki}, y_k + h_k \sum_{j=1}^s a_{ij} K_{kj}) - K_{ki}, \\ 0 = y^0 - y_0, \end{array}\right. \tag{DP2}$$

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:

$$\begin{array}{c} \Phi(y_N) + p^0 \cdot (y^0 - y_0) \\ + \displaystyle\sum_{k=0}^{N-1} \left\{ p_{k+1} \cdot \left(h_k \sum_{i=1}^s b_i K_{ki} + y_k - y_{k+1} \right) + \sum_{i=1}^s \xi_{ki} \cdot (f(u_{ki}, y_{ki}) - K_{ki}) \right\}. \end{array}$$

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 (remember the definition of y_{ki}) :

$$\begin{array}{rcl} p_N & = & \Phi'(y_N), \quad p_1 = p^0, \\ p_k - p_{k+1} & = & \sum_{i=1}^s f_y(u_{ki}, y_{ki})^\top \xi_{ki}, \\ 0 & = & h_k b_i p_{k+1} + h_k \displaystyle\sum_{j=1}^s a_{ji} f_y(u_{kj}, y_{kj})^\top \xi_{kj} - \xi_{ki}, \\ 0 & = & f_u(u_{ki}, y_{kj})^\top \xi_{ki}, \quad k = 0 \ldots N-1, \quad i = 1 \ldots s. \end{array}$$

Using now the *restrictive 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

$$\left\{\begin{array}{rcl} y_{k+1} & = & y_k + h_k \sum_{i=1}^s b_i f(u_{ki}, y_{ki}), \\ y_{ki} & = & y_k + h_k \sum_{j=1}^s a_{ij} f(u_{kj}, y_{kj}), \\ p_{k+1} & = & p_k - h_k \sum_{i=1}^s \hat{b}_i H_y(u_{ki}, y_{ki}, p_{ki}), \\ p_{ki} & = & p_k - h_k \sum_{j=1}^s \hat{a}_{ij} H_y(u_{kj}, y_{kj}, p_{kj}), \\ 0 & = & H_u(u_{ki}, y_{ki}, p_{ki}), \\ y_0 & = & y^0, \qquad p_N \;=\; \Phi'(y_N), \end{array}\right. \tag{DOC}$$

where coefficients $\hat{b}$ and $\hat{a}$ are defined by the following relations:

$$\hat{b}_i := b_i, \quad \hat{a}_{ij} := b_j - \frac{b_j}{b_i} a_{ji}, \quad i = 1, \ldots, s \;\; j = 1, \ldots, s. \tag{3.3.14}$$

If the algebraic constraints $H_u(u_{ki}, y_{ki}, p_{ki}) = 0$ are locally equivalent to relations of the form $u_{ki} = \phi(y_{ki}, p_{ki})$, then (DOC) is equivalent to the same PRK scheme applied to the reduced system (3.1.3).

The following diagram commutes, when we use the above discretization:

$$\begin{array}{ccc} (P) & \xrightarrow{\text{discretization}} & (DP) \\ \text{optimality conditions} \downarrow & & \text{optimality conditions} \downarrow \\ (OC) & \xrightarrow{\text{discretization}} & (DOC) \end{array} \qquad (D)$$

We will in section 3.3.4 give the definition of a symplectic (continuous or discrete) flow. It is said that a PRK scheme (or more generally any one step scheme) is symplectic if the corresponding flow is symplectic. We will see that PRK schemes satisfying (3.3.14) are symplectic, and exploit this property in order to study the error orders. Before that, we recall some basic tools for studying the error order of a RK scheme.

3.3.1 Trees and B-series

Consider the solution of an ordinary differential equation (ODE)

$$\dot{y}(t) = f(y(t)), \quad t \geq 0; \quad y(0) = y_0. \tag{3.3.15}$$

It is known that the global error order of a one-step scheme is nothing but the maximum order up to which the Taylor expansions of the ODE coincides with the one of the scheme. So we need to manipulate these Taylor expansions at any order. It is not difficult to compute the first derivatives of $y(t)$ at time 0. We adopt here the notation x for time, since t will denote trees, f', f'', etc, for the derivatives of f, and use e.g; $f'f(y_0)$ for $f'(f(y_0))$:

$$\begin{array}{ll} \dot{y}(0) & = f(y_0); \quad \ddot{y}(t) = f'f(y_0), \\ y^{(3)}(0) & = f''(f,f)(y_0) + f'f'f(y_0), \\ y^{(4)}(0) & = f'''(f,f,f)(y_0) + 3f''(f'f,f)(y_0) + f'f''(f,f)(y_0) + f'f'f'f(y_0). \end{array} \tag{3.3.16}$$

We see by induction that for any integer k, the expression of $y^{(k)}(0)$ is a linear combination with positive weights of "elementary differentials" which are compositions of $f^{(i)}$, for $i = 0$ to k, in which the symbol f appears k times, and $f^{(i)}$ has i arguments. Let us, with this type of expression, associate rooted trees, i.e., undirected graphs without cycles, with one "special" node called the root. Taking all possible paths from the root to the leaves (nodes other than the root having only one neighbor) induces a directions over edges, so we can see a rooted tree as a directed tree. Each node has one "parent", and none, one or several "sons".

With such a tree, associate the expression obtained by replacing each node with i "sons" by $f^{(i)}$, whose arguments are the i "sons". A more formal definition follows.

Definition 3.3 (Rooted trees) The set T of rooted trees is generated as follows: the tree $\tau = \bullet$ with only one vertex belongs to T, and given $t_1, \ldots, t_m$ in T, the graph obtained by connecting their roots to a new vertex, which will be the root of the resulting tree, belongs to T. This operation is denoted as $t = [t_1, \ldots, t_m]$.

Note that some of the trees among $t_1, \dots, t_m$ may be identical and that $t = [t_1, \dots, t_m]$ does not depend on the order of $t_1, \dots, t_m$.

We now show how to associate with a rooted tree a elementary differential expression as well as some integer coefficients.

Definition 3.4 (Elementary differentials) (i) The elementary differential of a rooted tree is an expression defined inductively by $F(\bullet) = f(y)$ and

$$F([t_1, \dots, t_m])(y) = f^{(m)}(y)(F(t_1)(y), \dots, F(t_m)(y)). \tag{3.3.17}$$

(ii) The order $\rho(t)$ of a tree is the number of its vertices.
(iii) The Taylor coefficient $\alpha(t)$ is defined inductively by $\alpha(\bullet) = 1$ and

$$\alpha([t_1, \dots, t_m]) = \begin{pmatrix} \rho(t) - 1 \\ \rho(t_1), \dots, \rho(t_m) \end{pmatrix} \frac{\alpha(t_1) \times \cdots \times \alpha(t_m)}{\mu_1! \mu_2! \dots} \tag{3.3.18}$$

where the integers μ_i count equal trees among $t_1, \dots, t_m$.
We remind that $\begin{pmatrix} n \\ n_1, \dots, n_m \end{pmatrix} := n!/(n_1! \cdots n_m!)$.

We will see in theorem 3.7 that $\alpha(t)$ is the integer appearing when computing the derivatives of $y(t)$ as in (3.3.16).

Since the expansion of the solution of a RK scheme acn be expressed using elementary differentials, but with different coefficients, we introduce formal power series[1] of a time step h.

Definition 3.5 (B-series) Given a "weight" mapping $w : T \cup \{\emptyset\} \to \mathbb{R}$, we call B-series a formal series of the form

$$B(w, y) = w(\emptyset) + \sum_{t \in T} \frac{h^{\rho(t)}}{\rho(t)!} \alpha(t) w(t) F(t)(y) \tag{3.3.19}$$

We need to compute the (formal) image of a B-series by f.

Lemma 3.6 *Let $w : T \cup \{\emptyset\} \to \mathbb{R}$ satisfy $w(\emptyset) = 1$. Then*

$$hf\left(B(w, y)\right) = B(w', y), \tag{3.3.20}$$

where $w'(\emptyset) = 0$, $w'(\bullet) = 1$, and for all $t \in T$ other than $\bullet$,

$$w'([t_1, \dots, t_m]) = \rho(t) w(t_1) \times \cdots \times w(t_n). \tag{3.3.21}$$

Proof. That $w(\emptyset) = 1$ implies that (formally) $B(w, y) = y + O(h)$. Therefore we can compute an expansion of $f\left(B(w, y)\right)$ around y. We have, denoting $t = [t_1, \dots, t_m]$, and

[1] A formal power series $\sum_{m \in \mathbb{N}} c_m h^m$ is noting but the sequence c_m. The set of formal power series is endowed with the operations of addition, multiplication, and composition, following the rules for usual power series.

the integers μ_i counting equal trees among $t_1, \ldots, t_m$:

$$\begin{aligned} hf\left(B(w,y)\right) &= h\sum_{m\geq 0}\frac{1}{m!}f^m\left(B(w,y)-y\right)^m \\ &= h\sum_{m\geq 0}\frac{1}{m!}\sum_{t_1,\ldots,t_m\in T}\prod_{i=1}^{m}\frac{h^{\rho(t_i)}\alpha(t_i)w(t_i)}{\rho(t_i)!}f^m(y)(F(t_1)(y),\ldots,F(t_m)(y)) \\ &= \sum_{m\geq 0}\sum_{t_1,\ldots,t_m\in T}\frac{h^{\rho(t)}}{(\rho(t)-1)!}\alpha(t)\frac{\mu_1!\cdots\mu_m!}{m!}w(t_1)\cdots w(t_m)F(t)(y) \\ &= \sum_{t\in T}\frac{h^{\rho(t)}}{(\rho(t)-1)!}\alpha(t)w(t_1)\cdots w(t_m)F(t)(y) = B(w',y). \end{aligned}$$

In the fourth equality we use the fact that there are $\left(\begin{array}{c} m \\ \mu_1,\ldots,\mu_m \end{array}\right)$ possibilities for writing t in the form $[t_1,\ldots,t_m]$. ■

Let us compute the coefficients of the B-series expansion of the solution of the differential equation.

Theorem 3.7 *The solution of the differential equation $\dot{y}(x) = f(y(x))$ can be expressed as a B-series: $y(x_0+h) = B(e,y_0)$, where $y_0 = y(x_0)$, and*

$$e(t) = 1, \quad \textit{for all } \ t\in T\cup\{\emptyset\}. \tag{3.3.22}$$

Proof. We that have already established that $y(x_0+h)$ has a B-series expansion. Obviously $e(\emptyset) = 1$ and $e(\bullet) = 1$. By lemma 3.6, this series is solution of $h\dot{y}(x_0+h) = hf(y(x_0+h))$ iff

$$\sum_{t\in T}\frac{h^{\rho(t)}}{(\rho(t)-1)!}\alpha(t)e(t)F(t)(y_0) = \sum_{t\in T}\frac{h^{\rho(t)}}{\rho(t)!}\alpha(t)e'(t)F(t)(y_0). \tag{3.3.23}$$

This holds iff, for all $t = [t_1,\ldots,t_m]$ (of order greater than one)

$$\rho(t)e(t) = e'(t) = \rho(t)e(t_1)\cdots e(t_m). \tag{3.3.24}$$

By induction we deduce that (3.3.22) holds. ■

We now consider the expansion of the solution of a Runge-Kutta scheme.

Theorem 3.8 *The solution of a Runge-Kutta scheme can be expressed as a B-series $y_1 = B(w,y_0)$, where $w(\emptyset) = 1$ and*

$$w(t) = \gamma(t)\sum_{i=1}^{s}b_i\Phi_i(t), \quad \textit{for all } \ t\in T, \tag{3.3.25}$$

where the integer coefficient $\gamma(t)$ and the expression $\Phi_i(t)$ are defined inductively by

$$\gamma(\bullet) = 1; \quad \Phi_i(\bullet) = 1, \ \ i = 1,\ldots,s \tag{3.3.26}$$

and, for all $t_1, \ldots, t_m$ in T:

$$\gamma([t_1, \ldots, t_m]) = \rho(t)\gamma(t_1)\cdots\gamma(t_m), \tag{3.3.27}$$

$$\Phi_i([t_1, \ldots, t_m]) = \sum_{j_1,\ldots,j_m=1}^{s} a_{ij_1}\cdots a_{ij_m}\Phi_i(t_1)\cdots\Phi_i(t_m). \tag{3.3.28}$$

Proof. The amount

$$hK_i = hf(y_0 + h\sum_{j=1}^{s} a_{ij}K_j)$$

has, for $i = 1$ to s, a B-series expansion that we denote Ψ_i. So we have

$$B(\Psi_i, y_0) = hK_i = hf(y_0 + \sum_{j=1}^{s} a_{ij}B(\Psi_j, y_0)) = hf(B(\sum_{j=1}^{s} a_{ij}\Psi_j, y_0)), \tag{3.3.29}$$

where we have set

$$\left(\sum_{j=1}^{s} a_{ij}\Psi_j\right)(\emptyset) = 1; \quad \left(\sum_{j=1}^{s} a_{ij}\Psi_j\right)(t) = \sum_{j=1}^{s} a_{ij}\Psi_j(t), \quad \text{for all } t \in T. \tag{3.3.30}$$

By lemma 3.6, we have that

$$\Psi_i(t) = \left(\sum_{j=1}^{s} a_{ij}\Psi_j\right)'(t) = \rho(t)\prod_{k=1}^{m}\left(\sum_{j=1}^{s} a_{ij}\Psi_j\right)(t_k), \tag{3.3.31}$$

or equivalently

$$\Psi_i(t) = \rho(t)\sum_{j_1,\ldots,j_m=1}^{s} a_{ij_1}\cdots a_{ij_m}\Psi_j. \tag{3.3.32}$$

It follows then from the definition of $\gamma(t)$ that $\Psi_i(t) = \gamma(t)\Phi_i(t)$. Since

$$y_1 = y_0 + \sum_{i=1}^{s} b_i hK_i = B(\sum_{i=1}^{s} b_i\Psi_i, y_0), \tag{3.3.33}$$

(3.3.25) follows. ■

Corollary 3.9 *A Runge-Kutta scheme has order p (for all possible smooth dynamics f) iff*

$$\sum_{i=1}^{s} b_i\Phi_i = \frac{1}{\gamma(t)}, \qquad \textit{whenever } \rho(t) \leq p. \tag{3.3.34}$$

Proof. If (3.3.34) is satisfied, then the B-series expansions of the exact and approximate solution coincide up to order p, so that their difference is of local order $p + 1$, and we know that this implies that the global order is p. For the converse implication, see [26]. ■

3.3.2 Partitioned Runge-Kutta methods

We have noticed that the optimality systems obtained when discretizing by a Runge-Kutta method may be interpreted as partitioned Runge-Kutta methods. In this section we show how to compute the order of these methods.

We first consider an ordinary differential equation in partitioned form (omitting the time argument):

$$p' = f(p,q); \quad q' = g(p,q). \tag{3.3.35}$$

Let (a,b) and $(\hat{a},\hat{b})$ be the coefficients of two Runge-Kutta methods having the same number s of inner steps. The associated partitioned Runge-Kutta method for solving (3.3.35) is, for k in $\mathbb{N}$:

$$\begin{array}{rcl} p_{k+1} & = & p_k + h_k \sum_{i=1}^s b_i f(p_{ki}, q_{ki}), \\ q_{k+1} & = & q_k + h_k \sum_{i=1}^s \hat{b}_i g(p_{ki}, q_{ki}), \\ p_{ki} & = & p_k + h_k \sum_{j=1}^s a_{ij} f(p_{kj}, q_{kj}), i = 1, \ldots, s \\ q_{ki} & = & q_k + h_k \sum_{j=1}^s \hat{a}_{ij} g(p_{kj}, q_{kj}), i = 1, \ldots, s \\ p_0 & = & p^0, \quad q_0 = q^0, \end{array} \tag{3.3.36}$$

The first derivatives of the p component of the solution of the ODE are as follows (omitting the arguments (p_0, q_0)):

$$\begin{array}{rcl} \dot{p} & = & f \\ \ddot{p} & = & f_p f + f_q g \\ p^{(3)} & = & f_{pp}(f,f) + 2f_{pq}(f,g) + f_{qq}(g,g) \\ & & + f_p f_p f + f_p f_q g + f_q g_p f + f_q g_q g. \end{array} \tag{3.3.37}$$

The expressions holds for the derivative of q are obtained by exchanging f and g. We see again that these derivatives are linear combinations with positive weights of elementary differentials, the novelty being that in the latter we have to distinguish between the symbols f and g. This leads to the introduction of rooted bi-colored trees. The vertices of the latter are either black or white, corresponding to f and g respectively. The set TP of rooted bi-colored trees is constructed as follows: it containts the two graphs

$$\tau_p = \bullet, \quad \tau_q = \circ, \tag{3.3.38}$$

and is closed under the operations

$$[t_1, \ldots, t_m]_p \quad \text{and} \quad [t_1, \ldots, t_m]_q, \tag{3.3.39}$$

of connecting the rooted bi-colored trees $t_1, \ldots, t_m$ to a black or white root. Those with a black (white) root are denoted TP_p (TP_q) and correspond to the derivatives of p (resp. q). The elementary differentials are constructed accordingly. Again for $t \in TP$, its order $\rho(t)$ is the number of vertices of t, and its coefficient $\alpha(t)$ is defined inductively by $\alpha(\tau_p) = \alpha(\tau_q) = 1$ and for $t = [t_1, \ldots, t_m]_p$ or $t = [t_1, \ldots, t_m]_q$,

$$\alpha(t) = \left(\begin{array}{c} \rho(t) - 1 \\ \rho(t_1), \ldots, \rho(t_m) \end{array} \right) \frac{\alpha(t_1) \times \cdots \times \alpha(t_m)}{\mu_1! \mu_2! \ldots} \tag{3.3.40}$$

where the integers μ_i count equal trees among $t_1, \ldots, t_m$. Although the formula is formally identical to the one of definition 3.4, the value of α depends on the coloring of the vertices. For instance,

$$\alpha([\tau_p, \tau_p]_p) = 1; \quad \alpha([\tau_p, \tau_q]_p) = 2. \tag{3.3.41}$$

We then define P-series (partitioned series) as series of the form

$$P(w,(p,q)) = \begin{pmatrix} P_p(w,(p,q)) \\ P_q(w,(p,q)) \end{pmatrix} = \begin{pmatrix} a(\emptyset)_p + \sum_{i \in TP_p} \frac{h^{\rho(t)}}{\rho(t)!} \alpha(t) w(t) F(t)(p,q) \\ a(\emptyset)_q + \sum_{i \in TP_q} \frac{h^{\rho(t)}}{\rho(t)!} \alpha(t) w(t) F(t)(p,q) \end{pmatrix} \tag{3.3.42}$$

The proof of the next results being similar to the non partitioned case, we omit them.

Lemma 3.10 *Let $w : TP \cup \{\emptyset_p, \emptyset_q\} \to \mathbb{R}$ satisfy $w(\emptyset_p) = w(\emptyset_q) = 1$. Then*

$$h \begin{pmatrix} f(P(w,(p,q)) \\ g(P(w,(p,q)) \end{pmatrix} = P(w',(p,q)), \tag{3.3.43}$$

where $w'(\emptyset_p) = w'(\emptyset_q) = 0$, $w'(\tau_p) = w'(\tau_q) = 1$, and for all $t = [t_1, \ldots, t_m]_p$ or $t = [t_1, \ldots, t_m]_q$:

$$w'(t) = \rho(t) a(t_1) \cdots a(t_m). \tag{3.3.44}$$

Theorem 3.11 *The solution of the ODE* (3.3.35) *has the P-series expansion*

$$(p(x_0 + h), q(x_0 + h)) = P(e,(p_0,q_0)),$$

where

$$e((\emptyset_p) = e(\emptyset_q) = 1; \quad e(t) = 1, \quad \text{for all } t \in TP. \tag{3.3.45}$$

Theorem 3.12 *The solution of the partitioned Runge-Kutta scheme* (3.3.36) *has the P-series expansion*

$$(p_1, q_1) = P(w,(p_0,q_0)),$$

where $w((\emptyset_p) = w(\emptyset_q) = 1$ and

$$w(t) = \begin{cases} \gamma(t) \sum_{i=1}^s b_i \Phi_i(t), & \text{for } t \in TP_p, \\ \gamma(t) \sum_{i=1}^s \hat{b}_i \Phi_i(t), & \text{for } t \in TP_q. \end{cases} \tag{3.3.46}$$

Here the coefficient $\gamma(t)$, defined as in (3.3.26)*, does not depend on the coloring of t, and $\Phi_i(t)$ is defined inductively by $\Phi_i(\tau_p) = \Phi_i(\tau_q) = 1$, and for $t = [t_1, \ldots, t_m]_p$ or $t = [t_1, \ldots, t_m]_q$:*

$$\Phi_i(t) = \Psi_i(t_1) \cdots \Psi_i(t_m), \qquad \text{with} \qquad \Psi_i(t_k) = \begin{cases} \sum_{j_k=1}^s a_{ij_k} \Phi_{j_k}(t_k) & \text{if } t \in TP_p \\ \sum_{j_k=1}^s \hat{a}_{ij_k} \Phi_{j_k}(t_k) & \text{if } t \in TP_q \end{cases} \tag{3.3.47}$$

Comparing the P-series expansions of the solution of the ODE and of the partitioned Runge-Kutta scheme, we obtain the order conditions.

Theorem 3.13 *The partitioned Runge-Kutta scheme* (3.3.36) *has (global) order r iff*

$$\begin{aligned} \sum_{i=1}^s b_i \Phi_i(t) &= \frac{1}{\gamma(t)}, \quad \text{for all } t \in TP_p, \ \rho(t) \le r, \\ \sum_{i=1}^s \hat{b}_i \Phi_i(t) &= \frac{1}{\gamma(t)}, \quad \text{for all } t \in TP_q, \ \rho(t) \le r. \end{aligned} \tag{3.3.48}$$

3.3.3 Quadratic invariants

For the study of symplectic flows, we need the following result on the conservation of quadratic invariants by partitioned Runge-Kutta schemes. Note that the hypotheses on the coefficients are satisfied in the case of discretized optimality systems.

Proposition 3.14 *If the coefficients of a partitioned Runge-Kutta scheme satisfy*

$$b_i\hat{a}_{ij} + \hat{b}_j a_{ij} = b_i\hat{b}_j, \quad \textit{for all } i = j = 1, \ldots, s, \tag{3.3.49}$$

$$b_i = \hat{b}_i, \quad \textit{for all } i = 1, \ldots, s, \tag{3.3.50}$$

then it conserves quadratic invariants of the form

$$Q(p, q) := p \cdot Dq, \tag{3.3.51}$$

where D is an arbitrary matrix.

Proof. Set

$$K_i = f(p_{0i}, q_{0i}); \quad \hat{K}_i = g(p_{0i}, q_{0i}), \quad i = 1, \ldots, s. \tag{3.3.52}$$

In view of the expression (3.3.36) of the partitioned scheme, we have that

$$p_1 \cdot Dq_1 - p_0 \cdot Dq_0 = h \sum_{i=1}^{s} b_i\, K_i \cdot Dq_0 + h \sum_{j=1}^{s} \hat{b}_j\, p_0 \cdot D\hat{K}_j + h^2 \sum_{i,j=1}^{s} b_i\hat{b}_j\, K_i \cdot D\hat{K}_j. \tag{3.3.53}$$

Now eliminate p_0 and q_0 from the right hand side of the above equality, using the expressions

$$p_{0\ell} = p_0 + h \sum_{j=1}^{s} a_{\ell j} f(p_{0j}, q_{0j}); \quad q_{0\ell} = q_0 + h \sum_{j=1}^{s} \hat{a}_{\ell j} g(p_{0j}, q_{0j}), \quad \ell = 1, \ldots, s. \tag{3.3.54}$$

Taking ℓ equal to the index i or j of the sums in (3.3.53), obtain

$$\begin{aligned} p_1 \cdot Dq_1 - p_0 \cdot Dq_0 = \; & h \sum_{i=1}^{s} b_i\, K_i \cdot Dq_{0i} + h \sum_{j=1}^{s} \hat{b}_j\, p_{0j} \cdot D\hat{K}_j \\ & + h^2 \sum_{i,j=1}^{s} \left(b_i\hat{b}_j - b_i\hat{a}_{ij} - \hat{b}_j a_{ji} \right) K_i \cdot D\hat{K}_j. \end{aligned} \tag{3.3.55}$$

Since $p \cdot Dq$ is an invariant, $f(p, q) \cdot Dq + p \cdot Dg(p, q) = 0$, so that for all $i = 1, \ldots, s$, $K_i \cdot Dq_{0i} + p_{0i} \cdot D\hat{K}_i = 0$. Using (3.3.49)-(3.3.50), we conclude that the r.h.s. of (3.3.55) is zero, as was to be proved. ■

3.3.4 Symplectic transformations

Consider the $2n \times 2n$ matrix

$$J := \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}, \tag{3.3.56}$$

where I denotes the identity matrix in $\mathbb{R}^n$. Given $H(p,q):\mathbb{R}^{2n}\to\mathbb{R}$ be of class C^2, the associated Hamiltonian system[2]

$$\dot{p}=-H_q(p,q);\quad \dot{q}=H_p(p,q) \tag{3.3.57}$$

can be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} p\\ q\end{pmatrix}=J^{-1}DH(p,q), \tag{3.3.58}$$

and the variational equation may be written as

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} Z_p\\ Z_q\end{pmatrix}=J^{-1}D^2H(p,q)\begin{pmatrix} Z_p\\ Z_q\end{pmatrix}. \tag{3.3.59}$$

Definition 3.15 (i) A linear mapping $A:\mathbb{R}^{2n}\to\mathbb{R}^{2n}$ is called *symplectic* if it satisfies $A^\top JA=J$.
(ii) A differentiable function $\varphi:\mathbb{R}^{2n}\to\mathbb{R}^{2n}$ is called *symplectic* at $(p,q)\in\mathbb{R}^{2n}$,if the Jacobian matrix is symplectic, i.e., if

$$\varphi'(p,q)J\varphi'(p,q)=J. \tag{3.3.60}$$

We say that φ is symplectic if it is symplectic at all points.

The next theorem is due to Poincaré.

Theorem 3.16 *Let $H(p,q):\mathbb{R}^{2n}\to\mathbb{R}$ be of class C^2. Then the associated Hamiltonian flow is symplectic.*

Proof. Denote the flow by φ_t. The amount $\Psi_t:=\partial\varphi_t(y_0)/\partial y_0$ is solution of the variational equation, and hence, skipping the arguments of H:

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\Psi_t\cdot J\Psi_t\right)=\dot{\Psi}_t\cdot J\Psi_t+\Psi_t\cdot J\dot{\Psi}_t=\Psi_tD^2HJ^{-\top}J\Psi_t+\Psi_t\cdot JD^2H\Psi_t=0,$$

since $J^{-\top}=-J^{-1}$ is the identity matrix. Therefore $\Psi_t\cdot J\Psi_t$ is invariant along the trajectory, and hence, equal to its intial value J, as was to be proved. ■

Definition 3.17 A numerical one-step method $y_1=\Phi_h(y_0)$ is said to be symplectic if, when applied to a smooth Hamiltonian system, the mapping Φ_h is symplectic.

Theorem 3.18 *A partitioned Runge-Kutta scheme satisfying* (3.3.49)-(3.3.50) *is symplectic.*

Proof. Let Ψ denote the solution of the variational system, and Ψ_h denote the solution of linearized partitioned Runge-Kutta scheme with initial value identity, i.e., the variational system of the PRK scheme. We may interpret the scheme for (p,q,Ψ) as a PRK scheme with the same value, and $\Psi\cdot J\Psi$ is a quadratic invariant for the original ODE. So by proposition 3.14, we have that $\Psi\cdot J\Psi$ is also an invariant of the scheme. The conclusion follows. ■

[2]In the application to optimal control, p and q correspond to the costate and state, respectively.

3.3.5 Order conditions for symplectic PRK methods

We have presented in section 3.3.2 the theory of error orders for PRK methods. For order r we have as many conditions as distinct rooted trees of order up to r. Let us remind the reader that for the tree $t = [t_1, \ldots, t_m]$, the order of branches $t_1, \ldots, t_m$ does not matter. The expression of these conditions has been given in theorem 3.13. Since $\hat{b} = b$, they reduce to

$$\sum_{i=1}^{s} b_i \Phi_i(t) = \frac{1}{\gamma(t)}, \quad \text{for all } \; t \in TP_p \cup TP_q, \quad \rho(t) \leq r, \tag{3.3.61}$$

where the integer numbers $\Phi_i(t)$ follow from (3.3.47).

Let us now show that the number of conditions can be reduced, using the special form of symplectic PRK methods.

Definition 3.19 (i) Oriented free trees, i.e., trees with oriented edges and without root, are denoted H-trees. They are defined by a pair (V, E), the sets of vertices and edges resp., such that $\#E = \#V - 1$.
(ii) Bi-colored graphs are endowed with a mapping $c(\cdot)$ that with each vertex v associates a colour $c(v)$, of value B or W (black or 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.

Given a H-tree (V, E), let us identify its set of vertices to $\{1, \ldots, \#V\}$. We denote

$$\tilde{a}_{i_k i_\ell} \;\; = \;\; a_{ij} \;\; \text{if vertex} \;\; \ell \;\; \text{is white,} \;\; \hat{a}_{ij} \;\; \text{otherwise.} \tag{3.3.62}$$

Below i_k associates with each vertex $k \in \{1, \ldots, \#V\}$ a number varying from 1 to s.

Definition 3.20 For a given *oriented bi-colored graph* $g = (V, E, c)$ we define the *elementary weight* as follows :

$$\Phi(g) := \sum_{i_v = 1, v \in V}^{s} \prod_{k \in V} b_{i_k} \prod_{(k,\ell) \in E} \tilde{a}_{i_k i_\ell} / b_{i_\ell}, \tag{3.3.63}$$

where $\sum_{i_v=1, v\in V}^{s} = \sum_{i_1=1}^{s} \cdots \sum_{i_{\#V}=1}^{s}$. Note that the graph is not necessarily connected.

In the case of bi-colored rooted trees, for which edges are oriented with the origin vertex closest to the root, all coefficient b_{i_ℓ} cancel, except those associated with the root, and we recover the left-hand-side of (3.3.61).

Lemma 3.21 *The elementary weight of a bi-colored oriented graph $g = (V, E, c)$ is the product of elementary weights of its connected components $\{g^q, q \in Q\}$, i.e.,*

$$\Phi(g) = \prod_{q \in Q} \Phi(g^q). \tag{3.3.64}$$

Proof. The product term in $\Phi(g)$ may be factored on terms depending on each connected component:

$$\Phi(g) = \sum_{i_v = 1, v \in V}^{s} \prod_{q \in Q} \left(\prod_{k \in V_q} b_{i_k} \prod_{(k,\ell) \in E_q} \frac{\tilde{a}_{i_k i_\ell}}{b_{i_\ell}} \right).$$

Denote by V_q the set of vertices in the qth connected component. We may then conclude by rewriting this sum of products as products of sums:

$$\Phi(g) = \prod_{q\in Q}\left(\sum_{i_v=1, v\in V}^{s}\left(\prod_{k\in V_q} b_{i_k}\prod_{(k,\ell)\in E_q}\frac{\tilde{a}_{i_k i_\ell}}{b_{i_\ell}}\right)\right).$$

■

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

$$F^{\top} := \{(x,y)\in V\times V;\ \ (y,x)\in F\}. \tag{3.3.65}$$

Theorem 3.22 *The elementary weight of a bi-colored oriented graph $g=(V,E,c)$, when* (3.3.14) *holds, satisfies*

$$\Phi(g) = \textstyle\sum_{\hat{E}_B\in\mathcal{P}(E_B)}(-1)^{\#\hat{E}_B}\Phi(V, E_W\cup\hat{E}_B^{\top}), \tag{3.3.66}$$

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 $\hat{a}$ in (3.3.14), we may write the elementary weight (3.3.63) as follows:

$$\Phi(g) = \sum_{i_v=1, v\in V}^{s}\prod_{k\in V} b_{i_k}\prod_{(k,\ell)\in E_W}\frac{a_{i_k i_\ell}}{b_{i_\ell}}\prod_{(k,\ell)\in E_B}\left(1-\frac{a_{i_\ell i_k}}{b_{i_k}}\right). \tag{3.3.67}$$

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 (3.3.66), the graphs $(V, E_W\cup\hat{E}_B^{\top})$ on the right hand side, are mono-colored oriented graph. Let h be a bi-colored H-tree. Then the only connected graph in sum of the right-hand-side of (3.3.66) is the one with $\hat{E}_B = E_B$. Observe that given an (mono-colored) H-tree h, we can reconstruct a bi-colored rooted tree t^h of the same order having h in its expansion in (3.3.66), 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-colored rooted trees, we can rewrite (3.3.66), separating the principal term for the others, as

$$\phi(t^h) = \sigma_0\Phi(h) + \textstyle\sum_{i\in I}\sigma_i\Phi(g_i^h), \tag{3.3.68}$$

where

$$I = \mathcal{P}(E_B)\setminus E_B,\ g_i^h = (V, E_W\cup i^{\top}),\ \sigma_i = (-1)^{\#i},\ \sigma_0 = (-1)^{\#E_B}. \tag{3.3.69}$$

Theorem 3.23 *For an SPRK scheme, the conditions for global order n are given by :*

$$\Phi(h) = \delta(h) \tag{3.3.70}$$

for all H-trees h of order not more than n, with $(I, g_i^h, \sigma_i, \sigma_0)$ defined by (3.3.69)*, and $\delta(h)$ inductively defined as*

$$\delta(h) = \sigma_0 \left(\tfrac{1}{\gamma(t^h)} - \textstyle\sum_{i\in I} \sigma_i \prod_{j\in J_i} \delta(h_i^j) \right). \tag{3.3.71}$$

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 3.21 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 (3.3.69) and h_i^j are its connected components. We conclude with (3.3.68). ■

Remark 3.24 We recover the result of Murua [34]: 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-colored rooted tree is already in [34]. 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.

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 [25]. We display the order conditions up to order 5 (order 6 conditions are displayed in [12]).

Table 3.1: Ordre 1

Graph	Condition
	$\sum b_i = 1$

Table 3.2: Ordre 2

Graph	Condition
	$\sum d_j = \frac{1}{2}$

Table 3.3: 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 3.4: 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 3.5: Ordre 5

Graph	Condition	Graph	Condition
	$\sum \frac{b_i}{b_k} a_{ik} a_{il} d_k c_l = \frac{3}{40}$		$\sum b_i a_{ik} a_{ij} c_j c_k = \frac{1}{20}$
	$\sum a_{lk} a_{kj} c_j d_l = \frac{1}{120}$		$\sum b_i a_{ik} a_{kj} c_i c_j = \frac{1}{30}$
	$\sum \frac{b_i}{b_k} a_{lk} a_{il} c_i d_k = \frac{11}{120}$		$\sum \frac{b_i}{b_k} a_{lk} a_{ik} c_i d_l = \frac{3}{40}$
	$\sum \frac{b_i b_j}{b_k} a_{jk} a_{ik} c_i c_j = \frac{2}{15}$		$\sum \frac{b_i}{b_l b_m} a_{im} a_{il} d_l d_m = \frac{2}{15}$
	$\sum \frac{1}{b_k} a_{ml} a_{lk} d_k d_m = \frac{1}{30}$		$\sum \frac{1}{b_k} a_{mk} a_{lk} d_l d_m = \frac{1}{20}$
	$\sum \frac{1}{b_l b_m} a_{lm} d_l^2 d_m = \frac{1}{15}$		$\sum \frac{1}{b_k} a_{kl} d_k^2 c_l = \frac{1}{60}$
	$\sum \frac{1}{b_l^2} a_{ml} d_l^2 d_m = \frac{1}{10}$		$\sum \frac{b_i}{b_l^2} a_{il} c_i d_l^2 = \frac{3}{20}$
	$\sum \frac{1}{b_k} a_{lk} d_k c_l d_l = \frac{7}{120}$		$\sum a_{jk} c_j d_j c_k = \frac{1}{40}$
	$\sum \frac{1}{b_k} a_{lk} c_k d_k d_l = \frac{1}{40}$		$\sum \frac{b_i}{b_k} a_{ik} c_i c_k d_k = \frac{7}{120}$
	$\sum \frac{b_i}{b_k} a_{ik} c_i^2 d_k = \frac{3}{20}$		$\sum b_i a_{ij} c_i^2 c_j = \frac{1}{10}$
	$\sum a_{kj} c_j^2 d_k = \frac{1}{60}$		$\sum b_i a_{ij} c_i c_j^2 = \frac{1}{15}$
	$\sum c_j^3 d_j = \frac{1}{20}$		$\sum b_i c_i^4 = \frac{1}{5}$
	$\sum \frac{1}{b_k} c_k^2 d_k^2 = \frac{1}{30}$		$\sum \frac{1}{b_l^2} c_l d_l^3 = \frac{1}{20}$
	$\sum \frac{1}{b_m^3} d_m^4 = \frac{1}{5}$		

Table 3.6: 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

3.4 Notes

The theory of rooted trees for computing expansions of differential equations has its origin in Cayley (On the theory of the analytic forms called trees, Phil. Magazine XIII(1857), 172-176). Its use for computing order conditions of Runge-Kutta methods was developed by Butcher, see [15]. Our presentation of the theory follows Hairer et al. [26]; the latter gives an extensive analysis of numerical integration of Hamiltonian systems. Murua [34] dealt with partitioned symplectic methods. He introduced the correspondance between bi-colored rooted trees and oriented free trees, and deduced the number of conditions to be satisfied for a given order. Sofroniou and Oevel [39] obtained order conditions for symplectic, but non partitioned RK schemes, using a specific parametrizatic of coefficients a and b.

The Runge-Kutta methods for optimal control problems and its transformed adjoint system were introduced in Hager [25], who stated the order conditions up to order 4. The link the the theory of partitioned Runge-Kutta methods, and the method of computation of order conditions is due to [12].

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. [4, 5, 11]).

Second-order accuracy for control constrained problems was obtained in [25] and Dontchev, Hager and Veliov [21]. A uniform accuracy of order $h^{3/2}$ is obtained for first-order state constrained problems in [20].

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. *AIAA J. of Guidance, Control and Dynamics*, 30(5):1228–1238, 2007.

[4] J.T. Betts. Survey of numerical methods for trajectory optimization. *AIAA J. of Guidance, Control and Dynamics*, 21:193–207, 1998.

[5] J.T. Betts. *Practical methods for optimal control using nonlinear programming*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2001.

[6] 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.

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

[8] 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*, 2007. Online First DOI 10.1007/s10107-007-0167-8.

[9] J.F. Bonnans and A. Hermant. Well-posedness of the shooting algorithm for state constrained optimal control problems with a single constraint and control. *SIAM J. Control Optimization*, 46(4):1398–1430, 2007.

[10] J.F. Bonnans and A. Hermant. Stability and sensitivity analysis for optimal control problems with a first-order state constraint. *ESAIM:COCV*, 2008. Available at http://hal.inria.fr/inria-00087573.

[11] J.F. Bonnans and G. Launay. Large scale direct optimal control applied to a re-entry problem. *AIAA J. of Guidance, Control and Dynamics*, 21:996–1000, 1998.

[12] J.F. Bonnans and J. Laurent-Varin. Computation of order conditions for symplectic partitioned runge-kutta schemes with application to optimal control. *Numerische Mathematik*, 103(1):1–10, 2006.

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

[14] B. Bonnard, L. Faubourg, and E. Trélat. *Mécanique céleste et contrôle des véhicules spatiaux*. Springer-Verlag, 2006.

[15] J.C. Butcher. *The numerical analysis of ordinary differential equations*. A Wiley-Interscience Publication. John Wiley & Sons Ltd., Chichester, 2003.

[16] F.H. Clarke. *Methods of dynamic and nonsmooth optimization*, volume 57 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia, 1989.

[17] F.H. Clarke, Yu. S. Ledyaev, R.J. Stern, and P. Wolenski. *Nonsmooth analysis and control theory*, volume 178 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1998.

[18] 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.

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

[20] A.L. Dontchev and W.W. Hager. The Euler approximation in state constrained optimal control. *Mathematics of Computation*, 70:173–203, 2001.

[21] A.L. Dontchev, W.W. Hager, and V.M. Veliov. Second-order Runge-Kutta approximations in control constrained optimal control. *SIAM Journal on Numerical Analysis*, 38:202–226 (electronic), 2000.

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

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

[24] H. Frankowska. Value function in optimal control, 2001. Lecture notes, Summer School on Mathematical Control Theory, Trieste.

[25] W. Hager. Runge-Kutta methods in optimal control and the transformed adjoint system. *Numerische Mathematik*, 87(2):247–282, 2000.

[26] E. Hairer, Ch. Lubich, and G. Wanner. *Geometric numerical integration*, volume 31 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2002.

[27] 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`.

[28] 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.

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

[30] 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.

[31] 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.

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

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

[34] A. Murua. On order conditions for partitioned symplectic methods. *SIAM J. on Numerical Analysis*, 34:2204–2211, 1997.

[35] 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.

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

[37] R.T. Rockafellar and R.J.-B. Wets. *Variational Analysis*. Springer-Verlag, New York, 1997.

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

[39] M. Sofroniou and W. Oevel. Symplectic Runge-Kutta schemes. I. Order conditions. *SIAM J. on Numerical Analysis*, 34:2063–2086, 1997.