RENDICONTI DEL SEMINARIO MATEMATICO

Università e Politecnico di Torino

Control Theory and Stabilization, I

CONTENTS

Volume 63, N. 4 **2005**

Preface

The papers collected in this issue are a selection of the contributions presented at the "Second Junior European Meeting on Control Theory and Stabilization", held in Torino (Italy) December 3-5, 2003.

The meeting was the second of a series of junior conferences whose aim is to stimulate contacts between European young researchers working in Mathematical Control Theory.

The articles cover many challenging subjects of researches in Mathematical Control Theory: from hybrid systems and quantum systems to various aspects of optimal control.

F. Ceragioli, F. Fagnani

Rend. Sem. Mat. Univ. Pol. Torino - Vol. 63, 4 (2005)
Control Theory and Stabil., I

L. Abrunheiro* – M. Camarinha*

RIEMANNIAN CUBIC POLYNOMIALS

Abstract. This paper gives an analysis of the Riemannian cubic polynomials, with special interest in the Lie group $SO(3)$, based on the study of a second order variational problem. The corresponding Euler-Lagrange equation gives rise to an interesting system of nonlinear differential equations. Motivated by the problem of the motion of a rigid body, the reduction of the essential size and the separation of the variables of the system are obtained by means of invariants along the cubic polynomials.

1. Introduction

The question of generalizing cubic polynomials to non-Euclidean spaces has been the object of intensive investigation in the context of interpolation theory. It is well known that the cubic polynomials are solutions of a variational problem in the Euclidean space, whose Lagrangian function is the squared norm acceleration. A generalization of this notion to Riemannian manifolds was introduced in 1989 [8] and explored from a dynamic interpolation perspective in 1995 [5]. In the last years it has been developed an interesting geometric theory related to the cubic polynomials which is surprisingly close to the Riemannian theory of geodesics [5, 4]. An analytical theory for cubic polynomials (Ljusternik -Schnirelman theory, Morse theory) has also been established [6]. After that, it became important to find interesting cubic polynomials examples, namely the $SO(3)$ case.

The main idea of the present paper is a qualitative analysis of the cubic polynomials on $SO(3)$ and is motivated by the particular importance of the notion of cubic polynomials on Lie groups to the development of interpolation theory, for instance in trajectory planning for rigid body motion with applications to aeronautics and robotics. An analogous analysis, but for the particular case of the so-called null cubic polynomials, was produced in [9]. The paper starts with the definition of the generalized Riemannian cubic polynomials and the deduction of some invariants along the cubic polynomials. It follows a special analysis of the problem of the motion of a rigid body as a motivation to section 4, which is devoted to the essential case of $SO(3)$. In this specific situation, a reduction of the cubic polynomial equation and a property related with a reparametrization of the cubic polynomial are obtained.

2. Cubic polynomials on a Riemannian manifold

Consider a Riemannian manifold M, with Riemannian metric $\langle ., . \rangle$. Denote the symmetric connection on M, which is compatible with this metric, by ∇, and the covariant

*Work partially supported by ISR, FCT (Portuguese Science and Technology Foundation) project posi/sri/41618/2001.

derivative along a curve x in M by DX/dt where X is a vector field along x. Moreover, denote the curvature tensor field by R and the covariant differential of R by ∇R.

In order to generalize the notion of cubic polynomials to a Riemannian manifold, the following second order variational problem in M was formulated [8, 5]

$$\min_{x\in\Omega} \frac{1}{2}\int_0^T \left\langle \frac{D^2x}{dt^2}\, ,\, \frac{D^2x}{dt^2} \right\rangle dt$$

where Ω is the class of C^1 piecewise smooth curves $x : [0, T] \to M$, satisfying

$$x(0) = p, \quad \frac{dx}{dt}(0) = \upsilon, \quad x(T) = q, \quad \frac{dx}{dt}(T) = u,$$

with $(p, \upsilon), (q, u) \in TM$ and $T \in \mathbb{R}^+$.

Cubic polynomials on M are smooth curves $x : [0, T] \to M$ that are solutions of the Euler-Lagrange equation

$$\frac{D^4x}{dt^4} + R\left(\frac{D^2x}{dt^2}\, ,\, \frac{dx}{dt}\right)\frac{dx}{dt} = 0 \tag{1}$$

Consider x a cubic polynomial and denote the velocity vector field along x, $\dfrac{dx}{dt}$, by V.

LEMMA 1. *[2] The following expression is invariant along the cubic polynomial*

$$\left\langle \frac{D^2V}{dt^2}\, ,\, V \right\rangle - \frac{1}{2}\left\langle \frac{DV}{dt}\, ,\, \frac{DV}{dt} \right\rangle.$$

Proof. The result follows from the integration of the inner product of (1) with V. □

LEMMA 2.

$$\frac{d}{dt}\left[\left\langle \frac{D^2V}{dt^2}, \frac{D^2V}{dt^2} \right\rangle - \left\langle \frac{D^3V}{dt^3}, \frac{DV}{dt} \right\rangle\right] = \left\langle (\nabla_V\, R)\left(\frac{DV}{dt}, V\right) V, \frac{DV}{dt} \right\rangle$$

Proof. Use the tensor curvature property $\langle R(X, Y)Z, W\rangle = \langle R(W, Z)Y, X\rangle$ and the definition of the covariant differentiation of the curvature tensor R, that is,

$$\begin{aligned}&(\nabla_X\, R)(Y, Z)W\\ &\quad = \nabla_X\, [R(Y, Z)W] - R(\nabla_X Y, Z)W - R(Y, \nabla_X Z)W - R(Y, Z)\nabla_X W,\end{aligned}$$

to get

$$\frac{d}{dt}\left\langle R\left(\frac{DV}{dt}V\right),\frac{DV}{dt}\right\rangle - \left\langle (\nabla R)\left(\frac{DV}{dt},V\right)V,\frac{DV}{dt}\right\rangle$$
$$= 2\left\langle R\left(\frac{DV}{dt},V\right)V,\frac{D^2V}{dt^2}\right\rangle.$$

Moreover, notice that

$$\frac{d}{dt}\left\langle \frac{D^2V}{dt^2}\,,\;\frac{D^2V}{dt^2}\right\rangle = 2\left\langle \frac{D^3V}{dt^3}\,,\;\frac{D^2V}{dt^2}\right\rangle$$

In order to complete the proof, it is sufficient to make the inner product of (1) with $\frac{D^2V}{dt^2}$ and apply the above equalities in the obtained equation. □

REMARK 1. In Riemannian locally symmetric manifolds, the lemma 2 gives a second invariant along the cubic polynomial,

$$\left\langle \frac{D^2V}{dt^2},\frac{D^2V}{dt^2}\right\rangle - \left\langle \frac{D^3V}{dt^3},\frac{DV}{dt}\right\rangle.$$

3. Motion of a rigid body

The Lie group $SO(3)$ is the configuration space for the motion of a rigid body with no external forces and fixed centre of mass. This problem is an interesting motivation to the study of cubic polynomials on $SO(3)$, since the motion can be described as motion along geodesics on the Lie group $SO(3)$ provided with a left-invariant Riemannian metric ([1], appendix 2).

The motion of a rigid body is described by the equation

$$\frac{d}{dt}(\mathrm{J}\,y) = (\mathrm{J}\,y)\times y \tag{2}$$

where $y = (y_1, y_2, y_3)^t$ is the angular velocity and J is the diagonal matrix whose elements are the principal moments of inertia in the body. The following two invariants are satisfied

$$\langle \mathrm{J}\,y\,,\;y\rangle = c \qquad \text{and} \qquad \langle \mathrm{J}^2\,y\,,\;y\rangle = m \tag{3}$$

with c and m real constants.

PROPOSITION 1. *The rigid body motion equation (2) reduces to*

$$\frac{d^2y}{dt^2} = D_1y + D_2\begin{bmatrix} {y_1}^3 \\ {y_2}^3 \\ {y_3}^3 \end{bmatrix}$$

with D_1 and D_2 diagonal matrices.

Proof. Use the property $u \times v = \dfrac{J}{|J|}\,[(Ju) \times (Jv)]$ to rewrite (2) as $\dfrac{dy}{dt} = \dfrac{1}{|J|}(J^2\, y) \times (J\, y)$. Now differentiate this, use (2) and apply the cross product property

$$u \times (v \times w) = \langle u, w\rangle\, v - \langle u, v\rangle\, w \tag{4}$$

to obtain

$$\frac{d^2y}{dt^2} = \langle J\, y\,,\ y\rangle\, J^{-1}\, y \;+\; \frac{1}{|J|}\,\langle J^2\, y\,,\ y\rangle\, J\, y \;-\; \left(\langle y\,,\ y\rangle \;+\; \frac{1}{|J|}\,\langle J^2\, y\,,\ J\, y\rangle\right) y$$

To conclude the proof use in the above equation the following property and the invariants (3)

$$\langle J^2 y\,,\ J\, y\rangle = |J|\langle y\,,\ y\rangle - \mathrm{tr}(J_{\mathrm{cof}})\langle Jy\,,\ y\rangle + \mathrm{tr}J\langle J^2 y\,,\ y\rangle$$

where $|.|$ denotes the matrix determinant, tr the matrix trace and $._{\mathrm{cof}}$ the cofactor matrix. □

REMARK 2. The proposition 1 gives an alternative proof of the well known result (see [7]) that the rigid body equations can be solved in terms of the Jacobian elliptic functions. Indeed, each coordinate of y is a solution of the differential equation $(dz/dt)^2 = az^4 + bz^2 + c$ [10].

4. Cubic polynomials on $SO(3)$

Consider the Lie group $SO(3)$ equipped with a bi-invariant Riemannian metric and the corresponding Lie algebra $(so(3), [.,.])$. Let $\hat{.}$ denotes the isomorphism between the Lie algebras $(\mathbb{R}^3, \times)$ and $(so(3), [.,.])$.

THEOREM 1. *[9] A smooth curve $x : I \to SO(3)$ is a cubic polynomial if and only if it satisfies*

$$(dL_{x^{-1}})_x \frac{dx}{dt} = \hat{v} \tag{5}$$

$$\frac{d^3v}{dt^3} + v \times \frac{d^2v}{dt^2} = 0 \tag{6}$$

The smooth curves $v : I \to \mathbb{R}^3$, solutions of the equation (6), are called *Lie quadratics*.

LEMMA 3. *If v is a Lie quadratic, then the following invariants along v are* *satisfied:*

$$\langle \frac{d^2 v}{dt^2}\ ,\ v\rangle - \frac{1}{2}\langle \frac{dv}{dt}\ ,\ \frac{dv}{dt}\rangle = I_1 \tag{7}$$

$$\langle \frac{d^2 v}{dt^2}\ ,\ v \times \frac{dv}{dt}\rangle + \frac{1}{2}\langle v \times \frac{dv}{dt}\ ,\ v \times \frac{dv}{dt}\rangle = I_2 \tag{8}$$

$$\langle \frac{d^2 v}{dt^2}\ ,\ \frac{d^2 v}{dt^2}\rangle = I_3 \tag{9}$$

Proof. The result follows from the integration of the inner product of the equation (6) with v, $v \times \dfrac{dv}{dt}$ and $\dfrac{d^2v}{dt^2}$, respectively. □

Observe that (7) and (8) correspond to the invariants presented in Section 2.

PROPOSITION 2. *Let v be a Lie quadratic. Then $f = \langle v, v\rangle$ satisfies*

$$f^{(4)} + ff'' - \frac{3}{4}(f')^2 - 2I_1 f - 6(I_2 + I_3) = 0. \tag{10}$$

Furthermore, if $y = f(t)$ is a solution of the differential equation (10), then the equation (6) reduces to

$$\frac{d^5 v}{dt^5}\ +\ f\ \frac{d^3 v}{dt^3}\ +\ \frac{3}{2}\ f'\ \frac{d^2 v}{dt^2}\ -\ (\frac{1}{6}\ f''\ +\ \frac{2}{3}\ I_1)\ \frac{dv}{dt}\ -\ \frac{1}{3}\ f^{(3)}\ v = 0.$$

Proof. From (7) it follows $\dfrac{d^2}{dt^2}\langle v, v\rangle = 3\langle \dfrac{dv}{dt}, \dfrac{dv}{dt}\rangle + 2I_1$. On the other hand, (8) and (9) imply

$$\frac{d^2}{dt^2}\langle \frac{dv}{dt}, \frac{dv}{dt}\rangle = \langle \frac{dv}{dt}, v\rangle^2 - \langle v, v\rangle\langle \frac{dv}{dt}, \frac{dv}{dt}\rangle + 2(I_2 + I_3).$$

Conjugating the above equalities the differential equation (10) yields. In order to complete the proof, it suffices to differentiate twice the equation (6) and use (4). □

REMARK 3. Note that (6) is equivalent to $\dfrac{d^2 v}{dt^2} + v \times \dfrac{dv}{dt} = C$, $C \in \mathbb{R}^3$. The *null Lie quadratics* ($C = 0$) were studied in [9] and Proposition 2 was proved to this case ([9], Lemma 3).

REMARK 4. The proposition 2 motivates interesting problems, such as numerical methods applications or integrability study of the corresponding Hamiltonian system. The cubic polynomials may be expressed as extremals of sub-Riemannian optimal control problems on $TSO(3)$, whose Hamiltonian function is exactly the first invariant (7) (see [3] for details)

$$\langle \frac{d^2 v}{dt^2}\ ,\ v\rangle - \frac{1}{2}\langle \frac{dv}{dt}\ ,\ \frac{dv}{dt}\rangle.$$

It is well known from the theory of geodesics, that only the linear reparametrization preserves the curve as a geodesic. This result follows from the fact that the velocity vector field length is an invariant along the geodesic. The invariant (8) has a similar role in the analysis of the freedom for cubic polynomial reparametrization. However, for the cubic polynomial, the study of the degenerated case turns out to be particularly interesting.

The following proposition analyses the freedom for cubic polynomial reparametrization.

PROPOSITION 3. *Let $x : I \to SO(3)$ be a cubic polynomial and $\upsilon : I \to \mathbb{R}^3$ the corresponding Lie quadratic. If the invariant (8) along υ is no zero, a reparametrization of x is a cubic polynomial if and only if it is linear.*

Proof. It is enough to note that, if $y : I' \to SO(3)$ is a smooth curve obtained by a reparametrization of x, $y = x \circ s$, and $w : I' \to \mathbb{R}^3$ is the corresponding Lie curve defined by $(dL_{y^{-1}})_y \dfrac{dy}{dt} = \hat{w}$, then

$$\langle \frac{d^2 w}{dt^2}, \; w \times \frac{dw}{dt} \rangle + \frac{1}{2} \langle w \times \frac{dw}{dt}, \; w \times \frac{dw}{dt} \rangle$$
$$= (s')^6 \left(\langle \frac{d^2 \upsilon}{dt^2}, \; \upsilon \times \frac{d\upsilon}{dt} \rangle + \frac{1}{2} \langle \upsilon \times \frac{d\upsilon}{dt}, \; \upsilon \times \frac{d\upsilon}{dt} \rangle \right).$$

□

References

[1] ARNOLD V.I., *Mathematical methods of classical mechanics*, Graduate Texts in Mathematics **60**, Springer-Verlag, New York 1989.

[2] CAMARINHA M., *The geometry of cubic polynomials on Riemannian manifolds*, Ph. D. Thesis in Pure Mathematics, University of Coimbra, Portugal 1996.

[3] CAMARINHA M., CROUCH P. AND SILVA LEITE F., *Hamiltonian structure of generalized cubic polynomials*, Preprints of the Workshop of Lagrangian and Hamiltonian Methods for Nonlinear Control, Princeton, New Jersey, USA, (2000), 13–18.

[4] CAMARINHA M., CROUCH P. AND SILVA LEITE F., *On the geometry of Riemannian cubic polynomials*, Differential Geom. Appl. **15** (2001), 107–135.

[5] CROUCH P. AND SILVA LEITE F., *The dynamic interpolation problem on Riemannian manifolds, Lie groups and symmetric spaces*, J. Dynam. Control Systems **I2** (1995), 177–202.

[6] GIAMBÒ R., GIANNONI F. AND PICCIONE P., *An analytical theory for Riemannian cubic polynomials*, IMA J. Math. Control Inform. **19** (2002), 445–460.

[7] HERMANN R., *Differential geometry and the calculus of variations*, Academic Press Inc., New York 1968.

[8] NOAKES L., HEINZINGER G. AND PADEN B, *Cubic splines on curved spaces*, IMA J. of Math. Control Inform. **6** (1989), 465–473.

[9] NOAKES L., *Null cubics and Lie quadratics*, J. Math. Phys. **44** 3 (2003), 1436–1448.

[10] PRASOLOV V. AND SOLOVYEV Y., *Elliptic functions and elliptic integrals*, Translations of Mathematical Monographs **170**, American Math. Society, Providence R.I. 1997.

AMS Subject Classification: 53B20, 58E10, 49S05.

Lígia ABRUNHEIRO, Instituto Superior de Contabilidade e Administração, Universidade de Aveiro, R. Ass. Hum. dos Bombeiros de Aveiro Apartado 58, 3811-953 Aveiro, PORTUGAL
e-mail: ligia.abrunheiro@isca.ua.pt

Margarida CAMARINHA, Departamento de Matemática, Universidade de Coimbra, FCTUC Apartado 3008, 3001-454 Coimbra, PORTUGAL
e-mail: mmlsc@mat.uc.pt

Rend. Sem. Mat. Univ. Pol. Torino - Vol. 63, 4 (2005)
Control Theory and Stabil., I

C. Altafini

ON THE EXACT UNITARY INTEGRATION OF TIME-VARYING QUANTUM LIOUVILLE EQUATIONS

Abstract. In this paper, the Dyson series corresponding to the time-varying Hamiltonian of a finite dimensional quantum mechanical system is expanded in terms of products of exponentials of a complete basis of commutator superoperators in the corresponding Liouville space. The Cayley-Hamilton theorem and the Wei-Norman formula allow to express explicitly the functional relation between the Dyson series and the product of exponentials via a set of first order differential equations. Since the method is structure preserving, it can be used for the exact unitary integration of the driven Liouville-von Neumann equation.

1. Introduction

The general solution of a quantum Liouville equation for time-varying Hamiltonians is given by the Dyson series. Normal procedure for its practical use is to truncate this expansion and work with the corresponding approximation. Beside providing approximate solutions, the main drawback of such truncations is that the unitarity of the time evolution is not necessarily preserved [17]. The method we present here relies on the formalism of the canonical coordinates of the first and second kind of the adjoint representation of the unitary group, and on the relation between them. In fact, the differential operator governing the Liouville equation can be related to a product of exponentials of the noncommuting operators corresponding to a complete basis of the adjoint representation of the Lie algebra via a set of nonlinear differential equations known as Wei-Norman formulæ. Such formulæ allows one to express the unitary evolution of the density operator *exactly* in terms of the product of exponentials. For $N \times N$ density operators, this is best understood in Liouville space. Once the parameterization of the density matrix is given in terms of the $N^2 - 1$ basis elements of $\mathrm{ad}_{\mathfrak{su}(N)}$ (i.e. the commutator superoperator of the Liouville space) plus the identity operator, then the method corresponds to solving a time-varying system of ODEs and is one of the most popular structure preserving algorithm used by the numerical algebra community [8, 14]. The algorithm preserves unitarity, as the real time-varying parameters are multiplied by skew-hermitian matrices (the corresponding infinitesimal generators in the basis) and then exponentiated. Particular cases of the formula we use have already appeared in the literature to treat $\mathfrak{su}(2)$-systems like spin $\frac{1}{2}$ or two-level systems [15, 17], examples which we also discuss below. The method, however, is absolutely general for all finite dimensional unitary operator algebras and for all "generalized" Euler angles one can choose on such algebras. Furthermore, it can be used for both pure and mixed states. A couple of applications, other than exact numerical simulation, are as follows. It can be used to reconstruct the behavior of a driven Hamiltonian from sequences of pulses or, on the other direction, to decompose a time-varying Hamiltonian into pulse sequences. This last will be treated in Section 4.

2. Time-independent Hamiltonians

In quantum mechanics, Liouville equations are very common to describe the time evolution of density operators or of observables in the Heisenberg picture. If H is a constant finite dimensional Hamiltonian, the density operator differential equation

$$\dot{\rho}(t) = -i[H,\ \rho] = -i\,\mathrm{ad}_H(\rho) \tag{1}$$

is solved by

$$\rho(t) = e^{-itH}\rho(0)e^{itH} = \mathrm{Ad}_{e^{-itH}}\rho(0) = e^{-it\mathrm{ad}_H}\rho(0) \tag{2}$$

If $-iH \in \mathfrak{su}(N)$, then in (2) $-i\mathrm{ad}_H$ is a so-called commutator superoperator i.e. a linear operator in the N^2 dimensional Liouville space obtained by expanding the density operator in a complete set of basis operators like the one obtained by choosing the N-dimensional Pauli matrices $\lambda_1, \ldots, \lambda_{N^2-1}$ (see [10] for an explicit expression for these matrices) plus the identity matrix $\lambda_0 = N^{-\frac{1}{2}}I$: $\rho = \sum_{j=0}^{n}\rho_j\lambda_j$. As is well-known for this parameterization, the coefficient ρ_0 along λ_0 is a constant fixed by the $\mathrm{tr}(\rho) = 1$ condition to $\rho_0 = N^{-\frac{1}{2}}$. Thus the evolution represented by (1) occurs along an hyperplane of the Liouville space, see [6]. Call $n = N^2 - 1$ the dimension of such hyperplane (equal to $\dim\mathfrak{su}(N)$). On the vector of n real components ρ_j, call it $\vec{\rho}$, the action of $-i\mathrm{ad}_H$ is linear:

$$\dot{\vec{\rho}} = -i\,\mathrm{ad}_H\,\vec{\rho} \tag{3}$$

The $\{\lambda_j\}$ basis of $\mathfrak{su}(N)$ corresponds to purely imaginary structure constants. For the scope of this paper, it is convenient to choose a skew-hermitian basis for $\mathfrak{su}(N)$, call it $A_1, \ldots A_n$ for which we have all real structure constants $[A_i,\ A_j] = \sum_{k=1}^{n} c_{ij}^k A_k$, $c_{ij}^k \in \mathbb{R}$. Then $-iH = \sum_{j=1}^{n} u_j A_j$ with $u_j \in \mathbb{R}$. The corresponding basis in the adjoint representation is given by the $n \times n$ matrices $\mathrm{ad}_{A_1}, \ldots \mathrm{ad}_{A_n}$ and $-i\mathrm{ad}_H = \sum_{j=1}^{n} u_j\mathrm{ad}_{A_j}$, where the ad_{A_i} have matrix elements $\left(\mathrm{ad}_{A_i}\right)_{jk} = c_{ij}^k$. The $n \times n$ matrices $\mathrm{ad}_{A_1}, \ldots, \mathrm{ad}_{A_n}$ are real and skew-symmetric and as such they are part of a basis of $\mathfrak{so}(n)$. Since $\dim\mathfrak{so}(n) = \frac{n(n-1)}{2} = \frac{N^4-3N^2+2}{2}$, for $N > 2$ the n matrices $\mathrm{ad}_{A_1}, \ldots, \mathrm{ad}_{A_n}$ span only a proper subalgebra of $\mathfrak{so}(n)$. For example for $N = 3$ $n = \dim\mathfrak{su}(3) = 8$ while $\dim\mathfrak{so}(8) = 28$! Just like the time evolutor of the Schrödinger equation is unitary, $|\psi\rangle = U(t)|\psi(0)\rangle$, $U(t) \in SU(N)$, for the Liouville equation (3) the adjoint representation giving matrices on $\mathfrak{so}(n)$, the propagator for $\vec{\rho}$ is an orthogonal matrix:

$$\vec{\rho}(t) = O(t)\vec{\rho}(0), \qquad O(t) \in SO(n).$$

For a generic (i.e. not necessarily diagonal) H, there exist many ways to compute the exponential $e^{-it\mathrm{ad}_H}$ other than its infinite series expansion:

$$e^{-it\mathrm{ad}_H} = \sum_{k=0}^{\infty}\frac{(-it)^k}{k!}\mathrm{ad}_H^k \tag{4}$$

see the classical survey [12] and the recent "classroom notes" of SIAM Review [7, 9] and references therein. Here we use a method based on the Cayley-Hamilton theorem. The method consists in expressing the series expansion of $e^{-i\,\mathrm{ad}_H}$ in terms of the first $n-1$ powers of ad_H with suitable coefficients depending on the coefficients of the characteristic polynomial of $-i\,\mathrm{ad}_H$ and on t. It is most suited for the adjoint representation, as the powers of the basis elements ad_{A_i} are immediately expressed in terms of the structure constants of the Lie algebra. The r-th power of ad_{A_i} is in fact given by

$$\mathrm{ad}^r_{A_i} = \left(\mathrm{ad}^r_{A_i}\right)_{kj} = \sum_{l_1,\ldots l_{r-1}=1}^{n} c^{k}_{1\,l_1} c^{l_1}_{1\,l_2} \ldots c^{l_{r-2}}_{1\,l_{r-1}} c^{l_{r-1}}_{1\,j} \tag{5}$$

If the characteristic polynomial is $\det(sI - (-i\,\mathrm{ad}_H)) = s^n - a_{n-1}s^{n-1} - \ldots - a_1 s - a_0$, with coefficients $a_{n-1} = \mathrm{tr}(-i\,\mathrm{ad}_H), \ldots, a_0 = (-1)^n \det(-i\,\mathrm{ad}_H)$, the Cayley-Hamilton theorem affirms that $-i\,\mathrm{ad}_H$ satisfies its own characteristic equation, i.e.

$$-i\,\mathrm{ad}^n_H = a_0 I + a_1(-i\,\mathrm{ad}_H) + a_2(-i\,\mathrm{ad}_H)^2 + \ldots + a_{n-1}(-i\,\mathrm{ad}_H)^{n-1} \tag{6}$$

and the infinite sum (4) can always be written as

$$e^{-it\,\mathrm{ad}_H} = \sum_{k=0}^{n-1} \beta_k (-i\,\mathrm{ad}_H)^k \tag{7}$$

for suitable $\beta_k = \beta_k(a_0, \ldots, a_{n-1}, t)$, computed in detail in [3].

The procedure is valid also for the the skew-hermitian basis elements $A_i \in \mathfrak{su}(N)$, with the powers of ad_{A_i} expressed in terms of the structure constants as in (5). Using the notation $\beta^{[i]}_0, \beta^{[i]}_1, \ldots \beta^{[i]}_{n-1}$ for the coefficients corresponding to (7), we have

$$\begin{aligned} e^{\gamma_i \mathrm{ad}_{A_i}} &= \beta^{[i]}_0 \delta^k_j + \beta^{[i]}_1 c^k_{i\,j} + \sum_{l_1=1}^{n} \beta^{[i]}_2 c^k_{i\,l_1} c^{l_1}_{i\,j} + \ldots + \sum_{l_1,\ldots l_{n-2}=1}^{n} \beta^{[i]}_{n-1} c^k_{i\,l_1} c^{l_1}_{i\,l_2} \ldots c^{l_{n-2}}_{i\,j} \\ &= \sum_{r=0}^{n-1} \sum_{l_1,\ldots l_{r-1}=1}^{n} \beta^{[i]}_r c^k_{i\,l_1} c^{l_1}_{i\,l_2} \ldots c^{l_{r-1}}_{i\,j} \end{aligned} \tag{8}$$

where it is intended that

$$\sum_{l_1,\ldots l_{r-1}=1}^{n} c^k_{i\,l_1} c^{l_1}_{i\,l_2} \ldots c^{l_{r-1}}_{i\,j} = \delta^k_j \quad \text{for} \quad r = 0$$

and

$$\sum_{l_1,\ldots l_{r-1}=1}^{n} c^k_{i\,l_1} c^{l_1}_{i\,l_2} \ldots c^{l_{r-1}}_{i\,j} = c^k_{ij} \quad \text{for} \quad r = 1$$

(the lower index in $\beta^{[i]}_k$ gives the number of times the structure constants $c^*_{i\,*}$ appear in the corresponding term).

3. Time-varying Hamiltonians

For time varying Hamiltonians, the linearity of (3) as a differential equation in the $\vec{\rho}$ coordinates implies that there are two standard ways to express its local solution, similarly to what happens for all linearly time-varying systems of differential equations [19]. The situation is obviously specular to the case of the time-varying Schrödinger equation, which has already been studied via similar techniques in [3].

When $H = H(t)$ (i.e. $u_j = u_j(t)$ in the $\mathfrak{su}(N)$ basis), the local solution of (1) can be expressed in terms of infinite formal series in the style of chronological calculus [1] or of the Dyson series, as it is commonly referred to in quantum physics [4]. Using the Dyson time-ordering operator T for $-i\,\mathrm{ad}_{H(t)} = -i\,\mathrm{ad}_{(u_1(\tau)A_1+\ldots+u_n(\tau)A_n)} = -i(\mathrm{ad}_{u_1(\tau)A_1} + \ldots + \mathrm{ad}_{u_n(\tau)A_n})$:

(9)
$$\begin{aligned}\vec{\rho}(t) &= T\exp\left(\int_0^t -i\,\mathrm{ad}_{H(\tau)}d\tau\right)\vec{\rho}(0)\\ &=\left(I + \sum_{k=1}^{\infty}\int\int\cdots\int_{0\leq\tau_1<\ldots<\tau_k\leq t}(-i\,\mathrm{ad}_{H(\tau_1)})\ldots(-i\,\mathrm{ad}_{H(\tau_k)})d\tau_k\ldots d\tau_1\right)\vec{\rho}(0)\end{aligned}$$

Alternatively, it can be written as a product of exponentials over all basis elements on the adjoint representation with arbitrarily fixed order, here the cardinal order:

$$\vec{\rho}(t) = e^{\gamma_1(t)\mathrm{ad}_{A_1}}\ldots e^{\gamma_n(t)\mathrm{ad}_{A_n}}\vec{\rho}(0) \tag{10}$$

with $\gamma_j(t)$ real valued parameters expressing the time-dependence of the solution. Notice that, as said above, $\mathrm{ad}_{\mathfrak{su}(N)} \subseteq \mathfrak{so}(n)$ ($\subsetneq$ for $N > 2$), but $\mathrm{ad}_{\mathfrak{su}(N)}$ is a subalgebra and therefore it is closed under commutation. Hence all integral curves of (9) will belong to $\exp(\mathrm{ad}_{\mathfrak{su}(N)})$ and the canonical coordinates of the second kind i.e. the product of exponentials (10) can be restricted to this subalgebra of $\mathfrak{so}(n)$ only.

The relation between (9) and (10) is given by the so-called Wei-Norman formula, [18], which is the Jacobian of the (locally invertible) transformation from (9) to (10) and is obtained by comparing (3) and the derivative of (10)

$$\begin{aligned}\left(\sum_{j=1}^{n}u_j\mathrm{ad}_{A_j}\right)\vec{\rho} &= \frac{d}{dt}\left(e^{\gamma_1(t)\mathrm{ad}_{A_1}}\ldots e^{\gamma_n(t)\mathrm{ad}_{A_n}}\right)\vec{\rho}(0)\\ &= \sum_{j=1}^{n}\left(\prod_{k=1}^{j}e^{\gamma_k\mathrm{ad}_{A_k}}\right)\dot{\gamma}_j\mathrm{ad}_{A_j}\vec{\rho}(t)\end{aligned}$$

along each of the $\mathfrak{su}(N)$ basis directions ad_{A_j}. Equation (8) can be used to compute in a closed form the $e^{\gamma_k\mathrm{ad}_{A_k}}$. The result is a set of n differential equations nonlinear in the

$\gamma_j(t)$ but linear in the $u_j(t)$ that relate the two sets of parameters:

$$\begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} = \Xi(\gamma_1, \ldots, \gamma_n) \begin{bmatrix} \dot{\gamma}_1 \\ \vdots \\ \dot{\gamma}_n \end{bmatrix} \tag{11}$$

which can be (locally) inverted to give:

$$\begin{bmatrix} \dot{\gamma}_1 \\ \vdots \\ \dot{\gamma}_n \end{bmatrix} = \Xi(\gamma_1, \ldots, \gamma_n)^{-1} \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix} \tag{12}$$

See [3, 2] for issues related to the non-globality of the used parameterization. The two sets of parameters live on the Lie algebra $\mathrm{ad}_{\mathfrak{su}(N)}$. Its skew-hermitian structure plus the exponentiation operation guarantee that unitarity is preserved in both the single exponential (9) and the product of exponentials (10).

While the matrix Ξ can always be obtained explicitly, see [3], except for a few simple cases like two-level systems, see [13, 15, 16], the analytic solution of (12) becomes quickly prohibitive with the dimension N. However, the systems of ODEs (11) or (12) can be numerically integrated in a structure preserving fashion using ordinary simulation tools.

4. Application: time-varying Hamiltonian and pulse sequences for a spin $\frac{1}{2}$ system

In NMR spectroscopy, a nuclear spin is manipulated by the application of suitable pulses along different axes, see [5, 11]. One of the main issues is then the reconstruction of the time-varying Hamiltonian which would correspond to a sequence of pulses and viceversa.

For sake of simplicity, we work in the case:

1. there is no constant magnetic field applied, i.e. the free Hamiltonian is zero;
2. the pulses are gaussian in shape;

In particular the first assumption means that only the interaction part of the Hamiltonian is considered and that we can work in the laboratory frame. Choosing Pauli-like skew-Hermitian matrices

$$A_1 = \frac{1}{2}\begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix} \quad A_2 = \frac{1}{2}\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \quad A_3 = \frac{1}{2}\begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix} \tag{13}$$

we get the adjoint basis for $\mathfrak{su}(2)$

$$\mathrm{ad}_{A_1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix} \quad \mathrm{ad}_{A_2} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix} \quad \mathrm{ad}_{A_3} = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Eq. (12) will look like:

$$\Xi^{-1} = \begin{bmatrix} 1 & \sin\gamma_1 \tan\gamma_2 & -\cos\gamma_1 \tan\gamma_2 \\ 0 & \cos\gamma_1 & \sin\gamma_1 \\ 0 & -\sec\gamma_2 \sin\gamma_1 & \cos\gamma_1 \sec\gamma_2 \end{bmatrix} \tag{14}$$

which corresponds to the inverse of equation (6) of [15]. Notice that, while on (10) the cardinal order is followed, changing the ordering (and also using repeated generators along the same direction) will lead to still admissible formulæ, see [2] for details.

Pulses of known shape are applied along the X and Y directions in different ways. The shape of the k-th gaussian pulse of amplitude $\frac{A_k}{\sigma_k\sqrt{2\pi}}$ and centered at $\tau_k = t_{k-1} + \frac{\Delta t_k}{2}$, where Δt_k is the time support of the k-th pulse and σ^2 its "variance", is

$$u_k(t) = \frac{A_k}{\sigma_k\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\tau_k}{\sigma_k}\right)^2} \tag{15}$$

Under the assumptions above, the Magnus expansion

$$\vec{\rho}(t) = T\exp\left(\int_0^t \left(\mathrm{ad}_{u_1(\tau)A_1} + \mathrm{ad}_{u_2(\tau)A_2}\right) d\tau\right) \vec{\rho}(0) \tag{16}$$

maps the (pure or mixed) density operator $\vec{\rho}_0$ to

$$\vec{\rho}(t_3) = e^{\gamma_1 \mathrm{ad}_{A_1}} e^{\gamma_2 \mathrm{ad}_{A_2}} e^{\gamma_3 \mathrm{ad}_{A_3}} \vec{\rho}_0 \tag{17}$$

with the $\gamma_j = \gamma_j(t)$ obtained numerically from the ODEs (12). In all the simulations below, the model is adimensional: $\hbar$ and the gyromagnetic ratio are set to 1 and the time scales and amplitudes refer to arbitrary units.

4.1. Case I: simultaneous pulses

Along X and Y apply simultaneously identical gaussian pulses with $\sigma_1 = \sigma_2 = 1$, $A_1 = A_2 = 1$ and centered at $\tau_1 = \tau_2 = 3$. The u_1 and u_2 coordinates of the Magnus expansion (16) and the corresponding time-varying coordinates γ_j, $j = 1, 2, 3$ in the product of exponentials representation (17) are shown respectively Fig. 1 and Fig. 2. While $u_3 \equiv 0$, $\gamma_3 \neq 0$ because of noncommutativity.

4.2. Case II: pulses with disjoint support

In (15), if $A_k = 1$ and we consider the "classical" 3σ case, i.e. choose $\frac{\Delta t_k}{2} = 3\sigma$, then 99.7% of the pulse area is contained in the interval $[t_{k-1}, t_k]$. This is the case normally considered for example in NMR, as the approximation of (16) with a product of exponentials is acceptable if the pulses are disjoint in time. In particular, a sequence XY of pulses (i.e. first a pulse along Y, then along X) with $\sigma_1 = \sigma_2 = 1$, $A_1 = A_2 = 1$

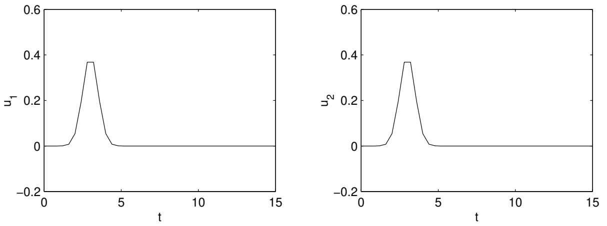


Figure 1: Case I: u_1 and u_2 coordinates

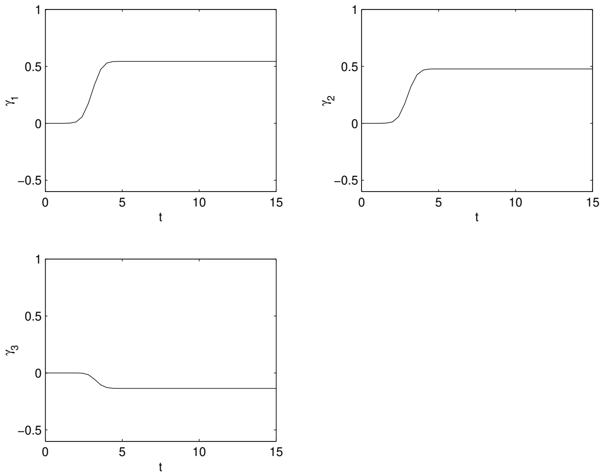


Figure 2: Case I: γ_1, γ_2 and γ_3 coordinates

and respectively centered at $\tau_1 = 9$ and $\tau_2 = 3$ implies that

$$\begin{aligned}\vec{\rho}(t) &= T\exp\left(\int_0^t \left(\mathrm{ad}_{u_1(\tau)A_1} + \mathrm{ad}_{u_2(\tau)A_2}\right) d\tau\right)\vec{\rho}(0) \\ &\simeq \exp\left(\int_0^t \mathrm{ad}_{u_1(\tau)A_1} d\tau\right)\exp\left(\int_0^t \mathrm{ad}_{u_2(\tau)A_2} d\tau\right)\vec{\rho}(0) = e^{\gamma_1 \mathrm{ad}_{A_1}} e^{\gamma_2 \mathrm{ad}_{A_2}}\vec{\rho}_0\end{aligned}$$

In this case, the coordinate along the Z direction remains constantly zero also in the product of exponentials coordinates. See Fig. 3 and Fig. 4 for the values of the time-varying parameters u_j and γ_j.

However, the Wei-Norman formula is a coordinate dependent expression and also the order in which the basis elements are taken in (10) matters. For example, if instead of the XY sequence of Fig. 3 and Fig. 4 we apply the same pulses but in the opposite order (YX: first along X then along Y) then the result changes when represented in the basis ordering given by the cardinality, as in (10) and (17). The

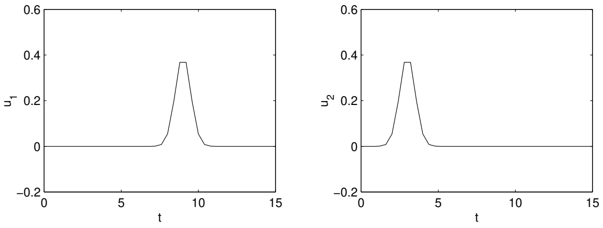


Figure 3: Case II: u_1 and u_2 coordinates

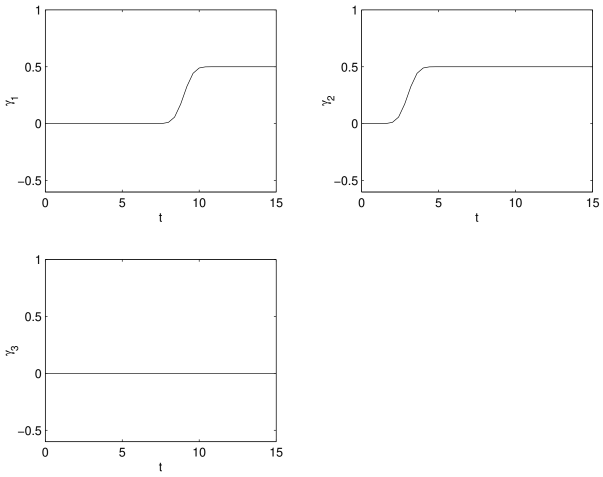


Figure 4: Case II: γ_1, γ_2 and γ_3 coordinates

time-evolutions of the γ_j in this case are represented in Fig. 6. As can be seen, also the γ_3 component becomes nonnull. Obviously, when a constant magnetic field is applied along the Z direction, H splits into constant and time-varying parts: $-iH = \bar{u}_3 A_3 + (u_1(t)A_1 + u_2(t)A_3)$, $\bar{u}_3 = \text{const}$, and an interaction representation has to be used to recover the results.

4.3. Hamiltonians from sequences of pulses: an outlook

The common way to reconstruct a time-varying Hamiltonian in NMR is through some form of averaging directly on the truncation of expressions like our product of exponentials [5]. When the method of this Section is applied to known sequences of pulses, i.e. to smooth time-varying coordinate functions $\gamma_j(t)$ in the product of exponentials (10), then it provides the time-depending functional expression of the parameters $u_j(t)$, from which an averaged expression for the Hamiltonian could be easily attained, with-

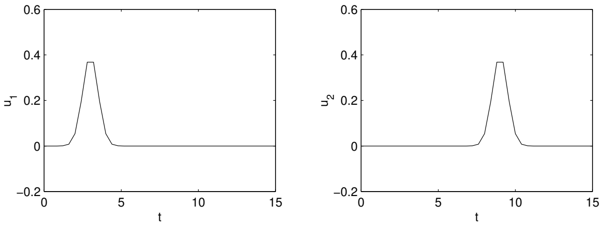


Figure 5: Case II bis: u_1 and u_2 coordinates

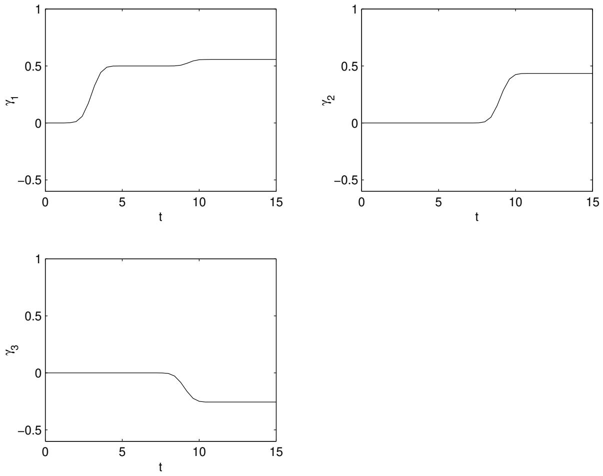


Figure 6: Case II bis: γ_1, γ_2 and γ_3 coordinates

out resorting to truncations.

5. Conclusion

If two-level systems like those of Section 4 are simple enough that explicit solutions are available [13, 15], the methodology presented here is general enough to describe realistic cases of laser-driven molecules or tensor products of nuclear spins in rf fields. It could be used to simulate the exact response of the systems to pulse shaping and for their coherent control.

References

[1] Agrachev A. and Gamkrelidze R. V., *Exponential representation of flows and a chronological enumeration*, Mat. Sb. **107** (1978), 467–532.

[2] ALTAFINI C., *On the generation of sequential unitary gates from continuous time Schrödinger equations driven by external fields*, Quantum Information Processing **1** (3) (2002), 207–224.

[3] ALTAFINI C., *Parameter differentiation and quantum state decomposition for time varying Schrödinger equations*, Reports on Mathematical Physics **52** (3) (2003), 381–400.

[4] DYSON F.J., *The S matrix in quantum electrodynamics*, Phys. Rev. **75** (1949), 1736–1755.

[5] ERNST R.R., BODENHAUSEN G. AND WOKAUN A., *Principles of magnetic resonance in one and two dimensions*, Oxford Science publications, 1987.

[6] FANO U., *Description of states in quantum mechanics by density matrix and operator techniques.* Reviews of Modern Physics **29** (1957), 74–93.

[7] HARRIS W.A., FILLMORE J.P AND SMITH D., *Matrix exponentials - another approach.* SIAM Review, **43** (4) (2001), 694–706.

[8] ISERLES A., MUNTHE-KAAS H.Z., NØRSETT S.P. AND ZANNA A., *Lie-group methods*, Acta Numerica **9** (2000), 215–365.

[9] LEONARD I., *The matrix exponential*, SIAM Review **38** (3) (1996), 507–512.

[10] LICHTENBERG D.B., *Unitary symmetry and elementary particles*, Academic Press, 1978.

[11] MEHRING M., *High resolution NMR in solids*, Springer-Verlag, 1976.

[12] MOLER C. AND VAN LOAN C., *Nineteen dubious ways to compute the exponential of a matrix*, SIAM Review **20** 4 (1978), 801–836.

[13] MURIEL A., *Driven two-level atom*, Phys. Rev. A, **50** (1994), 4286–4292.

[14] OWREN B. AND MARTHINSEN A., *Integration methods based on canonical coordinates of the second kind*, Numer. Math. **87** (4) (2001), 763–790.

[15] RAU A.R.P., *Unitary integration of Liouville-Bloch equations*, Physical Review Letters **81** (22) (1998), 4685–4789.

[16] ROYER A., *Driven two-level atom: a simplified approach.* Phys. Rev. A, **54** (1996), 3685–3686.

[17] SHADWICK B. AND BUELL W. F., *Unitary integration: a numerical technique preserving the structure of the quantum Liouville equation*, Physical Review Letters **79** (26) (1997), 5189–5193.

[18] WEI J. AND NORMAN E., *On the global representations of the solutions of linear differential equations as a product of exponentials.* Proc. of the Amer. Math. Soc. **15** (1964), 327–334.

[19] WILCOX R., *Exponential operators and parameter differentiation in quantum physics.* Journal of Mathematical Physics **8** (4) (1967), 962–982.

AMS Subject Classification: 81-08, 81Q05.

Claudio ALTAFINI, SISSA-ISAS, International School for Advanced Studies, via Beirut 2-4, 34014 Trieste, ITALY
e-mail: `altafini@sissa.it`

Rend. Sem. Mat. Univ. Pol. Torino - Vol. 63, 4 (2005)
Control Theory and Stabil., I

P. Bettiol

A REDUCTION METHOD IN OPTIMAL CONTROL FOR THE MAYER PROBLEM

Abstract. The Mayer Problem is treated looking for extremals in $W^{1,2}$ and using controls in L^2. We face the original problem finding critical points of the Action Functional related to the the pre-Hamiltonian $h = h(x, p, u)$. In this approach we show how it is possible to apply the Amann-Conley-Zehnder reduction involving not only the velocity of the curves, but also the controls: this permits us to study the solutions in terms of truncated Fourier series.

1. Introduction

We start from a classical model which is given by a control system

$$\dot{x} = f(x(t), u(t)),$$

where

$$f : \begin{array}{ccc} \mathbb{R}^n \times \mathbb{R}^m & \longrightarrow & \mathbb{R}^n \\ (x, u) & \longmapsto & f(x, u) \end{array}$$

is a $\mathcal{C}^2$ function in all variables. Let us consider the following set of *admissible controls*

$$\mathcal{U} = \{u : [0, T] \longrightarrow \mathbb{R}^m \text{ s.t. } u(\cdot) \text{ bounded and measurable}\}.$$

Given an initial condition $x_0 \in \mathbb{R}^n$ and a control $u(\cdot) \in \mathcal{U}$, we denote by $t \mapsto x(t, u(\cdot))$ the unique Carathéodory solution of the Cauchy problem

$$\begin{cases} \dot{x}(t) = f(x(t), u(t)) \\ x(0) = x_0. \end{cases} \tag{1}$$

For the existence and uniqueness of the Carathéodory solution of (1) we refer the reader, for example, to [10], [14] or [15].

Once we choose a function $\Psi \in \mathcal{C}^3(\mathbb{R}^n, \mathbb{R})$, we define the *cost functional* in the following way

$$\forall u(\cdot) \in \mathcal{U} \qquad J(u(\cdot)) := \Psi(x(T, u(\cdot))).$$

To solve the Mayer Problem with free terminal point (and with final time T) means that one has to find an optimal control $u^*(\cdot)$ which maximizes the functional $J(\cdot)$ among all $u(\cdot) \in \mathcal{U}$. The corresponding solution of the Cauchy problem (1), $x^*(t) = x(t, u^*(\cdot))$, is called optimal trajectory.

Let us consider the Hamiltonian function

$$\mathcal{H}(x, p, u) = \langle p, f(x, u)\rangle,$$

where $\langle p, f(x,u)\rangle := p \cdot f(x,u)$ is the usual scalar product in $\mathbb{R}^n$. The function $\mathcal{H} = \mathcal{H}(x, p, u)$ is also called pre-Hamiltonian in order to distinguish it from the maximized Hamiltonian given by $H(x, p) = \sup_{\omega\in\mathbb{R}^m} \mathcal{H}(x, p, \omega)$. But, here, we deal only with $\mathcal{H} = \mathcal{H}(x, p, u)$, that we will call simply Hamiltonian.

By using the Pontryagin Maximum Principle (see for instance [14] or more recent books [2] and [12]), we can associate a Hamiltonian system with boundary conditions mixed in time to the function $\mathcal{H} = \mathcal{H}(x, p, u)$ (cf. Section 2). The extremals satisfy this Hamiltonian system. Our purpose is to find the solutions of such system by studying the critical points of the Action Functional related to the Hamiltonian (see Section 3)

$$\int_0^T [p\dot{x} - \mathcal{H}(x, p, u)]\, dt.$$

In particular, these solutions satisfy the condition $\frac{\partial\mathcal{H}}{\partial u}(x, p, u) = 0$, which is weaker than the so-called maximality condition (see Section 2).

Our main idea is to follow techniques which are very common in symplectic geometry and mainly due to C. Viterbo (see [17], [18], [19] and cf. also [1]). The same techniques are applied in optimal control problems by the author and by F. Cardin in [4]. But, in [4] the necessary conditions given by the Pontryagin Maximum Principle are used in order to obtain the Hamiltonian system connected with the maximized Hamiltonian function; moreover, the Action Functional related to the maximized Hamiltonian is the main ingredient for the construction of the generating function of the initial Lagrangian submanifold $\Lambda = \{(x(0), p(0))\} \subset T^*\mathbb{R}^n$, which collects all the initial data. This procedure moves away from the controls. Nevertheless, in our work we also want to get some information on the controls, which we explicitly handle in finding critical points of the Action Functional. Finally, in our approach we deal with the Hamiltonian function h directly, which generally has a good regularity property; instead, the maximized Hamiltonian H is usually far from being regular.

In Section 4, we apply the so-called Amann-Conley-Zehnder reduction in order to obtain a reduced problem: roughly speaking, we look for stationary points of the Action Functional in the finite dimensional space of truncated Fourier series. Not always is it possible to simplify the problem in such a way; in fact, we can do it only when we get a condition on the u-component of a fixed point map, which plays a crucial role in the reduction.

In the last Section, we discuss how we apply this approach in the class of linear quadratic (L-Q) problems in order to understand some properties about the fixed point map and the reduced Action Functional. It is interesting to notice that some L-Q problems, singular as well (i.e., such that $\frac{\partial^2\mathcal{H}}{\partial u^2}$ fails to be strictly positive), can be reduced considering only trigonometrical polynomials. This result might be useful in applications when, in particular, the state equations of an optimal control problem are given in terms of the truncated Fourier series (see e.g. [9]).

2. Preliminaries and justifications

It is well known that the necessary conditions for optimality are classically given by the Pontryagin Maximum Principle, we write it here in a simple formulation (see books [14], [2] or [12]).

Pontryagin Maximum Principle. Let $u^*(\cdot)$ be an admissible control whose corresponding trajectory $x^*(t) = x(t, u^*(\cdot))$ is optimal. Then, there exists a vector-function $p(\cdot)$, $p(t) \in \mathbb{R}^n$, which is the solution of the adjoint linear system

$$\dot{p}(t) = -\frac{\partial \mathcal{H}}{\partial x}(x^*(t), p(t), u^*(t)), \qquad p(T) = \nabla\Psi(x^*(T)), \tag{2}$$

and, moreover, the *maximality condition*

$$\mathcal{H}(x^*(t), p(t), u^*(t)) = \sup_{\omega \in \mathbb{R}^m} \mathcal{H}(x^*(t), p(t), \omega) \tag{3}$$

holds true for almost every $t \in [0, T]$.

The triple $(x(\cdot), p(\cdot), u(\cdot))$ is said to satisfy PMP or *extremal* whenever $u(\cdot)$ is an admissible control, $x(\cdot)$ is the corresponding solution of system (1) and $p(\cdot)$ is such that (2)-(3) are satisfied.

We are going to investigate the solutions of the (controlled) Hamiltonian system associated to $\mathcal{H}$

$$\begin{cases} \dot{x} = \dfrac{\partial \mathcal{H}}{\partial p}(x, p, u) \\ \dot{p} = -\dfrac{\partial \mathcal{H}}{\partial x}(x, p, u) \\ 0 = \dfrac{\partial \mathcal{H}}{\partial u}(x, p, u)\ , \end{cases} \tag{4}$$

with a boundary condition which is mixed in time, that is

$$\begin{cases} x(0) = x_0 \\ p(T) = \nabla\Psi(x(T))\ . \end{cases}$$

Notice that the solutions of system (4) satisfy the condition $\frac{\partial \mathcal{H}}{\partial u}(x, p, u) = 0$, which is weaker than the maximality one (3). Therefore, we are looking for triples of functions $(x(\cdot), p(\cdot), u(\cdot))$ in a set bigger than extremals set.

Let us denote by $\mathbb{J}$ the $(2n + m) \times (2n + m)$-matrix

$$\mathbb{J} := \begin{pmatrix} \mathbb{O} & \mathbb{I}_{n\times n} & \mathbb{O} \\ -\mathbb{I}_{n\times n} & \mathbb{O} & \mathbb{O} \\ \mathbb{O} & \mathbb{O} & \mathbb{I}_{m\times m} \end{pmatrix}.$$

Thus, the Hamiltonian system (4) can be briefly written as follows

$$\begin{pmatrix} \dot{x} \\ \dot{p} \\ 0 \end{pmatrix} = \mathbb{J}\nabla\mathcal{H}(x, p, u),$$

where

$$\nabla\mathcal{H}(x,p,u) = \begin{pmatrix} \frac{\partial\mathcal{H}}{\partial x}(x,p,u) \\ \frac{\partial\mathcal{H}}{\partial p}(x,p,u) \\ \frac{\partial\mathcal{H}}{\partial u}(x,p,u) \end{pmatrix}.$$

In order to solve our problem, we first consider a canonical transformation $(x, p) \mapsto (\tilde{x}, \tilde{p})$ in $T^*\mathbb{R}^n$ given by

$$\begin{cases} p = \tilde{p} + \nabla\Psi(x) \\ x = \tilde{x}\ , \end{cases}$$

which produces the following transformed Hamiltonian:

$$\tilde{\mathcal{H}}(\tilde{x},\tilde{p},u) = \mathcal{H}(\tilde{x},\tilde{p}+\nabla\Psi(\tilde{x}),u).$$

REMARK 1. For any fixed control $u(\cdot)$, the characteristics of the vector field associated to the Hamiltonian $\tilde{\mathcal{H}}$ coincide with the characteristics of the vector field associated to $\mathcal{H}$ up to the above-mentioned canonical transformation. It allows us to study the characteristic curves $(\tilde{x}(\cdot), \tilde{p}(\cdot))$ which end at the zero-section of $T^*\mathbb{R}^n$ for $t = T$, instead of curves $(x(\cdot), p(\cdot))$, which end at Graph$(\nabla\Psi)$ at time $t = T$ (cf. also [4]).

The new boundary conditions in the $(\tilde{x}, \tilde{p})$-coordinates become

$$\begin{cases} \tilde{p}(T) = 0 \\ \tilde{x}(0) = x_0, \end{cases}$$

while the transformed Hamiltonian system is similar:

$$\begin{pmatrix} \dot{\tilde{x}} \\ \dot{\tilde{p}} \\ 0 \end{pmatrix} = \mathbb{J}\nabla\tilde{\mathcal{H}}(\tilde{x},\tilde{p},u).$$

Notations. We drop the "tilde" from the transformed quantities in order to simplify the notations, writing (x, p) instead of $(\tilde{x}, \tilde{p})$ again.

Hence, our purpose is to find the solutions of system

$$\begin{pmatrix} \dot{x} \\ \dot{p} \\ 0 \end{pmatrix} = \mathbb{J}\nabla\mathcal{H}(x,p,u), \qquad \begin{cases} p(T) = 0 \\ x(0) = x_0. \end{cases}$$

3. The action functional

Hereafter, we denote by $h = h(x, p, u)$ a $\mathcal{C}^2$ Hamiltonian function such that $|\nabla^2 h| \leq C$ for some positive constant C. In the following two Sections, we aim at looking for solutions $(x(\cdot), p(\cdot), u(\cdot))$ of system

$$\begin{pmatrix} \dot{x} \\ \dot{p} \\ 0 \end{pmatrix} = \mathbb{J}\nabla h(x, p, u), \qquad \begin{cases} p(T) = 0 \\ x(0) = x_0. \end{cases} \tag{5}$$

where $(x(\cdot), p(\cdot))$ belong to $W^{1,2}\big((0,T), \mathbb{R}^{2n}\big)$ and the function $u(\cdot)$ is chosen in the space $L^2\big((0,T), \mathbb{R}^m\big)$. We will call (simply) controls the elements of $L^2\big((0,T), \mathbb{R}^m\big)$. The strong assumption on the Hamiltonian function h to have bounded second derivatives provides that for any fixed $u(\cdot) \in L^2\big((0,T), \mathbb{R}^m\big)$, for any starting condition (x_0, p_0), the Cauchy problem

$$\begin{cases} \dot{x}(t) = \dfrac{\partial h}{\partial p}(x(t), p(t), u(t)) \\ \dot{p}(t) = -\dfrac{\partial h}{\partial x}(x(t), p(t), u(t)) \end{cases} \qquad \begin{cases} p(0) = p_0 \\ x(0) = x_0 \end{cases}$$

admits a unique Carathéodory solution. This condition guarantees also the existence of the Gâteaux derivatives of the below-defined functionals $\mathcal{A}$ and W. Finally, it plays a crucial role in order to obtain the existence and the regularity of a fixed point map that we will define later (see Lemma 2 and Remark 4 in Section 4).

REMARK 2. Notice that in this format the controls are not necessarily bounded. In fact, we are extending into L^2 controls the problem, we stated in previous Sections. By the way, once we are able to apply the Amann-Conley-Zehnder reduction, the u-component of the solution of (5) becomes a trigonometrical polynomial, which is bounded on $[0, T]$ and, hence, it is admissible control.

Let us introduce the Action Functional related to the Hamiltonian function h:

$$\begin{aligned} &\mathcal{A} : \Gamma \longrightarrow \mathbb{R} \\ &\gamma(\cdot) \longmapsto \mathcal{A}[\gamma(\cdot)] := \int_0^T \Big[p(t) \cdot \dot{x}(t) - h\big(x(t), p(t), u(t)\big)\Big]\, dt, \end{aligned} \tag{6}$$

where

$$\Gamma := \Big\{ \gamma(\cdot) = \big(x(\cdot), p(\cdot), u(\cdot)\big) \in W^{1,2}\big((0,T), \mathbb{R}^{2n}\big) \times L^2\big((0,T), \mathbb{R}^m\big) \ : \ p(T) = 0 \Big\}.$$

Thanks to the Sobolev Inequality Theorem (see for instance [8]), for any $\gamma(\cdot) = \big(x(\cdot), p(\cdot), u(\cdot)\big) \in \Gamma$ the (x, p)-components, namely $(x(\cdot), p(\cdot))$, provide a continuous curve in the cotangent fiber bundle $T^*\mathbb{R}^n$. Notice that the condition $p(T) = 0$ is justified by Remark 1. Moreover, we get a fibration

$$\begin{aligned} \pi : \Gamma &\longrightarrow \mathbb{R}^n \\ \gamma(\cdot) &\longmapsto \pi(\gamma(\cdot)) := x(0), \end{aligned}$$

where $x(0)$ is the starting point of the curve $x(\cdot)$. Indeed, a structure of vector space on the fibers $\pi^{-1}(x(0))$ with $x(0) \in \mathbb{R}^n$ is provided by the space of derivatives of the curves $(x(\cdot), p(\cdot))$ and the space of controls $u(\cdot)$; this is well expressed by means of the following bijection:

$$\text{(7)} \qquad \begin{array}{rccc} g: & \mathbb{R}^n \times L^2 & \longrightarrow & \Gamma \\ & (x(0); \phi) & \longmapsto & g\big(x(0); (\phi_x, \phi_p, \phi_u)\big)(\cdot), \end{array}$$

where

$$\begin{array}{rccc} g\big(x(0); \phi\big): & [0, T] & \longrightarrow & \mathbb{R}^{2n+m} \\ & t & \longmapsto & \Big(x(0) + \int_0^t \phi_x(s)\, ds, -\int_t^T \phi_p(s)\, ds, C_N \phi_u(t)\Big), \end{array}$$

$L^2 := L^2\big((0, T), \mathbb{R}^{2n} \times \mathbb{R}^m\big)$, $C_N := \frac{T}{2\pi N}$ and $\phi = (\phi_x, \phi_p, \phi_u)$; then one can immediately prove that g is injective and surjective. Roughly speaking, once we fix the initial point $x(0)$, the (x, p)-components of $\gamma(\cdot)$, $(x(\cdot), p(\cdot))$, are given by integrating the velocities (ϕ_x, ϕ_p), obtaining a continuous curve ending at the zero-section of $T^*\mathbb{R}^n$ at time $t = T$; while we simply multiply the control by a suitable constant.

An important fact is that the solutions of the Hamiltonian system (5) are the stationary points of the function $\mathcal{A}$ defined above in (6). This connection is well explained by the following Lemma.

LEMMA 1. *A curve $\gamma(\cdot) \in \Gamma$ solves the Hamiltonian system (5) if and only if*

$$\delta\mathcal{A}[\gamma]\delta\gamma = 0, \qquad \forall\, \delta\gamma \in \Gamma \quad \textit{such that} \;\; \delta x(0) = 0,$$

where by δ we denote the Gâteaux derivative.

Proof. We follow a classical scheme (cf. [1], [4] or [5]); notice that here we have one term more: the derivative of h with respect to u. For any $\delta\gamma \in T_\gamma\Gamma = \Gamma$ ($\gamma =$

(x, p, u)), let us consider

$$
\begin{aligned}
\delta\mathcal{A}[\gamma]\delta\gamma &= \frac{d\mathcal{A}}{d\lambda}(\gamma + \lambda\delta\gamma)|_{\lambda=0} = \\
&= \int_0^T \left(\delta p \cdot \dot{x} + p \cdot \delta\dot{x} - \frac{\partial h}{\partial x}(\gamma) \cdot \delta x - \frac{\partial h}{\partial p}(\gamma) \cdot \delta p - \frac{\partial h}{\partial u}(\gamma) \cdot \delta u\right) dt = \\
&\quad \left(\text{integrating by parts } \int_0^T p \cdot \delta\dot{x}\, dt\right) \\
&= \int_0^T \left(\dot{x} - \frac{\partial h}{\partial p}(\gamma)\right) \cdot \delta p\, dt - \int_0^T \left(\dot{p} + \frac{\partial h}{\partial x}(\gamma)\right) \cdot \delta x\, dt + \\
&\quad - \int_0^T \frac{\partial h}{\partial u}(\gamma) \cdot \delta u\, dt + p(T) \cdot \delta x(T) - p(0) \cdot \delta x(0) = \\
&= - \int_0^T \left[\begin{pmatrix} \dot{x} \\ \dot{p} \\ 0 \end{pmatrix} - \mathbb{J}\nabla h(\gamma)\right] \cdot \delta\gamma\, ds + p(0) \cdot \delta x(0),
\end{aligned}
$$

which immediately proves the Lemma. □

Composing the bijection g defined above with the Action Functional, we obtain the functional $W = -\mathcal{A} \circ g$:

$$
\begin{array}{rccl}
W: & \mathbb{R}^n \times L^2 & \longrightarrow & \mathbb{R} \\
& \big(x(0), \phi\big) & \longmapsto & W\big(x(0), \phi\big) := -\mathcal{A} \circ g\big(x(0), \phi\big) = -\mathcal{A}[g(x(0), \phi)].
\end{array}
$$

Writing W explicitly, we have ($\gamma = (x, p, u) = g(x(0), \phi)$):

$$
\begin{aligned}
&W\big(x(0), \phi\big) = \\
&\qquad = - \int_0^T \big(p \cdot \dot{x} - h(x, p, u)\big)\, dt = \\
&\qquad = - \int_0^T \Bigg[\phi_x(t) \int_T^t \phi_p(s)\, ds + \\
&\qquad\qquad - h\left(x(0) + \int_0^t \phi_x(s)\, ds, - \int_t^T \phi_p(s)\, ds, C_N \phi_u(t)\right)\Bigg]\, dt.
\end{aligned}
$$

Let us compute the Gâteaux derivative of W with respect to ϕ:

$$
\begin{aligned}
\frac{DW}{D\phi}\,\delta\phi = &- \int_0^T \Bigg[\delta\phi_x(t) \int_T^t \phi_p(r)\, dr + \phi_x(t) \int_T^t \delta\phi_p(r)\, dr + \\
&- \frac{\partial h}{\partial x}(\gamma) \int_0^t \delta\phi_x(r)\, dr - \frac{\partial h}{\partial p}(\gamma) \int_T^t \delta\phi_p(r)\, dr - C_N \frac{\partial h}{\partial u}(\gamma)\delta\phi_u\Bigg]\, dt.
\end{aligned}
$$

By using the equality

$$
\int_0^T \delta\phi_x(t) \int_t^T \phi_p(r)\, dr\, dt = \int_0^T \phi_p(t) \int_0^t \delta\phi_x(r)\, dr\, dt,
$$

in the beginning of the expression of $DW/D\phi$, we obtain

$$
\begin{aligned}
&\frac{DW}{D\phi}\,\delta\phi = \\
&\quad = -\int_0^T \Big[-\phi_p(t)\int_0^t \delta\phi_x(r)\,dr - \phi_x(t)\int_t^T \delta\phi_p(r)\,dr + \\
&\qquad - \frac{\partial h}{\partial x}(\gamma)\Big(\int_0^t \delta\phi_x(r)\,dr\Big) - \frac{\partial h}{\partial p}(\gamma)\int_T^t \delta\phi_p(r)\,dr - C_N\frac{\partial h}{\partial u}(\gamma)\delta\phi_u\Big]\,dt = \\
&\quad = \int_0^T \left[\begin{pmatrix} \dot{x} \\ \dot{p} \\ 0 \end{pmatrix} - \mathbb{J}\nabla h(\gamma)\right]\delta\gamma\ dt,
\end{aligned}
$$

where $\delta\gamma(t) = \Big(-\int_t^T \delta\phi_p(r)\,dr, \int_0^t \delta\phi_x(r)\,dr, C_N\delta\phi_u(t)\Big) \in \Gamma$.
Therefore, we proved the following result.

PROPOSITION 1. . *Choose $x(0) \in \mathbb{R}^n$. An element $\phi \in L^2$ is a stationary point of $W(x(0),\cdot)$ if and only if $\gamma(\cdot) = g(x(0),\phi)(\cdot) \in \Gamma$ satisfies the Hamiltonian system (5).*

We conclude this Section considering the derivatives of W with respect to $x(0)$ and a consequent remark:

$$
\begin{aligned}
\frac{\partial W}{\partial x(0)}\Big|_{\frac{DW}{D\phi}=0}\delta x(0) &= -\int_0^T \Big(-\frac{\partial h}{\partial x}(\gamma)\Big)\cdot\delta x(0)\,dt = \\
&= -\int_0^T \dot{p}(t)\cdot\delta x(0)\,dt = \\
&= p(0)\cdot\delta x(0).
\end{aligned}
$$

REMARK 3. The functional W can be considered as a global generating function of $\Lambda \subset T^*\mathbb{R}^n$ with ∞-dimensional space of auxiliary parameters, where

$$
\Lambda := \left\{(x(0), p(0)) :\ x(0) \in \mathbb{R}^n,\ p(0) = \frac{\partial W}{\partial x(0)}(x(0),\phi^*),\ \frac{DW}{D\phi}(x(0),\phi^*) = 0\right\}.
$$

In fact, in the original scheme given by C. Viterbo (see [17], [18], [19] and cf. also [1], [5] and [4]) the Action Functional constitutes the main ingredient for constructing a global generating function for some Lagrangian submanifold related to a given Hamiltonian flow (see Appendix for basic definitions and properties on Lagrangian submanifolds). Notice that in singular linear quadratic problems the set $\Lambda \subset T^*\mathbb{R}^n$ might fail to be a Lagrangian submanifold; while Λ turns out to be a Lagrangian submanifold in regular L-Q problems (cf. [12] and, for regular cases, see also [4]).

4. The Amann-Conley-Zehnder reduction

The reduction method, introduced by H. Amann, C. Conley and E. Zehnder in [3] and [7], transforms an infinite dimensional variational problem involving the Action Functional into a finite dimensional one.

In the space L^2 we consider the orthonormal basis $\{e^{i\frac{2\pi k}{T}t}\}_{k\in\mathbb{Z}}$. Hence, denoting by $e_k(t) := e^{i\frac{2\pi k}{T}t}$, for all $\phi \in L^2$, we have the Fourier expansion

$$\phi(t) = \sum_{k\in\mathbb{Z}} \phi_k \, e_k(t).$$

For any $N \in \mathbb{N}$ fixed, we can define the projection operator $\mathbb{P}_N$ on the $K(n, m, N)$ central components of ϕ, where $K(n, m, N) := (2n + m)(2N + 1)$,

$$\mathbb{P}_N\phi(t) := \sum_{|k|\leq N} \phi_k \, e_k(t),$$

and the projection operator $\mathbb{Q}_N$ on the remaining infinite external components

$$\mathbb{Q}_N\phi(t) := \sum_{|k|>N} \phi_k \, e_k(t).$$

Take an element $\phi \in L^2 = \mathbb{P}_N L^2 \oplus \mathbb{Q}_N L^2$, we denote by $\mu := \mathbb{P}_N\phi$ (and by $\eta := \mathbb{Q}_N\phi$ respectively) the central (and the external respectively) components of ϕ.

We show that for a suitable N only the finite dimensional space $\mathbb{P}_N\phi$ is sufficient to find stationary points of W (and to construct a generating function of Λ); indeed, by a fixed point argument, we prove that $\mathbb{P}_N\phi$ alone uniquely determines $\mathbb{Q}_N\phi$.

LEMMA 2. *For a suitably large $N \in \mathbb{N}$ the map*

$$\begin{array}{rcll} \mathcal{G}: & \mathbb{Q}_N L^2 & \longrightarrow & \mathbb{Q}_N L^2 \\ & \eta & \longmapsto & \mathbb{Q}_N \mathbb{J} \nabla h\Big(g\big(x(0), \mu + \eta\big)\Big); \end{array} \tag{8}$$

is a contraction map, for any $x(0) \in \mathbb{R}^n$ and $\mu \in \mathbb{P}_N L^2$.

Proof. First, we recall that by the assumptions on the Hamiltonian h there exists $C > 0$

such that $|\nabla^2 h| \leq C$. For any $\eta_1, \eta_2 \in \mathbb{Q}_N L^2$, we obtain

$$
\begin{aligned}
&\Big|\Big|\mathcal{G}(\eta_2) - \mathcal{G}(\eta_1)\Big|\Big|_{L^2} = \\
&=\Big|\Big|\mathbb{Q}_N \mathbb{J}\nabla h\Big(g\big(x(0), \mu + \eta_2\big)\Big) - \mathbb{Q}_N \mathbb{J}\nabla h\Big(g\big(x(0), \mu + \eta_1\big)\Big)\Big|\Big|_{L^2} \leq \\
&\leq C\Big|\Big|g\big(x(0), \mu + \eta_2\big) - g\big(x(0), \mu + \eta_1\big)\Big|\Big|_{L^2} = \\
&= C\Big|\Big|\Big(\int_0^t \sum_{|k|>N} x_k\, e^{i\frac{2\pi k}{T}r} dr,\ -\int_t^T \sum_{|k|>N} p_k\, e^{i\frac{2\pi k}{T}r} dr,\ C_N \sum_{|k|>N} u_k e^{i\frac{2\pi k}{T}t}\Big)\Big|\Big|_{L^2} = \\
&= C\Big|\Big|\Big(T\sum_{|k|>N} \frac{e^{i\frac{2\pi k}{T}t}}{i2\pi k} x_k,\ T\sum_{|k|>N} \frac{e^{i\frac{2\pi k}{T}t}}{i2\pi k} p_k,\ C_N \sum_{|k|>N} u_k e^{i\frac{2\pi k}{T}t}\Big) + \\
&\quad - \Big(\sum_{|k|>N} \frac{1}{i2\pi k} x_k,\ \sum_{|k|>N} \frac{1}{i2\pi k} p_k,\ 0\Big)\Big|\Big|_{L^2},
\end{aligned}
$$

where $(x_k, p_k, u_k) = \eta_k$ is the k^{th} Fourier coefficient of $\eta = (x, p, u) = \eta_2 - \eta_1$.

$$
\begin{aligned}
\Big|\Big|\mathcal{G}(\eta_2) - \mathcal{G}(\eta_1)\Big|\Big|_{L^2} &\leq C\Big(C_N\big|\big|\eta\big|\big|_{L^2} + T\Big|\Big|\sum_{|k|>N} \frac{(x_k, p_k, 0)}{i2\pi k}\Big|\Big|_{L^2}\Big) \leq \\
&\leq C\Big(C_N\big|\big|\eta\big|\big|_{L^2} + \big|\big|\langle (x, p, 0), \mathbb{Q}_N \mathrm{id}_{[0,T]}\rangle_{L^2}\big|\big|_{L^2}\Big) \leq \\
&\leq C\Big(C_N\big|\big|\eta\big|\big|_{L^2} + \big|\big|\eta\big|\big|_{L^2}\ \big|\big|\mathbb{Q}_N \mathrm{id}_{[0,T]}\big|\big|_{L^2}\Big) \leq \\
&\leq C\Big(C_N\big|\big|\eta\big|\big|_{L^2} + \frac{T}{2\pi N}\sqrt{2N}\big|\big|\eta\big|\big|_{L^2}\Big) \leq \\
&\leq CC_N(1 + \sqrt{2N})\big|\big|\eta\big|\big|_{L^2} = \\
&= CC_N(1 + \sqrt{2N})\big|\big|\eta_2 - \eta_1\big|\big|_{L^2}.
\end{aligned}
$$

Hence, we get a contraction if we choose N such that

$$CC_N(1 + \sqrt{2N}) = \frac{TC}{2\pi N}(1 + \sqrt{2N}) < 1.$$

□

By the Banach-Caccioppoli contraction Lemma (see for example [15] or [10]) applied to $\mathcal{G}$ defined in (8), once we choose $x(0) \in \mathbb{R}^n$ and $\mu \in \mathbb{P}_N L^2$, we obtain one and only one fixed point of $\mathcal{G}$, denoted by $q(x(0), \mu)$, that satisfies

$$q(x(0), \mu) = \mathbb{Q}_N \mathbb{J}\nabla h(g(x(0), \mu + q(x(0), \mu))). \tag{9}$$

REMARK 4. Thanks to the fact that the Hamiltonian function $h \in \mathcal{C}^2$ and by using the implicit function (Dini) Theorem, the fixed point map

$$
\begin{array}{rlcl} q: & \mathbb{R}^n \times \mathbb{P}_N L^2 & \longrightarrow & \mathbb{Q}_N L^2 \\ & (x(0), \mu) & \longmapsto & q(x(0), \mu), \end{array} \tag{10}
$$

is continuously differentiable (see [1] or [4]).

Now, suppose that the u-component of the fixed point map $q(x(0), \mu)$ vanishes, namely $q_u(x(0), \mu) \equiv 0$. Then, this allows us to restrict our functional W to a finite-dimensional space of auxiliary parameters, $\mathbb{P}_N L^2 \cong \mathbb{R}^{K(n,m,N)}$. Indeed, if $x(0) \in \mathbb{R}^n$ and $\mu \in \mathbb{P}_N L^2$ are fixed, we can consider the curve

$$
\gamma(\cdot) = (x(\cdot), p(\cdot), u(\cdot)) = g(x(0), \mu + q(x(0), \mu))(\cdot) \in \Gamma,
$$

such that

$$
\begin{pmatrix} \dot{x} \\ \dot{p} \\ \frac{u}{C_N} \end{pmatrix} = \mu + q(x(0), \mu) = \begin{pmatrix} \mu_x + q_x(x(0), \mu) \\ \mu_p + q_p(x(0), \mu) \\ \mu_u + q_u(x(0), \mu) \end{pmatrix},
$$

where q_x, q_p and q_u are the x, p and u-components of $q(x(0), \mu)$ respectively. In particular, if $q_u(x(0), \mu) = 0$, notice that the equation

$$
\mathbb{Q}_N \left[\begin{pmatrix} \dot{x} \\ \dot{p} \\ 0 \end{pmatrix} - \mathbb{J}\nabla h(\gamma) \right] = 0 \tag{11}
$$

is satisfied by $\gamma(\cdot)$, because

$$
\mathbb{Q}_N(\mathbb{J}\nabla h(\gamma)) = q(x(0), \mu) = \mathbb{Q}_N \begin{pmatrix} \dot{x} \\ \dot{p} \\ 0 \end{pmatrix}.
$$

We summarize the result provided by the reduction machinery: let us consider a solution of system

$$
\mu = \mathbb{P}_N \mathbb{J}\nabla h\Big(g\big(x(0), \mu + q(x(0), \mu)\big)\Big) \tag{12}
$$

in the unknowns $\mu \in \mathbb{P}_N L^2 \cong \mathbb{R}^{k(n,m,N)}$; assume that the u-component of the fixed point map is zero, $q_u(x(0), \mu) = 0$, then we automatically obtain the solution of the projection of the Hamiltonian system (5) on $\mathbb{Q}_N L^2$ (thanks to (11)). Therefore, the curve $\gamma(\cdot) := g(x(0), \mu + q(x(0), \mu))(\cdot)$ solves the Hamiltonian system (5) (with the boundary conditions $x(0) = x_0$ and $p(T) = 0$).

For $K := K(n, m, N)$ where N is determined as in Lemma 2, we define the function

$$
\begin{array}{rlcl} \mathcal{F}: & \mathbb{R}^n \times \mathbb{R}^K & \longrightarrow & \mathbb{R} \\ & \big(x(0), \mu\big) & \longmapsto & \mathcal{F}(x(0), \mu) := W\big(x(0), \mu + q(x(0), \mu)\big). \end{array}
$$

The Amann-Conley-Zehender reduction permits us to find solutions of our original system (5) by studying critical points of function $\mathcal{F}$. We express also our main result in terms of generating functions; this well summarizes all richness of structure of the function $\mathcal{F}$.

THEOREM 1. *Let us suppose that $h \in \mathcal{C}^2$ and $|\nabla^2 h| \leq C$ for a positive constant C. If $\frac{\partial \mathcal{F}}{\partial \mu}(x(0), \mu) = 0$ and $q_u(x(0), \mu) = 0$, then we also obtain $\frac{DW}{D\phi}(x(0), \mu + q(x(0), \mu)) = 0$. Moreover, if the u-component of the fixed point q is identically zero, namely $q_u(x(0), \mu) \equiv 0$, then the function $\mathcal{F} = \mathcal{F}(x(0), \mu)$ is a (global) generating function of $\Lambda = \{(x(0), p(0))\} \subset T^*\mathbb{R}^n$ if and only if the functional $W(x(0), \phi)$ is a (global) generating function of Λ (with ∞-dimensional space of parameters).*

Proof. We use a classical argument based on the Amann-Conley-Zehender reduction. First of all let us compute the derivative with respect to μ:

(13)
$$\begin{aligned}\frac{\partial \mathcal{F}}{\partial \mu}(x(0), \mu) =& \frac{DW}{D\phi}\left(\frac{D\phi}{D\mu} + \frac{D\phi}{D\eta}\frac{Dq}{D\mu}\right) = \\ =& -\int_0^T \mathbb{P}_N \left[\left(\begin{array}{c}\dot{x}\\ \dot{p}\\ 0\end{array}\right) - \mathbb{J}\nabla h(\gamma)\right]\Bigg|_{\gamma = g(x(0), \mu + q(x(0),\mu))} dt + \\ & -\int_0^T \mathbb{Q}_N \left[\left(\begin{array}{c}\dot{x}\\ \dot{p}\\ 0\end{array}\right) - \mathbb{J}\nabla h(\gamma)\right]\Bigg|_{\gamma = g(x(0), \mu + q(x(0),\mu))} \frac{Dq}{D\mu}\, dt.\end{aligned}$$

The second integral in (13) vanishes by the properties of the fixed point $q(x(0), \mu)$ (cf. (11)). Hence, we get

$$\frac{\partial \mathcal{F}}{\partial \mu}(x(0), \mu) = -\int_0^T \mathbb{P}_N \left[\left(\begin{array}{c}\dot{x}\\ \dot{p}\\ 0\end{array}\right) - \mathbb{J}\nabla h(\gamma)\right]\Bigg|_{\gamma = g(x(0), \mu + q(x(0),\mu))} dt$$

Similarly, deriving $\mathcal{F}$ with respect to $x(0)$, we obtain

$$\begin{aligned}&\frac{\partial \mathcal{F}}{\partial x(0)}(x(0), \mu) = \\ &\quad = \frac{\partial W}{\partial x(0)} + \frac{DW}{D\phi}\frac{D\phi}{D\eta}\frac{Dq}{Dx(0)} = \\ &\quad = \frac{\partial W}{\partial x(0)}(x(0), \phi)\Big|_{\gamma = g(x(0), \mu + q(x(0),\mu))} + \\ &\quad\quad -\int_0^T \mathbb{Q}_N \left[\left(\begin{array}{c}\dot{x}\\ \dot{p}\\ 0\end{array}\right) - \mathbb{J}\nabla h(\gamma)\right]\Bigg|_{\gamma = g(x(0), \mu + q(x(0),\mu))} \frac{Dq}{Dx(0)}\, dt = \\ &\quad = \frac{\partial W}{\partial x(0)}(x(0), \phi)\Big|_{\phi = \mu + q(x(0),\mu)}.\end{aligned}$$

We conclude observing that if $(x(0), \phi) \in \mathbb{R}^n \times L^2$ satisfies the system

$$(14) \qquad \begin{cases} p(0) = \dfrac{\partial W}{\partial x(0)}(x(0), \phi) \\ \quad 0 = \dfrac{DW}{D\phi}(x(0), \phi) \,, \end{cases}$$

then, considering the projection $\mu = \mathbb{P}_N \phi$, the couple $(x(0), \mu) \in \mathbb{R}^n \times \mathbb{R}^K$ satisfies

$$\begin{cases} p(0) = \dfrac{\partial \mathcal{F}}{\partial x(0)}(x(0), \mu) \\ \quad 0 = \dfrac{\partial \mathcal{F}}{\partial \mu}(x(0), \mu) \,. \end{cases}$$

Vice versa, if one computes the solution of (26), $(x(0), \mu) \in \mathbb{R}^n \times \mathbb{R}^K$, then, completing μ with $q(x(0), \mu)$ in $\phi = \mu + q(x(0), \mu)$, the couple $(x(0), \phi) \in \mathbb{R}^n \times L^2$ solves (14). $\square$

REMARK 5. i) The condition $q_u(x(0), \mu) = 0$ in the hypothesis of Theorem 1 is strong, but it is satisfied in very simple examples, for instance some linear quadratic cases (see the next section). By using the Theorem above, we are able to study our Mayer Problem only considering a suitable truncation of Fourier series as far as it concerns the control parameters and also the derivatives of state variables (instead of whole L^2). Moreover, the controls turn out to be admissible because they are trigonometrical polynomials (cf. Remark 2).

ii) Suppose that $\mu \mapsto \mathcal{F}(x(0), \mu)$ is weakly quadratic at infinity, namely out of a compact set it is a quadratic form (even degenerate), then we can apply the Ljusternik-Schnirelman theory, which is a powerful tool to get lower bounds on the critical points of a given function (see for example the books [1] or [16] for general theory and [4] for the degenerate case).

5. Linear quadratic examples

In this Section, we apply our results to the well known linear quadratic (L-Q) optimal problem. Let us consider the linear control system

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ x(0) = x_0 \end{cases}$$

and the running cost

$$\ell(x, u) = \langle Px, u \rangle + \frac{1}{2}\langle Ru, u \rangle + \frac{1}{2}\langle Qx, x \rangle,$$

where $x \in \mathbb{R}^n,\ u \in \mathbb{R}^m$ and A, B, P, Q, R are constant matrices; in particular, Q and R are symmetric. We underline the fact that here R can be any definite non-negative matrix, so also the *singular case* is included. Here, we have not a final target for the controlled trajectory $x(t, u(\cdot))$. We are going to minimize the functional

$$\int_0^T \ell(x(t,u), u(t))\, dt\ .$$

It is well known that the above Bolza Problem can be recast in a Mayer form, introducing the auxiliary variable

$$x_{n+1}(t) = \int_0^t \ell(x(s,u), u(s))\, ds\ ,$$

and defining $\Psi(x_1, \ldots, x_n, x_{n+1}) = -x_{n+1}$. Hence, we have a $(n+1)$-dimensional system

$$\begin{cases} (\dot{x}_i)_{i\in\{1,\ldots,n\}} = Ax + Bu \\ \qquad\quad \dot{x}_{n+1} = \ell(x,u)\ . \end{cases}$$

In this case the Hamiltonian is

$$h(x,p,u) = \langle p, Ax\rangle + \langle p, Bu\rangle - (\langle Px, u\rangle + \frac{1}{2}\langle Ru, u\rangle + \frac{1}{2}\langle Qx, x\rangle),$$

hence

$$\nabla h(x,p,u) = \begin{pmatrix} \frac{\partial h}{\partial x}(x,p,u) \\ \frac{\partial h}{\partial p}(x,p,u) \\ \frac{\partial h}{\partial u}(x,p,u) \end{pmatrix} = \begin{pmatrix} pA - uP - Qx \\ Ax + Bu \\ pB - Px - Ru \end{pmatrix}.$$

Take $N \in \mathbb{N}$ as in Lemma 2, once $x(0)$ and μ are chosen, then the fixed point map defined in (9)-(10) has the property

$$q(x(0), \mu) = \mathbb{Q}_N \mathbb{J} \nabla h(g(x(0), \mu + q(x(0), \mu))).$$

Let us denote by η_x, η_p and η_u the components of the fixed point $q(x(0), \mu)$

$$q(x(0), \mu) = \begin{pmatrix} \eta_x \\ \eta_p \\ \eta_u \end{pmatrix} \in \mathbb{Q}_N L^2,$$

then we obtain a curve $\gamma(\cdot) = (x(\cdot), p(\cdot), u(\cdot)) \in \Gamma$ by means of the function g, defined in (7):

$$\gamma(\cdot) = g(x(0), \mu + q(x(0), \mu))(\cdot) \in \Gamma.$$

Moreover, we can explicitly write the components of $\gamma(\cdot)$ as follows

$$\begin{cases} x(t) = x(0) + \displaystyle\int_0^t (\mu+\eta)_x(s)\,ds \\ p(t) = -\displaystyle\int_t^T (\mu+\eta)_p(s)\,ds \\ u(t) = C_N(\mu+\eta)_u, \end{cases}$$

namely

$$\begin{cases} x(t) = \hat{x}_0 + T \displaystyle\sum_{0\neq|k|\leq N} \frac{\mu_{x,k}+\mu_{x,0}}{i2\pi k} e_k(t) + T\sum_{|k|>N} \frac{\eta_{x,k}+\mu_{x,0}}{i2\pi k} e_k(t) \\ p(t) = \hat{p}_0 + T \displaystyle\sum_{0\neq|k|\leq N} \frac{\mu_{p,k}+\mu_{p,0}}{i2\pi k} e_k(t) + T\sum_{|k|>N} \frac{\eta_{p,k}+\mu_{p,0}}{i2\pi k} e_k(t) \\ u(t) = C_N(\displaystyle\sum_{|k|\leq N} \mu_{u,k}e_k(t) + \sum_{|k|>N} \eta_{u,k}e_k(t)), \end{cases}$$

where $(\mu_{x,k}, \eta_{p,k}, \mu_{u,k}) = \mu_k$ and $(\eta_{x,k}, \eta_{p,k}, \eta_{u,k}) = \eta_k$ are the k^{th} Fourier coefficients of μ and η respectively (here we have to use a more detailed notation with respect to that we used in the proof of Lemma 2); $\hat{x}_0$ and $\hat{p}_0$ are suitable real numbers and recall that $e_k(t) = e^{i\frac{2\pi k}{T}t}$. By simple computations we get

$$\mathbb{Q}_N \mathbb{J}\nabla h(\gamma) =$$
$$\begin{pmatrix} A(T\sum_{|k|>N}\frac{\eta_{x,k}+\mu_{x,0}}{i2\pi k}e_k) + B(C_N\sum_{|k|>N}\eta_{u,k}e_k) \\ T\Big(Q(\sum_{|k|>N}\frac{\eta_{x,k}+\mu_{x,0}}{i2\pi k}e_k) - (\sum_{|k|>N}\frac{\eta_{p,k}+\mu_{p,0}}{i2\pi k}e_k)A\Big) + (C_N\sum_{|k|>N}\eta_{u,k}e_k)P \\ T\Big((\sum_{|k|>N}\frac{\eta_{p,k}+\mu_{p,0}}{i2\pi k}e_k)B - P(\sum_{|k|>N}\frac{\eta_{x,k}+\mu_{x,0}}{i2\pi k}e_k)\Big) - R(C_N\sum_{|k|>N}\eta_{u,k}e_k) \end{pmatrix}.$$

REMARK 6. Notice that in general $(\eta_x, \eta_p, \eta_u) = 0$ is not a fixed point for the map $\mathcal{G}$ defined in (8). But, if $\mu_{x,0} = 0$ and $\mu_{p,0} = 0$ (the so-called zero mean case), then, by the linearity, we obtain that the fixed point vanishes:

$$q(x_0, \mu) = 0.$$

EXAMPLE 1. In order to understand some properties on the fixed point map and on the critical points of W and $\mathcal{F}$, let us consider a very simple control system in $\mathbb{R}^2$ with $u = (u_1, u_2) \in \mathbb{R}^2$:

$$\begin{cases} \dot{x}_1 = u_1 \\ \dot{x}_2 = x_2 + u_2. \end{cases}$$

Once we choose a starting point $x(0) = x_0 \in \mathbb{R}^2$, we want to minimize the functional

$$\frac{1}{2}\int_0^T (u_1^2(t) - x_1^2(t))\,dt,$$

where $T = \pi$. The Hamiltonian function is

$$h(x, p, u) = p_2 x_2 + p_1 u_1 + p_2 u_2 - \frac{1}{2}u_1^2 + \frac{1}{2}x_1^2,$$

while the (controlled) Hamiltonian system related to h is given by

$$\begin{cases} \dot{x}_1 = u_1 \\ \dot{x}_2 = x_2 + u_2 \\ \dot{p}_1 = -x_1 \\ \dot{p}_2 = -p_2 \\ u_1 = p_1 \\ p_2 = 0, \end{cases}$$

with the boundary condition

$$\begin{cases} x(0) = x_0 \\ p(T) = 0. \end{cases}$$

It is straightforward to see that for any starting point $x(0) \in \mathbb{R}^2$, called $\mu \in P_N L^2$ the solution of (12), the u-component of the fixed point map vanishes: $q_u(x(0), \mu) \equiv 0$. Instead, the x_2-component of $q(x(0), \mu)$ is an infinite series, therefore $q \neq 0$. Notice that to solve the reduced problem, namely to find stationary points of $\mathcal{F}$, implies getting solutions of the ∞-dimensional problem (for W) or, equivalently, for the Hamiltonian system; but the opposite implication is not true: for any $u_2(\cdot) \in L^2((0, T), \mathbb{R})$ (not only for $u_2(\cdot) \in \mathbb{P}_N L^2((0, T), \mathbb{R})$) we have a solution of the Hamiltonian system. Finally, we underline the fact that for $x(0) = 0$ the function $\mu \mapsto \mathcal{F}(0, \mu)$ is a quadratic form degenerate with respect to the u_2-component (cf. ii) of Remark 5).

6. Appendix: Generating functions of Lagrangian submanifolds and symplectic structures

We recall some basic definitions and results which concern the Lagrangian submanifolds of the cotangent fiber bundle $T^*\mathbb{R}^n$ (cf. [20]). A differentiable manifold $\Lambda \subset T^*\mathbb{R}^n$ is Lagrangian if the following conditions hold true

1. $\dim \Lambda = n$
2. $\omega_{\mathbb{R}^n}|_\Lambda = 0,$

where $\omega_{\mathbb{R}^n} = d\theta_{\mathbb{R}^n}$ ($\theta_{\mathbb{R}^n}$ is the canonical 1-form of Liouville); in local coordinates we have $\theta_{\mathbb{R}^n} = \sum_{i=1}^n p_i dx^i$ and the 2-form $\omega_{\mathbb{R}^n} = dp \wedge dx = \sum_{i=1}^n dp_i \wedge dx^i$.
The Theorem of Maslov-Hörmander ([13], [11]) locally characterizes the Lagrangian manifolds $\Lambda \subset T^*\mathbb{R}^n$: a submanifold Λ is Lagrangian if and only if Λ is described by means of (local) functions $(x, v) \mapsto S(x, v)$, $S \in \mathcal{C}^2(\mathbb{R}^n \times \mathbb{R}^k, \mathbb{R})$ such that

$$(15) \qquad \Lambda = \left\{ (x,p) : \ p = \frac{\partial S}{\partial x}(x,\bar{v}), \quad 0 = \frac{\partial S}{\partial v}(x,\bar{v}) \ \ \exists\, \bar{v} \in \mathbb{R}^k \right\},$$

with the rank condition

$$(16) \qquad \mathrm{rk}\left(\frac{\partial^2 S}{\partial x \partial v}, \frac{\partial^2 S}{\partial v \partial v} \right)\Bigg|_{\left\{\frac{\partial S}{\partial v}=0\right\}} = \max = k.$$

Functions S satisfying (15)-(16) are called Morse Families for Λ; instead, we call S a *generating function* of Λ if (15) holds, but not necessarily (16).

Acknowledgments. I would like to thank F. Cardin, who introduced me in the subject, and the referee for his precious suggestions.

References

[1] AEBISCHER B. AND AL., *Symplectic geometry*, Progress in Mathematics **124**, Birkhäuser, Basel 1992.

[2] AGRACHEV A. AND SACHKOV YU.L., *Control theory from the geometric viewpoint*, Springer-Verlag, Berlin 2004.

[3] AMANN H. AND ZEHNDER E., *Periodic solutions of asymptotically linear Hamiltonian systems*, Manus. Math. **32** (1980), 149–189.

[4] BETTIOL P. AND CARDIN F., *Lagrangian submanifold landscapes of necessary conditions for maxima in optimal control: global parameterizations and generalized solutions*, Sovremennaya Matermatika I Ee Prilozheniya Prilozheniya (Contemporary Mathematics and its Applications) **21** (2004) (in russian), to appear in Journal of Mathematical Sciences.

[5] CARDIN F., *The global finite structure of generic envelope loci for Hamilton-Jacobi equations*, J. of Mathematical Physics **43** (1) (2002), 417–430.

[6] CARDIN F., *On viscosity and geometrical solutions of Hamilton-Jacobi equations*, Nonlinear Analysis, T.M.A. **20** (1993), 713–719.

[7] CONLEY C. AND ZEHNDER E., *Morse type index theory for flows and periodic solutions for Hamilton equations*, Comm. Pure Appl. Math. **37** (1984), 207–253.

[8] EVANS L.C., *Partial differential equations*, Graduate Studies in Mathematics **19** AMS, Providence R.I. 1998.

[9] ENDOW Y., *Optimal control via Fourier series of operational matrix of integration.*, IEEE Trans. Automat. Control **34** (7) (1989), 770–773.

[10] HALE J.K., *Ordinary differential equations*, second edition, Robert E. Krieger Publishing Co., Inc., Huntington, New York 1980.

[11] HÖRMANDER L., *Fourier integral operators I*, Acta Math. **127** (1971), 79–183.

[12] JURDJEVIC V., *Geometric control theory*, Cambridge University Press, Cambridge 1997.

[13] MASLOV V.P., *Théorie des perturbations et méthodes asymptotiques*, Editions de l'Université de Moscou, 1965 (russian version), Dunod-Gauthier-Villars, Paris 1971 (French version).

[14] PONTRYAGIN L.S., BOLTYANSKII V.G., GAMKRELIDZE R.V. AND MISCHENKO E.F., *The mathematical theory of optimal processes*, Wiley, New York 1962.

[15] SANSONE G. AND CONTI R. , *Non-linear differential equations*, International Series of Monographs in Pure and Applied Mathematics **67**, A Pergamon Press Book, The Macmillan Co., New York 1964.

[16] STRUWE M., *Variational methods. Applications to nonlinear partial differential equations and Hamiltonian systems.* Springer-Verlag, Berlin 1990.

[17] VITERBO C., *Intersection de sous-variétés lagrangiennes, fonctionnelles d'action et indice des systèmes hamiltoniens.* Bull. Soc. Math. France **115** (3) (1987), 361–390.

[18] VITERBO C., *Recent progress in periodic orbits of autonomous Hamiltonian systems and applications to symplectic geometry*, Lecture Notes in Pure and Appl. Math. **121**, Dekker, New York 1990, 227-250.

[19] VITERBO C., *Solutions of Hamilton-Jacobi equations and symplectic geometry*, addendum to: *Séminaire sur les Équations aux Dérivés Partielles. 1994-1995* Séminaire sur les Équations aux Dérivées Partielles, 1995-1996, École Polytech., Palaiseau 1996.

[20] WEINSTEIN A., *Lectures on symplectic manifolds.*, C.B.M.S. Conf. Series Amer. Math. Soc. **29**, Providence R.I. 1977.

AMS Subject Classification: 49K99, 49N10, 53D99, 93C15.

Piernicola BETTIOL, S.I.S.S.A. - I.S.A.S., Scuola Internazionale Superiore di Studi Avanzati - International School for Advanced Studies, via Beirut 2-4, 34013 Trieste, ITALY
e-mail: `bettiol@sissa.it`

Rend. Sem. Mat. Univ. Pol. Torino - Vol. 63, 4 (2005)
Control Theory and Stabil., I

M.I. Caiado* – A.V. Sarychev†

REMARKS ON STABILITY OF INVERTED PENDULA

Abstract. Using linearization principle and tools from chronological calculus one establishes a criteria for stabilization of, usually unstable, equilibrium position of Double Inverted Pendula when subject an arbitrary fast oscillation. Both, planar and spherical cases are considered.

1. Introduction

Problem of stability and stabilization of, usually unstable, upper equilibrium position of inverted pendulum has been intensively studied. Almost no bibliography is provided here. Some interesting references to the earlier work (starting from the beginning of 1900) can be found in [8]. Two publications introducing the readers into the field of vibrational control and vibrational mechanics are [5] and [11]. Here we just quote a "textbook result" from [10, Chap 5] concerning stability of an inverted pendulum of length l with vertically (harmonically) oscillating point of suspension. Stability of equilibrium position is assured whenever number of oscillations in one unit of time is greater or equal then $\frac{1}{a}\sqrt{\frac{3}{64}\omega}$, where $\omega^2 = g/l$ and a is the amplitude of oscillation of the suspension point.

In [6, 7], one of us considered problem of stability for time-periodic systems, in particular the problem of stabilizing equilibrium position of an inverted pendulum when its pivot is subject to an arbitrary fast oscillation. Conditions of stability were established there by application of technics from chronological calculus and averaging theory.

Following the techniques used in [4, 7, 6] one intends to derive similar conditions for double pendulum in planar and spherical cases.

The paper is organized as follows. Section 2 is devoted to chronological calculus and to some classical results on stability. In Section 3 equations for mechanical system (planar case) are set, the corresponding monodromy matrix is computed and the proof of main result for planar inverted pendulum is presented. The spherical case is treated in Section 4.

The authors express their gratitude to anonymous referee for various suggestions on improvement of the paper.

*Centro de Matemática da Universidade do Minho, Portugal
†Dipartimento di Matematica per le Decisione, Università di Firenze,Italia

2. Preliminaries

2.1. Variational formula of chronological calculus

Chronological calculus (see [1] and [9]) has been introduced by Agrachev and Gamkrelidze in [1] for nonlinear time-varying systems on smooth manifolds. Since the system to be considered in this work is linear, one presents some reformulations for (time-varying) linear case.

Consider the time-varying system of linear ordinary differential equations

$$\dot{z}(t) = A(t)\, z(t), \qquad z(0) = z_0 \tag{1}$$

where $z(t) \in \mathbb{R}^n$ and $A(t)$ is a matrix-valued function depending continuously on t. Considering its solutions $z(t; z_0)$ one can introduce a flow of linear maps $P^t : z_0 \mapsto z(t, z_0)$. Obviously, P^t is the unique solution to the matrix ordinary differential equation with initial condition

$$\dot{P}^t = A(t)\, P^t, \qquad P^0 = I_n.$$

Following [1], one calls the flow P^t right chronological exponential of $A(t)$ and denotes it

$$P^t = \overrightarrow{\exp} \int_0^t A(\tau)\, d\tau.$$

One can also define left chronological exponential $\overleftarrow{\exp} \int_0^t A(\tau)\, d\tau$ as solution of $\dot{P}^t = P^t\, A(t),\ P^0 = I_n$.

Using this notation, the solution of problem (1) is represented as

$$z(t) = \overrightarrow{\exp} \int_0^t A(\tau)\, d\tau\, z_0 = P^t\, z_0.$$

The flow $\overrightarrow{\exp} \int_0^t A(\tau)\, d\tau$ admits a series expansion in the form

$$\overrightarrow{\exp} \int_0^t A(\tau)\, d\tau = \mathrm{Id} + \sum_{n=1}^{\infty} \underset{0 \le \tau_n \le \ldots \le \tau_1 \le t}{\int \cdots \int} A(\tau_n) \ldots A(\tau_1)\, d\tau_n \ldots d\tau_1.$$

In (1), take $A(t) = B(t) + C(t)$, considering $B(t)$ as reference matrix of coefficients and $C(t)$ as its perturbation. There holds so called chronological calculus variational formula

$$\overrightarrow{\exp} \int_0^t (B(\tau) + C(\tau))\, d\tau = \overrightarrow{\exp} \int_0^t \left(\overrightarrow{\exp} \int_0^\tau \mathrm{ad}\, B(\theta)\, d\theta \right) C(\tau)\, d\tau \circ \circ\, \overrightarrow{\exp} \int_0^t B(\tau)\, d\tau \tag{2}$$

where “ad” is related to the matrix commutator as

$$\mathrm{ad}\, A(t)\, B(t) = -[A(t), B(t)] = B(t)\, A(t) - A(t)\, B(t)$$

and $Q^t = \overrightarrow{\exp} \int_0^t \mathrm{ad}\, B(\theta)\, d\theta\, d\tau$ is the solution of the operator differential equation

$$\frac{d}{dt}\, Q^t = Q^t \circ \mathrm{ad}\, B(t), \qquad Q^0 = Id.$$

2.2. Formal expansion for $\ln \overrightarrow{\exp} \int_0^t A_\tau d\tau$

Let $A(t) \stackrel{def}{=} A_t$ be a matrix-valued function with time-varying entries. One is interested in defining formal expansion

$$\Lambda_{0,\,t}(A_\tau) = \ln \overrightarrow{\exp} \int_0^t A_\tau d_\tau = \sum_{m=1}^{+\infty} \Lambda^{(m)}.$$

Here (see [1])

$$\Lambda^{(m)} = \int_0^t d\tau_1 \int_0^{\tau_1} d\tau_2 \ldots \int_0^{\tau_{m-1}} d\tau_m\, g_m(A_{\tau_1}, \ldots, A_{\tau_m}), \tag{3}$$

where, for each $m \geq 2$, $g_m(A_{\tau_1}, \ldots, A_{\tau_m})$ is an homogeneous polynomial of first degree in each A_{τ_i}. Moreover, it is a commutator polynomial in $A_{\tau_1}, \ldots, A_{\tau_m}$, i.e., it can be expressed as a linear combination of $A_{\tau_1}, \ldots, A_{\tau_m}$ and of their iterated commutators:

$$g_m(A_{\tau_1}, \ldots, A_{\tau_m}) = \sum_{\alpha=1}^{(2m-3)!!} b_{\upsilon_{1\alpha}} \ldots b_{\upsilon_{m\alpha}} w_\alpha(\tau_1, ..., \tau_m). \tag{4}$$

Each of w_α is an iterated Lie bracket of length $(m-1)$ of $A_{\tau_1}, \ldots, A_{\tau_m}$, υ_{ij} is the depth of A_{τ_i} in bracket w_j, $b_k = B_k/k!$, $k \geq 2$ and B_k are Bernoulli numbers.

Following [1], one briefly explains the meaning of “depth of A_{τ_i} in w_α”. Consider the symbols “ad" and $A_{\tau_1}, ..., A_{\tau_m}$ and all finite sequences of such symbols, which one calls *word* and denotes by w. One accepts repetitions of “ad” but no repetitions of $A_{\tau_1}, ..., A_{\tau_m}$ are allowed. A word is regular if by the introduction of suitable parenthesis it can be expressed as a commutator polynomial in $A_{\tau_1}, ..., A_{\tau_m}$ with the usual meaning of the symbol “ad”. Let w be a regular word. Assume, $w = \upsilon_1 ... \upsilon_l A_\tau \widetilde{w}$. Then, depth of A_τ in w is the number of regular words of the form $\upsilon_i ... \upsilon_l A_\tau$, $1 \leq i < l$. For more detailed description of polynomials g_m see [1]. Another method to compute $\ln P^t$ using the so called chronological product is presented in [2, 3].

The series

$$\Lambda_{0,\,t}(A_\tau) = \sum_{m=1}^{\infty} \Lambda^{(m)}$$

is known to be absolutely convergent for $\int_0^t \|A_\tau\|\, d\tau \leq 0.44$, [1, Prop. 5.2].

2.3. Classical results on stability

Consider linear fast-oscillating system

$$\dot{x}(t) = A(k\,t)\,x(t) \tag{5}$$

where $x \in \mathbb{R}^n$, $A(t), t \geq 0$, is matrix-valued function continuous and 1-periodic with respect to t and $k > 0$ is a large parameter.

A standard averaging result states that if all eigenvalues of corresponding averaged matrix $\int_0^1 A(t)\,dt$ are located in the open left complex half-plane then system (5) is asymptotically stable for all sufficiently large k.

Another stability condition uses monodromy matrix P^1 of time-periodic systems. If all eigenvalues of P^1 are located in the interior of the unit circle then the system is asymptotically stable; system is unstable if at least one eigenvalue lies outside the unit circle. In general it is difficult to compute spectrum of P^1.

It is more convenient to deal with the logarithm $\ln P^1$. In this case stability conditions can be formulated as: system (5) is asymptotically stable if all the eigenvalues of $\ln P^1$ are located in the open left complex half-plane and is unstable if at least one eigenvalue lies in the open right complex half-plane. System (5) is stable if all eigenvalues of $\ln P^1$ have non positive real part and purely imaginary eigenvalues are distinct.

3. Inverted double pendulum

Consider a mechanical system which consists of two mathematical inverted pendula (i.e., a pendulum in a gravitational field without friction and tension). Each pendulum is modeled by a mass point (the bob of mass m_i) and a massless beam of length r_i. The second pendulum is attached to the bob of the first one. One neglects the axial rotation of the beams, so there is one degree of freedom for each pendulum.

Assume that the pivot of first pendulum is subject to an arbitrary fast oscillation $\delta s(k\,t)$ where $\delta > 0$ is a small fixed parameter, k can be arbitrarily large and, in any case, $\delta k > 1$. Assume also that $s(t)$ is a $C^2(\mathbb{R})$ 1-periodic function and that $\dot{s}(0) = 0$. Let m_i and r_i be the mass and length of each pendulum. Consider a coordinate system with the origin at the pivot of the first pendulum. Let $\theta_i; i = 1, 2$ to be the angle between each pendulum and the positive part of vertical axis and g is the acceleration due to gravity.

From now on $\theta = (\theta_1, \theta_2)$. One uses standard matrix notation: 0_n will denote square matrix of order n with zero entries and I_n the identity matrix of order n.

The double inverted pendulum admits four equilibrium points: $(0, 0)$, $(\pi, 0)$, $(0, \pi)$ and (π, π). Upper position corresponds to $\theta = (0, 0)$. The (forced) Hamiltonian H^ℓ, corresponding to the linearized equations of the motion near this equilibrium point, is

$$H^\ell(t, p, \theta) = \frac{1}{2}\Big(\langle p, C\,p\rangle + \langle B(t)\,\theta, \theta\rangle\Big) \tag{6}$$

with $B(t) = -[g\,A + \delta k^2 \ddot{s}(kt)\,A]$, and A and C are constant matrices depending only on the parameters $m_1,\ m_2,\ r_1,\ r_2$:

$$A = \begin{pmatrix} (m_1 + m_2)r_1 & 0 \\ 0 & m_2 r_2 \end{pmatrix}, \quad C = \begin{pmatrix} (r_1^2 m_1)^{-1} & -(r_1 r_2 m_1)^{-1} \\ -(r_1 r_2 m_1)^{-1} & (m_1 + m_2)(r_2^2 m_2 m_1)^{-1} \end{pmatrix}.$$

Hamiltonian equations in matrix form are

$$\begin{pmatrix} \dot{\theta} \\ \dot{p} \end{pmatrix} = \begin{pmatrix} 0_2 & C \\ g\,A + \delta k^2 \ddot{s}(kt)\,A & 0_2 \end{pmatrix} \begin{pmatrix} \theta \\ p \end{pmatrix}. \tag{7}$$

Setting $\tau = k\,t$, denoting $z = (\theta, p)^T$ and $\dot{z} = dz/d\tau$, one transforms (7) into linear time-variant system

$$\dot{z}(\tau) = k^{-1} \begin{pmatrix} 0_2 & C \\ g\,A + \delta k^2 \ddot{s}(\tau)\,A & 0_2 \end{pmatrix} z(t)$$

or

$$\dot{z}(\tau) = Q(\tau) z(\tau). \tag{8}$$

To establish stability result one has to study monodromy matrix P^1 for linear time-varying system (8) corresponding to the Hamiltonian (6). One accomplishes this by chronological calculus variational formula introduced in Section 2.1. The main work consists of the study of the series expansion for $\ln P^1$. One studies the convergence of the series, estimates the rest term of its truncation, and derive stability result from the spectral information regarding this truncation.

3.1. Monodromy matrix

Apply variational formula (2) to the system (8) written as

$$\dot{z}(\tau) = [k^{-1} M + \delta k N(\tau)]\, z(\tau)$$

where

$$M = \begin{pmatrix} 0_2 & C \\ g\,A & 0_2 \end{pmatrix} \qquad \text{and} \qquad N(\tau) = \begin{pmatrix} 0_2 & 0_2 \\ \ddot{s}(\tau)\,A & 0_2 \end{pmatrix}.$$

One obtains

$$\begin{aligned} \overrightarrow{\exp} \int_0^1 \left(k^{-1} M + \delta k N_\tau\right) d\tau &= \overrightarrow{\exp} \int_0^1 \left(\overrightarrow{\exp} \int_0^\tau \operatorname{ad}(\delta k N_\sigma) d\sigma\; k^{-1} M \right) d\tau \circ \\ &\qquad \circ \overrightarrow{\exp} \int_0^1 \delta k N_\tau\, d\tau \\ &= \overrightarrow{\exp} \int_0^1 D_\tau\, d\tau \circ \overrightarrow{\exp} \left(\delta k \int_0^1 N_\tau\, d\tau \right). \end{aligned}$$

From 1−periodicity of $s(\cdot)$ it follows that $\int_0^1 N_\tau d\tau = 0_4$. Besides

$$\overrightarrow{\exp}\left(\delta k \int_0^1 N_\tau d\tau\right) = \mathrm{e}^{\delta k \int_0^1 N_\tau d\tau} = I_4.$$

On the other hand,

$$D_\tau = \overrightarrow{\exp}\left(-\delta k \,\mathrm{ad} \int_0^\tau N_\sigma d\sigma\right) k^{-1} M = \mathrm{e}^{-\delta k \,\mathrm{ad} \int_0^\tau N_\sigma d\sigma}\, k^{-1} M$$

$$= \left(I_4 + (-\delta k)\,\mathrm{ad} \int_0^\tau N_\sigma d\sigma + \frac{(-\delta k)^2}{2}\,\mathrm{ad}^2 \int_0^\tau N_\sigma d\sigma + \dots\right) k^{-1} M. \tag{9}$$

By direct computations

$$\mathrm{ad}\left(\int_0^\tau N_\sigma d\sigma\right) M = \begin{pmatrix} -CA\dot{s}(\tau) & 0_2 \\ 0_2 & AC\dot{s}(\tau) \end{pmatrix},$$

$$\mathrm{ad}^2\left(\int_0^\tau N_\sigma d\sigma\right) M = \begin{pmatrix} 0_2 & 0_2 \\ -2ACA\dot{s}^2(\tau) & 0_2 \end{pmatrix}$$

and

$$\mathrm{ad}^j\left(\int_0^\tau N_\sigma d\sigma\right) M = 0_4, \qquad \text{for } j \geq 3.$$

Therefore series in (9) ends at the second order term and monodromy matrix can be represented as

$$P^1 = \overrightarrow{\exp} \int_0^1 D_\tau \, d\tau$$

where

$$D_\tau = \begin{pmatrix} \delta\dot{s}(\tau)CA & k^{-1}C \\ k^{-1}g\,A - \delta^2 k\,\dot{s}^2(\tau)A\,CA & -\delta\dot{s}(\tau)AC \end{pmatrix}. \tag{10}$$

Obviously the matrix P^1 is uniquely determined, but the representation of P^1 as a chronological exponential $\overrightarrow{\exp}\int_0^1 D_\tau \, d\tau$ is not unique. Our construction of D_τ allows to establish stability results basing on the properties of the averaging of D_τ.

From what was said in Section 2.3, the system can not be asymptotically stable since $\ln P^1$ is traceless. It can be stable if all eigenvalues are imaginary numbers (two pairs of conjugate imaginary numbers).

3.2. Asymptotic expansion for the logarithm of monodromy matrix

Introduce the logarithm

$$\Lambda_{0,t}(D_\tau) = \ln \overrightarrow{\exp} \int_0^t D_\tau \, d\tau.$$

As it was mentioned in Section 2.2 $\Lambda_{0,t}(D_\tau)$ admits an expansion

$$\Lambda_{0,t}(D_\tau) = \sum_{m=1}^{\infty} \Lambda^{(m)} \tag{11}$$

where $\Lambda^{(m)}$ is defined by (3).

Since w_α in (4) is a Lie bracket of length $(m-1)$, one arranges an estimate for $[D_{\tau_m}, \dots, [D_{\tau_2}, D_{\tau_1}] \dots]$ where D_τ is defined by (10). Assume that $|\dot{s}(\tau)| \le \mu$ for all $\tau \in [0, 1]$.

LEMMA 1. *There exist positive constants σ and a such that, for all $\bar{\tau} = (\tau_1, \dots, \tau_m) \in [0,1]^m$, iterated Lie bracket of length m, $m > 2$, $[D_{\tau_m}, \dots, [D_{\tau_2}, D_{\tau_1}] \dots]$ can be represented as*

$$\delta^m \begin{pmatrix} \sigma_{11}^m(\bar{\tau})\,(CA)^m + \varepsilon^2\, C_{11}^m & \varepsilon\, \sigma_{12}^m(\bar{\tau}) =, C(AC)^{m-1} + \varepsilon^3\, C_{12}^m \\ \varepsilon^{-1}\, \sigma_{21}^m(\bar{\tau}) A\,(CA)^m + \varepsilon\, C_{21}^m & = \sigma_{11}^m(\bar{\tau})\,(AC)^m - \varepsilon^2\,(C_{11}^T)^m \end{pmatrix},$$

where $\varepsilon = (\delta k)^{-1}$, and $|\sigma_{ij}^m(\bar{\tau})| < \sigma^m$ and $\|C_{ij}^m\| < a^m$ for ; $i, j = 1, 2$.

Proof is done by induction on length of the Lie bracket.

LEMMA 2. *Series* (11) *is absolutely convergent if*

$$|\delta| < \frac{0.4432}{\hat{\sigma}\, a}$$

where $\hat{\sigma} = (\sigma^m + 1)^{1/m}$ and a, σ are the constants introduced in Lemma 1.

Proof. Proof is carried out by proving several inequalities.

For a matrix $A = (a_{ij})_{i,j=1}^n$, $a_{ij} \in \mathbb{R}$, one defines the norm by

$$\|A\| = \max_{1 \le j \le n} \sum_{i=1}^{n} |a_{ij}|.$$

In what follows $[D_{\tau_m}, \dots, [D_{\tau_2}, D_{\tau_1}] \dots]$ is denoted by D_m.

From Lemma 1 and since $\varepsilon < 1$

$$\begin{aligned} \|D_m\| &< \delta^m \, \max\{\sigma^m\, \|CA\|^m + \varepsilon^2\, \|C_{11}^m\| + \varepsilon^{-1}\, \sigma^m \|A\|\, \|CA\|^m + \varepsilon\, \|C_{21}^m\|; \\ &\qquad \varepsilon\, \sigma^m\, \|C\|\|AC\|^{m-1} + \varepsilon^3\, \|C_{12}^m\| + \sigma^m\, \|AC\|^m + \varepsilon^2\, \|(C_{11}^T)^m\|\} \\ &< \delta^m\, a^m\, \max\{\sigma^m + \varepsilon^2 + \varepsilon^{-1}\, \sigma^m + \varepsilon; \varepsilon\, \sigma^m + \varepsilon^3 + \sigma^m + \varepsilon^2\} \\ &< \delta^m\, a^m\, \max\{\sigma^m(\varepsilon^{-1} + 1) + \varepsilon(\varepsilon + 1);\ \sigma^m(\varepsilon + 1) + \varepsilon^2(\varepsilon + 1)\}, \\ &< \delta^m\, a^m\, [\sigma^m(\varepsilon^{-1} + 1) + \varepsilon(1 + \varepsilon)] < 2\,\delta^m\, a^m\, \varepsilon^{-1}\,(\sigma^m + 1) \\ &< 2\,\varepsilon^{-1}\, \delta^m\, a^m\, \hat{\sigma}^m, \end{aligned}$$

where $\hat{\sigma}^m > \sigma^m + 1$. Therefore,

$$\|D_m\| \le 2\,\varepsilon^{-1}\,\delta^m\,a^m\,\hat{\sigma}^m. \tag{12}$$

Now one sets an upper bound for $\|g_m(D_{\tau_1}, \ldots, D_{\tau_m})\|$. From the definition (4) of g_m it follows that

$$\|g_m(D_{\tau_1}, \ldots, D_{\tau_m})\| \le \left(\sum_{\alpha=1}^{(2m-1)!!} |b_{\upsilon_{1\alpha}} \ldots b_{\upsilon_{m\alpha}}| \right) \|D_m\| \le \chi_m \|D_m\|, \tag{13}$$

where χ_m are constants determined by the Taylor expansion of the function

$$\chi(z) = \frac{z}{2}(1 - \cot\frac{z}{2}) + 2 = \sum_{\alpha=0}^{\infty} |b_\alpha| z^\alpha, \qquad z \in \mathbb{C}.$$

Recall that this Taylor expansion converges for $|z| \le 2\pi$.

From (3), (12) and (13) it follows that

$$\begin{aligned}
\|\Lambda^{(m)}\| &\le \int_0^1 d\tau_1 \int_0^{\tau_1} d\tau_2 \ldots \int_0^{\tau_{m-1}} d\tau_m\, \|g_m(D_{\tau_1}, \ldots, D_{\tau_m})\| \\
&\le \chi_m \left(\int_0^1 d\tau_1 \int_0^{\tau_1} d\tau_2 \ldots \int_0^{\tau_{m-1}} d\tau_m\, \|D_m\| \right) \\
&\le 2\,\varepsilon^{-1}\,\chi_m\,\delta^m\,a^m\,\hat{\sigma}^m \left(\int_0^1 d\tau_1 \int_0^{\tau_1} d\tau_2 \ldots \int_0^{\tau_{m-1}} d\tau_m \right) \\
&\le 2\,\varepsilon^{-1}\,\chi_m\,\delta^m\,a^m\,\hat{\sigma}^m\,\frac{1}{m!}.
\end{aligned}$$

In [1] it is established that, for each $\gamma \in (0, 2\pi)$

$$\chi_m \le (m-1)!\frac{\gamma}{2}\left(\frac{2M(\gamma)}{\gamma}\right)^m$$

where $M(\gamma) = \max_{z\in\mathbb{C},|z|=\gamma} |\chi(z)|$. Therefore

$$\|\Lambda^{(m)}\| < 2\,\varepsilon^{-1}\,(m-1)!\frac{\gamma}{2}\left(\frac{2M(\gamma)}{\gamma}\right)^m \delta^m\,a^m\,\hat{\sigma}^m\,\frac{1}{m!} < \varepsilon^{-1}\,\gamma \left(\frac{2M(\gamma)}{\gamma}\,\delta\,a\,\hat{\sigma}\right)^m.$$

Series (11) converges absolutely if

$$\left|\,\delta\,a\,\hat{\sigma}\,\right| < \max_{\gamma\in(0,2\pi)} \frac{\gamma}{2\,M(\gamma)} = 0.4432.$$

□

3.3. Estimate for the rest term $\Lambda - \Lambda^{(1)}$

One would like to estimate the rest term $\Lambda - \Lambda^{(1)}$. Obviously,

$$\Lambda - \Lambda^{(1)} = \sum_{m=2}^{\infty} \Lambda^{(m)}$$

and by Lemma 2 the latter series is convergent for sufficiently small $\delta > 0$. The following result holds.

LEMMA 3. *For series* (11)*, there exists a constant $\xi > 0$ such that,*

$$\Lambda - \Lambda^{(1)} = \delta^2 R$$

where

$$R = \begin{pmatrix} R_{11} + \varepsilon^2 Q_{11} & \varepsilon R_{12} + \varepsilon^3 Q_{12} \\ \varepsilon^{-1} R_{21} + \varepsilon Q_{21} & -R_{11}^T - \varepsilon^2 Q_{11}^T \end{pmatrix},$$

and $\|R_{ij}\|,\ \|Q_{ij}\| < \xi;\ i, j = 1, 2.$

Notice that the block structure of $\Lambda^{(m)}$ has been established in Lemma 1. Let one set an estimate for the left-upper 2×2 block. One gets

$$\begin{aligned}
\|\Lambda_{11}^{(m)}\| &< \frac{\delta^m}{m!} \chi_m \|\sigma_{11}^m(\bar{\tau})\,(CA)^m + \varepsilon^2 C_{11}^m\| \\
&< \frac{\delta^m}{m} \frac{\gamma}{2} \left(\frac{2M(\gamma)}{\gamma}\right)^m (\sigma^m \|CA\|^m + \varepsilon^2 \|C_{11}^m\|) \\
&< \frac{\delta^m}{m} \frac{\gamma}{2} \left(\frac{2M(\gamma)}{\gamma}\right)^m a^m (\sigma^m + \varepsilon^2) \\
&< \delta^2 \frac{\gamma}{2} \left(\frac{2M(\gamma)}{\gamma} a\right)^2 \left[\left(\delta\, a\, \sigma\, \frac{2M(\gamma)}{\gamma}\right)^{m-2} + \varepsilon^2 \left(\delta\, a\, \frac{2M(\gamma)}{\gamma}\right)^{m-2}\right].
\end{aligned}$$

Thus the series $\sum_{m=2}^{\infty} \Lambda_{11}^{(m)}$ converges absolutely if $|\delta\, a\, \sigma| < \frac{\gamma}{2\,M(\gamma)}$ and $|\delta\, a| < \frac{\gamma}{2\,M(\gamma)}$. By Lemma 2, both conditions are true since $\sigma < \bar{\sigma}$ and $\bar{\sigma} > 1$.

The proof for other blocks is similar.

3.4. Stability of the first-order averaging

Eigenvalues of $\Lambda^{(1)}$ perform a crucial role in establishing stability conditions.

Computing $\Lambda^{(1)}$, which is the first averaging of D_τ, one obtains

$$\Lambda^{(1)} = \int_0^1 D_\tau\, d\tau = \begin{pmatrix} 0_2 & k^{-1} C \\ k^{-1} g\, A - \delta^2 k\, \bar{s} A\, C\, A & 0_2 \end{pmatrix}$$

with

$$\bar{s} = \int_0^1 \dot{s}^2(\tau)\, d\tau > 0.$$

Obviously $\Lambda^{(1)}$ is a Hamiltonian matrix of dimension 4.

Denoting $\delta^2 \bar{s} = \gamma$ and $k^{-1} g\, A - \gamma\, k\, ACA = \Sigma$ one computes the characteristic polynomial of $\Lambda^{(1)}$:

$$\det(\lambda I_4 - \Lambda^{(1)}) = \lambda^4 + p_2 \lambda^2 + p_0 \tag{14}$$

where

$$p_2 = -\operatorname{tr}(k^{-1}\Sigma\, C) = -k^{-2} g\, \operatorname{tr}(A\, C) + \gamma\, \operatorname{tr}(A\, C)^2 \tag{15}$$

and

$$p_0 = \det(k^{-1}\Sigma\, C) = \det(k^{-2} g\, I_2 - \gamma\, A\, C) \det(A\, C). \tag{16}$$

As one knows $\Lambda^{(1)}$ is a Hamiltonian matrix; its characteristic polynomial (14) is biquadratic. If $\Lambda^{(1)}$ is stable and possesses two distinct pairs of conjugate imaginary nonzero eigenvalues, then p_2, p_0 must be positive. Therefore there must hold:

$$p_2 = -\operatorname{tr}(k^{-1}\Sigma\, C) > 0 \tag{17}$$

and

$$p_0 = \det(k^{-1}\Sigma\, C) > 0. \tag{18}$$

Inequality (17) is equivalent to (see equation (15))

$$\delta^2 \int_0^1 \dot{s}^2(\tau)\, d\tau > k^{-2}\, g\, \frac{\operatorname{tr}(A\, C)}{\operatorname{tr}(A\, C)^2}. \tag{19}$$

To study the sign of $\det(k^{-1}\Sigma C)$ in (18) note that $\det(A\, C) > 0$. Then from (16), the sign of $\det(k^{-1}\Sigma C)$ depends only on the sign of $\det\left(k^{-2} g\, I_2 - \gamma\, A\, C\right)$, which is a quadratic polynomial in γ:

$$\det\left(k^{-2} g\, I_2 - \gamma\, A\, C\right) = \gamma^2 \det(A\, C) - \gamma\, k^{-2}\, g \operatorname{tr}(A\, C) + k^{-4} g^2. \tag{20}$$

The discriminant of this latter polynomial equals

$$\operatorname{tr}^2(A\, C) - 4\det(A\, C) > 0.$$

Therefore this polynomial has two real roots

$$\gamma_{\pm} = k^{-2}\, g\, \frac{-\operatorname{tr}(A\, C) \pm \sqrt{\operatorname{tr}^2(A\, C) - 4\det(A\, C)}}{2\det(A\, C)},$$

and the inequality (18) holds either for

$$\delta^2 \int_0^1 \dot{s}^2(\tau)d\tau > \gamma_+, \tag{21}$$

or for

$$\delta^2 \int_0^1 \dot{s}^2(\tau)\, d\tau < \gamma_-. \tag{22}$$

Inequality (19) is incompatible with (22), while (21) implies (19). Therefore (19)-(22)-(21) can be reduced to a single inequality (equivalent to (21))

$$\delta^2 \int_0^1 \dot{s}^2(\tau)\, d\tau > k^{-2} g \frac{\upsilon(r_1+r_2) + \sqrt{\upsilon^2(r_1+r_2)^2 - 4\upsilon\, r_1\, r_2}}{2\upsilon}, \tag{23}$$

where

$$\frac{m_1+m_2}{m_1} = \upsilon,\ \det(A\,C) = \frac{\upsilon}{r_1 r_2} > 0,\ \operatorname{tr}(A\,C) = \upsilon \frac{r_1+r_2}{r_1 r_2} > 0. \tag{24}$$

The following fact holds for the polynomial (14).

LEMMA 4. *The discriminant $p_2^2 - 4p_0$ of the biquadratic polynomial (14) is nonnegative for any choice of the parameters m_1, m_2, r_1, r_2 of the system. It is positive if the condition (23) holds.*

Proof. To simplify the notation denote $\operatorname{tr}(A\,C)$, $\operatorname{tr}(A\,C)^2$ by t_1, t_2 and $\det(A\,C)$ by d, correspondingly. Recall that $\delta^2\bar{s} = \gamma$, and put $\eta = k^{-2}g$. Then from formulae (15), (16) and (20) one concludes

$$\begin{aligned} p_2^2 - 4\,p_0 &= (-\eta\, t_1 + \gamma\, t_2)^2 - 4(\gamma^2 d - \gamma\, \eta\, t_1 + \eta^2)\, d \\ &= \eta^2\,(t_1^2 - 4\,d) - 2\gamma\, \eta\, t_1(t_2 - 2\,d) + \gamma^2\,(t_2 - 2\,d)(t_2 + 2\,d). \end{aligned}$$

For any (2×2)-matrix N the identity $\operatorname{tr} N^2 = (\operatorname{tr} N)^2 - 2\det N$ holds. Taking $N = AC$ one concludes with the identity $t_2 = t_1^2 - 2d$. Therefore $t_1^2 - 4\,d = t_2 - 2\,d$ and one obtains:

$$p_2^2 - 4\,p_0 = (t_1^2 - 4\,d)(\gamma\, t_1 - \eta)^2. \tag{25}$$

Substituting expressions (24) for $t_1 = \operatorname{tr}(A\,C)$ and $d = \det(A\,C)$ one gets

$$t_1^2 - 4\,d = \frac{\upsilon}{r_1^2\, r_2^2}\left(\upsilon\, r_1^2 + \upsilon\, r_2^2 + 2(\upsilon - 2)\, r_1\, r_2\right).$$

Since $\upsilon > 1$ for $m_2 > 0$ (non triviality condition), the latter expression is strictly positive for all non vanishing r_1, r_2 and hence the right hand side of (25) is nonnegative.

If condition (21) holds, then $\gamma > \eta\frac{r_1+r_2}{2}$ and (since $\upsilon > 1$)

$$\gamma\, t_1 - \eta = \gamma\, \upsilon \frac{r_1+r_2}{r_1 r_2} - \eta > \gamma \frac{r_1+r_2}{r_1 r_2} - \eta > \eta\left(\frac{(r_1+r_2)^2}{2r_1 r_2} - 1\right) = \eta \frac{r_1^2 + r_2^2}{2r_1 r_2} > 0.$$

Therefore $p_2^2 - 4p_0$ defined by (25) is positive provided that (21) holds. □

3.5. Stability of double inverted pendulum

Finally, one is able to establish the main result regarding the stability of double pendulum, or, the same of matrix Λ.

THEOREM 1. *For each $\epsilon > 0$ there exist $\delta_0 > 0, k_0 > 0$ such that upper equilibrium position $(\theta_1, \theta_2) = (0, 0)$ of inverted double pendulum is stable if $0 < \delta < \delta_0, k > k_0$ and*

$$\delta^2 \int_0^1 \dot{s}^2(\tau)\, d\tau > k^{-2}\, g\, \frac{v(r_1 + r_2) + \sqrt{v^2(r_1 + r_2)^2 - 4\, v\, r_1\, r_2}}{2v} + \epsilon \tag{26}$$

and unstable if $0 < \delta < \delta_0, k > k_0$ and

$$\delta^2 \int_0^1 \dot{s}^2(\tau)\, d\tau < k^{-2}\, g\, \frac{v(r_1 + r_2) + \sqrt{v^2(r_1 + r_2)^2 - 4\, v\, r_1\, r_2}}{2v} - \epsilon. \tag{27}$$

Proof. From Lemma 3, $\Lambda = \Lambda^{(1)} + \delta^2 R$ and one sees that

$$\det\left(\lambda I - (\Lambda^{(1)} + \delta^2 R)\right) = \det(\lambda I - \Lambda^{(1)}) + r(\lambda)$$

where $r(\lambda) = \rho_2\lambda^2 + \rho_0$, $\rho_2 = \mathcal{O}(\delta(\delta^2 + k^{-2}))$, $\rho_0 = \mathcal{O}(\delta(\delta^4 + k^{-4}))$, as $\delta + k^{-1} \to 0$. The characteristic polynomial of Λ can be written as

$$\det(\lambda I - \Lambda) = \lambda^4 + q_2\lambda^2 + q_0 \tag{28}$$

and, under condition (26), is close to the characteristic polynomial $\lambda^4 + p_2\lambda^2 + p_0$ of $\Lambda^{(1)}$. Namely, $p_0 \neq 0$, $p_2 \neq 0$ and $q_2 = p_2\,(1 + \mathcal{O}(\delta))$, $q_0 = p_0\,(1 + \mathcal{O}(\delta))$ as $\delta \to 0$.

Assume that condition (26) holds. Then $p_0 > 0$, $p_2 > 0$ and $p_2^2 - 4p_0 > 0$. Evidently for sufficiently small δ_0, and for $\delta < \delta_0$ and k satisfying (26) one gets $q_2 > 0$, $q_0 > 0$ and $q_2^2 - 4q_0 > 0$. Therefore Λ is stable.

If condition (27) holds, then $p_0 < 0$ and for sufficiently small δ_0, also $q_0 < 0$ and one gets instability of Λ. □

4. Spherical case

4.1. Simple spherical pendulum

Consider a simple inverted pendulum under the same settings as in Section 3. Assume that motion takes place on 3D-space.

As before, let m and r be, respectively, mass and length of the pendulum. Let θ be the angle between the pendulum and its projection on the xOz plane. Denote by ϕ the angle between the latter projection and the positive part of vertical axis z. The system has two degrees of freedom: θ and ϕ. Let q denote (θ, ϕ).

Proceeding as for planar case, an equilibrium point for system is $q = 0$. In a neighborhood of this equilibrium, the four-dimensional Hamiltonian system of first order equations arising from linearized Hamiltonian is, in matrix form,

$$\begin{pmatrix} \dot{q} \\ \dot{p} \end{pmatrix} = \begin{pmatrix} 0 & (m\,r^2)^{-1}\,I_2 \\ m\,r\,I_2\,[g + k^2\ddot{s}(k\,t)] & 0 \end{pmatrix} \begin{pmatrix} q \\ p \end{pmatrix}$$

or

$$\dot{z}(t) = \begin{pmatrix} 0 & C \\ A\,g + \delta\,k^2\ddot{s}(k\,t)\,A & 0 \end{pmatrix} z(t) = Q(t)\,z(t) \tag{29}$$

with $z(t) = (q, p)^T$, $A = m\,r\,I_2$ and $C = (m\,r^2)^{-1}\,I_2$.

System (29) is analogous to system (8) and one may apply the approach of previous sections. So, for simple spherical pendulum equilibrium position $q = 0$ is stable whenever, c.f. condition (23),

$$\delta^2 \int_0^1 \dot{s}^2(\tau)\,d\tau > k^{-2}\,g\,\frac{\mathrm{tr}(A\,C) + \sqrt{\mathrm{tr}^2(A\,C) - 4\det(A\,C)}}{2\det(A\,C)} + \epsilon.$$

One has $\mathrm{tr}(A\,C) = 2\,r^{-1}$ and $\det(A\,C) = r^{-2}$. Stability condition arises in a very simple form.

THEOREM 2. *For each $\epsilon > 0$ there exist $\delta_0 > 0$, $k_0 > 0$ such that upper equilibrium position $(\theta, \phi) = 0$ of simple spherical pendulum is stable if* $0 < \delta < \delta_0$, $k > k_0$ *and*

$$\delta^2 \int_0^1 \dot{s}^2(\tau)\,d\tau > k^{-2}\,g\,r + \epsilon$$

and unstable if $0 < \delta < \delta_0$, $k > k_0$ *and*

$$\delta^2 \int_0^1 \dot{s}^2(\tau)\,d\tau < k^{-2}\,g\,r - \epsilon.$$

4.2. Double spherical pendulum

Consider a double pendulum as presented in Section 3 describing a 3-D motion. Let θ_i; $i = 1, 2$ and ϕ_i; $i = 1, 2$ be as described in Section 4.1.

Proceeding as in two previous cases one linearizes Hamiltonian system at the equilibrium point and obtains eight equations on variables θ_1, θ_2, ϕ_1, ϕ_2, p_{θ_1}, p_{θ_2}, p_{ϕ_1} and p_{ϕ_2} which take the form

$$\begin{pmatrix} \dot{q} \\ \dot{p} \end{pmatrix} = \begin{pmatrix} 0_2 & C \\ -B(t) & 0_2 \end{pmatrix} \begin{pmatrix} q \\ p \end{pmatrix}$$

where

$$C = \begin{pmatrix} G & 0_2 \\ 0_2 & G \end{pmatrix}, \quad \text{with} \quad G = \frac{1}{m_1 m_2 r_1^2 r_2^2} \begin{pmatrix} r_2^2 & -m_2\,r_1\,r_2 \\ -m_2\,r_1\,r_2 & (m_1 + m_2)\,r_1^2 \end{pmatrix},$$

and

$$B(t) = \begin{pmatrix} -A & 0_2 \\ 0_2 & -A \end{pmatrix} [g + \delta k^2 \dddot{s}(k\,t)], \quad \text{with} \quad A = \begin{pmatrix} (m_1 + m_2)\, r_1 & 0 \\ 0 & m_2\, r_2 \end{pmatrix}.$$

This system of eight equations can be re-written as two decoupled systems of dimension four such that each system structure is analogous to system (7)

$$\begin{pmatrix} \dot{\theta} \\ \dot{p}_\theta \end{pmatrix} = \begin{pmatrix} 0 & G \\ g\,A + \delta k^2 \dddot{s}(k\,t)\,A & 0 \end{pmatrix} \begin{pmatrix} \theta \\ p_\theta \end{pmatrix} \tag{30}$$

and

$$\begin{pmatrix} \dot{\phi} \\ \dot{p}_\phi \end{pmatrix} = \begin{pmatrix} 0 & G \\ g\,A + \delta k^2 \dddot{s}(k\,t)\,A & 0 \end{pmatrix} \begin{pmatrix} \phi \\ p_\phi \end{pmatrix}. \tag{31}$$

Decoupled form (30)-(31) of spherical inverted double pendulum shows that one can study independently its projections on xOy (system on variables $\phi_i;\, i = 1, 2$) and yOz (system on variables $\theta_i;\, i = 1, 2$). Therefore stability conditions for spherical inverted double pendulum coincides with those presented in Theorem 1.

References

[1] AGRACHEV A.A. AND GAMKRELIDZE R.V. *The exponential representation of flows and the chronological calculus*, (in russian) Mat. Sb., **107** (1978), 467–532; english translation: Math. USSR Sb. **35** (1979), 727–785.

[2] AGRACHEV A.A. AND GAMKRELIDZE R.V. *Chronologica algebras and nonstationary vecotr field*, Journal of Soviet Mathematics, **17** (1981) 1650–1675; translation from: Itogi Nauki i Tekhniki, Seriya Problemy Geometrii **11** (1980) 135–176.

[3] AGRACHEV A.A., GAMKRELIDZE R.V., AND SARYCHEV A.V. *Local invariants of smooth control systems*, Acta Applicandae Mathematicae **14** (3) (1989), 191–237.

[4] AGRACHEV A.A. AND SARYCHEV A.V. *On reduction of a smooth system linear in the control*, Math. USSR Sbornik **58** (1) (1987) 15–30; translation from Mat. Sb., Nov. Ser., **130(172)** 1 (1986) 18-34.

[5] BELLMAN R.E., BENTSMAN J. AND MEERKOV S.M. *Vibrational control of nonlinear systems.* IEEE Trans. Autom. Control **31** (1986) 710–724.

[6] SARYCHEV A.V. *Lie - and chrolonologico-algebraic tools for studying stability of time-varying sytems*, Systems and control Letters **43** (1)(2001), 59–76.

[7] SARYCHEV A.V. *Stability criteria for time-periodic systems via high-order averaging techniques*, Nonlinear Control in Year 2000 **2** (2001), 365–377.

[8] SHIRIAEV A.S., EGELAND O., LUDVIGSEN H. AND FRADKOV A.L. *Vss-version of energy-based control for swinging up a pendulum*, Systems and Control Letters **44** (1) (2001) 45–56.

[9] AGRACHEV A.A. AND SACHKOV YU.L. *Control theory from the geometric viewpoint*, Springer, Berlin 2004.

[10] ARNOLD V.I. *Mathematical Methods of Classical Mechanics*, Graduate texts in Mathematics, Springer-Verlag, New York 1989.

[11] BLEKHMAN I. *Vibrational mechanics: nonlinear dynamic effects, general approach, applications*, World scientific, Singapure 1999.

AMS Subject Classification: 34E05, 34C29, 34D20.

Maria Isabel CAIADO, Centro de Matemática, Universidade do Minho, Campus de Gualtar, 4710-057 Braga, PORTUGAL
icaiado@math.uminho.pt

Andrey V. SARYCHEV, Dipartimento di Matematica per le Decisioni, Università di Firenze, Via C. Lombroso 6/17, 50134 - Firenze (FI), ITALIA
andrey.sarychev@dmd.unifi.it

Rend. Sem. Mat. Univ. Pol. Torino - Vol. 63, 4 (2005)
Control Theory and Stabil., I

R. Giambò*

AN ANALYTICAL THEORY FOR OPTIMAL CONTROLS ON RIEMANNIAN MANIFOLDS

Abstract. The problem of interpolation in a Riemannian manifold (M, g) discussed in [9] (using variational techniques on a suitable functional fra-mework) is here reviewed, stating the main existence results for minimizers and multiplicity and regularity results for critical points of the involved functional f.

1. Introduction

In recent years, a geometric theory of the so–called Riemannian cubic polynomials has been developed [3, 5, 8, 12]. With the aim of a generalization, variational problems with Lagrangians involving higher order derivatives can be encountered in Control Theory, especially in robotics (see for instance [2], where higher order interpolation in Riemannian manifold is exploited), and has been successively developed [1, 4, 6, 10].

This paper intends to give a brief survey of the results contained in [9], where intepolating polynomials of odd order $2k + 1$ on Riemannian manifolds are studied as critical points of a functional involving covariant derivatives of k–th order of the velocity. The path space where one looks for critical points consists of curves in the Sobolev space H^{k+1}, satisfying boundary conditions on position and velocity and all its covariant derivatives up to order $k - 1$ at the initial and final point. This set is shown to be a Hilbert manifold, and the functional defined on it is shown to satisfy a compactness property (Palais–Smale condition). In this way, classical techniques from Global Analysis can be applied to recover results of multiplicity of critical points, and also existence of curves that globally minimize the functional on the path space. It is interesting to note that, in spite of the sub–Riemannian nature of the variational problem, the class of Riemannian polynomials does not contain *abnormal* minimizers.

2. The variational framework

The space of paths satisfying higher order regularity assumptions and suitable boundary conditions, which is the domain of our variational problem, is introduced here in an abstract context. The main idea is to establish some topological properties of the space to be used later (Remark 2 and Section 4) where homotopic invariants play a crucial role.

Let $\mathfrak{M}$ be a functor that associates to each compact interval $I = [a, b]$ of $\mathbb{R}$ and to each differentiable manifold M a topological space, denoted by $\mathfrak{M}(I, M)$, consisting of curves $x : I \to M$. Since the purpose is to study sets of curves satisfying boundary

*Dipartimento di Matematica e Informatica, Università degli Studi di Camerino, Italy

conditions involving derivatives up to the k–th order (with $k \geq 1$), it is natural to consider the case that each curve in the class $\mathfrak{M}$ be of class $\mathcal{C}^k$. For the detailed axiomatic definition of $\mathfrak{M}$, we refer the reader to [9].

REMARK 1. For the purposes of this paper, the basic examples of $\mathfrak{M}$ are the functors $\mathcal{C}^k$ and $W^{k+1,p}$, with $p \geq 1$; note however that, depending on the variational problem to be studied, many other relevant examples of regularity might be considered, like for instance a regularity of Sobolev mixed type.

Let us now fix two arbitrary points $p, q \in M$. We have an inclusion

$$\mathfrak{M}_{p,q}\big([a,b], M\big) \subset \mathcal{C}^0_{p,q}\big([a,b], M\big), \tag{1}$$

where

$$\mathfrak{M}_{p,q}\big([a,b], M\big) = \Big\{x \in \mathfrak{M}([a,b], M) : x(a) = p,\ x(b) = q\Big\}, \tag{2}$$

$$\mathcal{C}^0_{p,q}\big([a,b], M\big) = \Big\{x \in \mathcal{C}^0\big([a,b], M\big) : x(a) = p,\ x(b) = q\Big\}. \tag{3}$$

Using a classical result by Palais [13], it can be proved that (1) is actually a homotopy equivalence.

Note however that, so far, no boundary conditions, other that the initial and final position of the curve, have been imposed. To involve higher order boundary conditions, jet spaces will be used. Let us simply recall (referring the reader to [14] for more details) that, fixed $x \in M$, we can set:

$$\mathfrak{J}^k(M)_p = \big\{x \in \mathcal{C}^k\big(\,]-\varepsilon, \varepsilon[\,, M\big) : x(0) = p\big\}\Big/ \sim$$

where $\sim$ is the equivalence relation:

$$x_1 \sim x_2 \quad \iff \quad \big(\varphi \circ x_1\big)^{(i)}(0) = \big(\varphi \circ x_2\big)^{(i)}(0),\ i = 1, \dots, k$$

for some (hence for all) local chart φ of M around p; we will denote by $[x]$ the equivalence class of the curve x by the above equivalence relation. The disjoint union $\bigcup_{p \in M} \mathfrak{J}^k(M)_p$ can be canonically given a topological space structure, and will be denoted by $\mathfrak{J}^k(M)$.

Observe now that for all $t_0 \in [a,b]$ we have a well defined continuous map

$$\mathfrak{J}^k_{t_0} : \mathfrak{M}\big([a,b], M\big) \longrightarrow \mathfrak{J}^k(M)_{x(t_0)}$$

that is obtained by sending a curve x to its equivalence class in $\mathfrak{J}^k(M)_{x(t_0)}$. It can be seen that imposing boundary conditions related to retracts of the jet spaces $J^k(M)_p$ and $J^k(M)_q$ does not affect the homotopy type of the paths space. In other words, we have the following proposition (see [9] for the proof):

PROPOSITION 1. *Let $\mathcal{S}_1 \subset \mathfrak{J}^k(M)_p$ and $\mathcal{S}_2 \subset \mathfrak{J}^k(M)_q$ be retracts, and set:*

$$\mathfrak{M}_{p,q}\big([a,b], M; \mathcal{S}_1, \mathcal{S}_2\big) = \Big\{x \in \mathfrak{M}_{p,q}\big([a,b], M\big) : \mathfrak{J}^k_a(x) \in \mathcal{S}_1,\ \mathfrak{J}^k_b(x) \in \mathcal{S}_2\Big\}. \tag{4}$$

Then, the inclusion $\mathfrak{M}_{p,q}\big([a,b],M;\mathcal{S}_1,\mathcal{S}_2\big) \hookrightarrow \mathfrak{M}_{p,q}\big([a,b],M\big)$ is an homotopy equivalence.

Using the homotopy equivalence (1) already stated, we can immediately deduce the following corollary:

COROLLARY 1. *Under the assumptions of Proposition 1, the spaces $\mathcal{C}^0_{p,q}\big([a,b],M\big)$ and $\mathfrak{M}_{p,q}\big([a,b],M;\mathcal{S}_1,\mathcal{S}_2\big)$ have the same homotopy type.*

3. Variational formulation for Riemannian polynomials

We will now proceed with the study of the variational problem arising from the interpolation problem. Let us consider a smooth m-dimensional manifold M and a complete Riemannian metric g on M. We will denote by ∇ the covariant derivative of the Levi–Civita connection of g and by R the curvature tensor of ∇ chosen with the following sign convention: $R(X,Y)=[\nabla_X,\nabla_Y]-\nabla_{[X,Y]}$. We will write indifferently $R(X,Y)Z$ or $R(X,Y,Z)$, the latter notation being more appropriate when dealing with covariant derivatives of R.

For all $k\in\mathbb{N}$ and all $p\in[1,+\infty]$, we say that a curve $x:[a,b]\to M$ is of class $W^{k,p}$ if for all local chart (U,φ) of M and all interval $[c,d]\subset x^{-1}(U)$, the composition $\varphi\circ x|_{[c,d]}$ is in the Sobolev space $W^{k,p}\big([c,d],\mathbb{R}^m\big)$ of C^{k-1} curves having p-integrable weak k-th derivative. It is well known that $W^{k,p}\big([a,b],M\big)$ is an infinite dimensional Banach manifold modelled on $W^{k,p}\big([a,b],\mathbb{R}^m\big)$. We will set $H^k=W^{k,2}$, since we will be mainly interested in $H^{k+1}\big([0,1],M\big)$, which is a complete Hilbert manifold.

For all $x\in H^{k+1}\big([0,1],M\big)$, the tangent space $T_xH^{k+1}\big([0,1],M\big)$ is identified with the Hilbert space of all vector fields along x of class H^k.

It is easy to see that the map:

$$H^{k+1}\big([0,1],M\big)\ni x\mapsto\Big(\dot x(0),\frac{\mathrm{D}}{\mathrm{d}t}\dot x(0),\ldots,\tfrac{\mathrm{D}^{k-1}}{\mathrm{d}t^{k-1}}\dot x(0);\dot x(1),\frac{\mathrm{D}}{\mathrm{d}t}\dot x(1),\ldots,\tfrac{\mathrm{D}^{k-1}}{\mathrm{d}t^{k-1}}\dot x(1)\Big),$$

where $\frac{\mathrm{D}}{\mathrm{d}t}$ denotes covariant differentiation along the curve x, is a submersion taking value in the cartesian product $TM^{(k)}\times TM^{(k)}$, where $TM^{(k)}$ is the Whitney sum of vector bundles $TM\oplus\ldots\oplus TM$ (k times) with itself. Thus, if p,q are fixed points of M, and if $v_1,\ldots,v_k\in T_pM$, $w_1,\ldots,w_k\in T_qM$ are given tangent vectors, the set

$$\Gamma:=H^{k+1}\big([0,1],M;p,v_1,\ldots,v_k;q,w_1,\ldots,w_k\big) \tag{5}$$

consisting of those curves $x\in H^{k+1}\big([0,1],M\big)$ such that:

$$\begin{aligned} &x(0)=p,\ \dot x(0)=v_1,\ldots,\tfrac{\mathrm{D}^{k-1}}{\mathrm{d}t^{k-1}}\dot x(0)=v_k,\\ &x(1)=q,\ \dot x(1)=w_1,\ldots,\tfrac{\mathrm{D}^{k-1}}{\mathrm{d}t^{k-1}}\dot x(1)=w_k \end{aligned} \tag{6}$$

is a smooth embedded closed (hence complete) submanifold of $H^{k+1}\big([0,1],M\big)$. For all $x \in \Gamma$, the tangent space $T_x\Gamma$ is identified with the closed subspace of $T_xH^{k+1}\big([0,1],M\big)$ consisting of those vector fields V such that:

$$\begin{aligned} V(0) &= \frac{\mathrm{D}}{\mathrm{d}t}V(0) = \ldots = \tfrac{\mathrm{D}^k}{\mathrm{d}t^k}V(0) = 0,\\ V(1) &= \frac{\mathrm{D}}{\mathrm{d}t}V(1) = \ldots = \tfrac{\mathrm{D}^k}{\mathrm{d}t^k}V(1) = 0. \end{aligned} \tag{7}$$

and is endowed with the Riemannian structure induced by the Hilbert space inner product $\langle V, W\rangle = \int_0^1 g\big(\frac{\mathrm{D}^{k+1}}{\mathrm{d}t^{k+1}}V, \frac{\mathrm{D}^{k+1}}{\mathrm{d}t^{k+1}}W\big)\,\mathrm{d}t$.

REMARK 2. The metric g induces a diffeomorphism Ψ between the fiber bundle $\mathfrak{J}^k(M)$ and the vector bundle $TM^{(k)}$, that for each point $\pi \in M$ is given by

$$\mathfrak{J}^k(M)_\pi \ni [x] \overset{\Psi_\pi}{\longmapsto} \big(\dot{x}(0), \frac{\mathrm{D}}{\mathrm{d}t}\dot{x}(0), \ldots, \tfrac{\mathrm{D}^{k-1}}{\mathrm{d}t^{k-1}}\dot{x}(0)\big) \in T_\pi M \oplus \ldots \oplus T_\pi M. \tag{8}$$

If $\mathcal{S}_1$ and $\mathcal{S}_2$ denote the counterimages of $(v_1,\ldots,v_k)$ via Ψ_p and $(w_1,\ldots,w_k)$ via Ψ_q respectively, then it is easily seen that $\mathcal{S}_1$ and $\mathcal{S}_2$ are retracts of $\mathfrak{J}^k(M)_p$ and $\mathfrak{J}^k(M)_q$ respectively. It follows that Corollary 1 applies to the functional space Γ defined in (5), to conclude that Γ has the same homotopy type of the based loop space $\mathcal{C}^0_{p,q}\big([a,b],M\big)$.

In view of the above discussion, we will cast our variational problem on the space Γ. Indeed, we now give the following definition:

DEFINITION 1. *A a critical point x of the functional*

$$f(x) = \frac{1}{2}\int_0^1 g\big(\tfrac{\mathrm{D}^k}{\mathrm{d}t^k}\dot{x}, \tfrac{\mathrm{D}^k}{\mathrm{d}t^k}\dot{x}\big)\,\mathrm{d}t. \tag{9}$$

in Γ will be called a polynomial curve of order $2k+1$ in M *satisfying the boundary conditions* (6).

We are now going to establish the key property that the functional f needs to satisfy in order to recover results stated in Section 4. First, let us recall a couple of basic definitions. Given a smooth functional $f : \mathcal{X} \to \mathbb{R}$ on the Banach manifold $\mathcal{X}$, and given a smoothly varying Banach space structure $\mathfrak{h}$ on each tangent space $T_x\mathcal{X}$, a *Palais–Smale sequence* for f is a sequence $(x_n)_n$ in $\mathcal{X}$ such that:

1. $\lim\limits_{n\to\infty} f(x_n) = c \in \mathbb{R}$;

2. $\lim\limits_{n\to\infty} \|\mathrm{d}f(x_n)\|_{x_n} = 0$, where $\|\cdot\|_x$ is the norm of bounded linear functionals on $T_x\mathcal{X}$ induced by $\mathfrak{h}$.

We also recall that the functional f is said to satisfy the *Palais–Smale condition* if every Palais–Smale sequence for f has a converging subsequence in $\mathcal{X}$.

Palais–Smale condition is the compactness property required to apply classical results from Global Analysis. Of course, choosing the suitable path space is a crucial step. For instance, let us study cubic polynomials ($k = 1$) on the sphere S^2. If we take $H^2_{p,q}([0,1], S^2)$ as functional space – i.e. without fixing initial and final velocity – one can define a sequence x_n of curves making n loops around the great circle at p and q. This is a Palais–Smale sequence, but obviously does not possess any converging subsequence. Indeed, in the case we are studying, the constraints imposed by boundary conditions (6) on derivatives make things work properly, so that the following key result holds:

THEOREM 1. *The functional $f : \Gamma \to \mathbb{R}$ satisfies the Palais–Smale condition.*

REMARK 3. It may be worthwhile noticing that Palais–Smale condition holds for a whole class of functionals on Γ, also involving lower order derivatives, that contains the case introduced in (9) as a particular situation.

4. Conclusions

We are now ready to state the main results, applying Global Analysis techniques. To begin, a result of existence of Riemannian polynomials can be obtained:

PROPOSITION 2. *Each homotopy class of curves joining p and q contains a polynomial curve x of order $2k+1$ satisfying boundary conditions* (6) *and that minimizes the action functional f in that homotopy class.*

Proof. It is an easy application of the so-called *Minimax Principle* (see for instance [15]) and of Remark 2. □

We recall that if $\mathcal{X}$ is a topological space and $\mathcal{Y} \subseteq \mathcal{X}$, then the *Ljusternik-Schnirelman category* $\mathrm{cat}_{\mathcal{X}}(\mathcal{Y})$ is an homotopic invariant, given by the (possibly infinite) minimal number of closed, contractible subsets of $\mathcal{X}$ that cover $\mathcal{Y}$; we set $\mathrm{cat}(\mathcal{X}) = \mathrm{cat}_{\mathcal{X}}(\mathcal{X})$. The core of the Ljusternik–Schnirelman theory is given by the following classical result [15]:

PROPOSITION 3. *Let $\mathcal{X}$ be a complete Banach manifold and let $f : \mathcal{X} \to \mathbb{R}$ be a functional bounded from below that satisfies the Palais–Smale condition. Then, f has at least* $\mathrm{cat}(\mathcal{X})$ *critical points; moreover, if* $\mathrm{cat}(\mathcal{X}) = +\infty$*, then f has arbitrarily large critical values.*

In this case, Ljusternik–Schnirelman theory gives us a lower bound for Riemannian polynomials:

PROPOSITION 4. *There exist at least* $\mathrm{cat}\big(\mathcal{C}^0_{p,q}([0,1], M)\big)$ *polynomial curves of order* $2k+1$ *in M satisfying the boundary conditions* (6)*. If M is not contractible, then there exists a sequence $(x_n)_{n\in\mathbb{N}}$ of polynomial curves of order* $2k+1$ *in M satisfying*

(6), *and such that:*

$$\lim_{n\to\infty} f(x_n) = +\infty.$$

Proof. It follows immediately from Remark 2, Theorem 1 and Proposition 3. If M is not contractible, a well known result of Fadell and Husseini (see [7]) states that $\mathcal{C}^0_{p,q}\big([0,1], M\big)$ has infinite Ljusternik–Schnirelman category. □

Regularity of Riemannian polynomials is ensured by standard variational arguments instead:

PROPOSITION 5. *A Riemannian polynomial of order* $2k+1$ *satisfying boundary conditions* (6) *is smooth, and satisfies a differential equation of the form:*

$$\tag{10} \tfrac{\mathrm{D}^{2k+1}}{\mathrm{d}t^{2k+1}}\dot{x} + \mathbf{G}\big(\dot{x}, \frac{\mathrm{D}}{\mathrm{d}t}\dot{x}, \dots, \tfrac{\mathrm{D}^{2k}}{\mathrm{d}t^{2k}}\dot{x}\big) = 0,$$

where $\mathbf{G} : TM^{(2k)} \to \mathbb{R}$ *is a map given by the sum of tensor fields over* M *obtained from* R *and its covariant derivatives.*

Proof. Computing the first variation of f at x is a lengthy but straightforward process that gives

$$\tag{11} \mathrm{d}f(x)V = \int_0^1 g\Big(\tfrac{\mathrm{D}^k}{\mathrm{d}t^k}\dot{x}, \tfrac{\mathrm{D}^{k+1}}{\mathrm{d}t^{k+1}}V + \mathbf{F}\big(\dot{x}, \frac{\mathrm{D}}{\mathrm{d}t}\dot{x}, \dots, \tfrac{\mathrm{D}^{k-1}}{\mathrm{d}t^{k-1}}\dot{x}, V, \frac{\mathrm{D}}{\mathrm{d}t}V, \dots, \tfrac{\mathrm{D}^{k-1}}{\mathrm{d}t^{k-1}}V\big)\Big)\,\mathrm{d}t,$$

where $\mathbf{F} = 0$ if $k = 0$ and $F : TM^{(2k)} \to \mathbb{R}$ is defined by:

$$\tag{12} \mathbf{F}\big(\dot{x}, \frac{\mathrm{D}}{\mathrm{d}t}\dot{x}, \dots, \tfrac{\mathrm{D}^{k-1}}{\mathrm{d}t^{k-1}}\dot{x}, V, \frac{\mathrm{D}}{\mathrm{d}t}V, \dots, \tfrac{\mathrm{D}^{k-1}}{\mathrm{d}t^{k-1}}V\big) = \sum_{j=0}^{k-1} \tfrac{\mathrm{D}^j}{\mathrm{d}t^j}\big(R(V,\dot{x})\tfrac{\mathrm{D}^{k-j-1}}{\mathrm{d}t^{k-j-1}}\dot{x}\big), \quad k \geq 1.$$

Then, Euler–Lagrange equations (10) are obtained by successive integration by parts in (11), keeping in mind that the conditions (7) imply that all the boundary terms arising from these integration by parts vanish. □

EXAMPLE 1. Euler–Lagrange equation (10) for the cubic case ($k = 1$) is

$$\tfrac{\mathrm{D}^3}{\mathrm{d}t^3}\dot{x} - R\big(\dot{x}, \frac{\mathrm{D}}{\mathrm{d}t}\dot{x}\big)\,\dot{x} = 0,$$

whereas fifth order polynomials ($k = 2$) satisfy

$$\tfrac{\mathrm{D}^5}{\mathrm{d}t^5}\dot{x} + R\big(\frac{\mathrm{D}}{\mathrm{d}t}\dot{x}, \tfrac{\mathrm{D}^2}{\mathrm{d}t^2}\dot{x}\big)\dot{x} - R\big(\dot{x}, \tfrac{\mathrm{D}^3}{\mathrm{d}t^3}\dot{x}\big)\dot{x} = 0.$$

Finally, Morse theory can be applied, at least whenever $(p, v_1, \dots, v_k)$ and $(q, w_1, \dots, w_k)$ are not *conjugate* by Riemannian polynomials, that is it does not exist a critical point $x \in \Gamma$ of f such that the second variation $\mathrm{d}^2 f(x)$ of f at x is degenerate. Let us recall that, if $f : \mathcal{X} \to \mathbb{R}$ is a $\mathcal{C}^2$ map on a Hilbert manifold, f is a

Morse function if all its critical points are *nondegenerate* (i.e. $d^2 f(x_0)$ is represented by an invertible self–adjoint operator on $T_{x_0}\mathcal{X}$, whenever x_0 is a critical point), and the *Morse index* $m(x_0)$ is defined to be the index of the bilinear form $d^2 f(x_0)$.

In general, the central result of Morse theory can be stated as follows [15, 11]:

PROPOSITION 6. *Let $\mathcal{X}$ be a complete Hilbert manifold, $f : \mathcal{X} \to \mathbb{R}$ a Morse function which is bounded from below and satisfies the Palais–Smale condition. Then, given any coefficient field $\mathbb{F}$ and denoted by $\beta_n(\mathcal{X};\mathbb{F})$ the n-th* Betti's number *of $\mathcal{X}$ (i.e., the dimension of the n-th singular homology vector space with coefficient in $\mathbb{F}$), and by $\mathfrak{P}(\mathcal{X};\mathbb{F})(z) = \sum_{n=0}^{\infty} \beta_n(\mathcal{X};\mathbb{F}) z^n$ the* Poincaré polynomial *of $\mathcal{X}$ with coefficients in $\mathbb{F}$, then there exists a formal power series $Q(z) = \sum_{n=0}^{\infty} q_n z^n$ with coefficients $q_n \in \mathbb{N} \bigcup \{+\infty\}$ such that the following identity between formal power series holds:*

$$\sum_{x \text{ critical point of } f} z^{m(x)} = \mathfrak{P}(\mathcal{X};\mathbb{F})(z) + (1+z)Q(z). \tag{13}$$

In our case, it can be showed [9] that $f : \Gamma \to \mathbb{R}$ is a Morse function if and only if $(p, v_1, \ldots, v_k)$ and $(q, w_1, \ldots, w_k)$ are not conjugate by Riemannian polynomials. Then, using Remark 2, Theorem 1 and the above proposition, one gets the following form for (13):

$$\sum_{n=0}^{\infty} \kappa_n z^n = \mathfrak{P}\big(\mathcal{C}^0_{p,q}\big([0,1], M\big);\mathbb{F}\big) + (1+z)Q(z), \tag{14}$$

where κ_n is the the number of Riemannian polynomials of order $2k+1$ in M satisfying the boundary conditions (6) and having Morse index n. In particular, if β_n^0 denotes the n–th Betti number of $\mathcal{C}^0_{p,q}\big([0,1], M\big)$, we can state the following result:

PROPOSITION 7. *Assuming that $(p, v_1, \ldots, v_k)$ and $(q, w_1, \ldots, w_k)$ are not conjugate by Riemannian polynomials of order $2k+1$, then there are at least $\sum_{n=0}^{\infty} \beta_n^0$ Riemannian polynomials of order $2k+1$ in M satisfying (6). Moreover, the number of such polynomials is either infinite or odd.*

Proof. It follows from the above discussion, whereas last statement comes from (14) setting $z = 1$. □

References

[1] ALTAFINI C., *Geometric control methods for nonlinear systems and robotic applications*, Ph. D. Thesis Division of Optimization and Systems Theory, Department of Mathematics, Royal Institute of Technology, Stockholm, Sweden 2001.

[2] BAILLIEUL J., Proceedings of the IEEE International Conference on Robotics and Automation, Nice 1992, 715–721.

[3] CAMARINHA M., *The Geometry of cubic polynomials on Riemannian manifolds*, Ph. D. Thesis in Pure Mathematics, University of Coimbra, Portugal 1996.

[4] CAMARINHA M., CROUCH P. AND SILVA LEITE F., *Splines of class C^k on non-Euclidean spaces*, IMA J. Math. Control Inform. **12** (4) (1995), 399–410.

[5] CAMARINHA M., CROUCH P. AND SILVA LEITE F., *On the geometry of Riemannian cubic polynomials*, Diff. Geom. Appl. **15** (2001), 107–135.

[6] CROUCH P. AND SILVA LEITE F., *The dynamic interpolation problem: on Riemannian manifolds, Lie groups, and symmetric spaces*, J. Dynam. Control Systems **1** (2) (1995), 177–202.

[7] FADELL E. AND HUSSEINI S., *Category of loop spaces of open subsets in Euclidean space*, Nonlinear Analysis: T.M.A. **17** (1991), 1153–1161.

[8] GIAMBÒ R., GIANNONI F. AND PICCIONE P., An analytical theory for Riemannian cubic polynomials, IMA J. Math. Control Inform. **19** (4) (2002), 445–460

[9] GIAMBÒ R., GIANNONI F. AND PICCIONE P., *Optimal control on Riemannian manifolds by interpolation*, Math. Control Signals Systems **16** (4) (2004), 278-296.

[10] HERMANN R., *The differential geometric structure of general mechanical systems from the Lagrangian point of view*, J. Math. Phys. **23** (1) (1982), 2077–2089.

[11] MERCURI F., PICCIONE P. AND TAUSK D.V., *Notes on Morse theory*, Publicações Matemáticas IMPA, Rio de Janeiro 2001.

[12] NOAKES L., HEINZINGER G. AND PADEN B., *Cubic splines on curved spaces*, IMA J. Math. Control Inform. **6** (1989), 465–473.

[13] PALAIS R., *Homotopy theory of infinite dimensional manifolds*, Topology **5** (1966), 1–16.

[14] PALAIS R., *Foundations of global nonlinear analysis*, W. A. Benjamin, New York–Amsterdam 1968.

[15] PALAIS R. AND TERNG CH.-L., *Critical point theory and submanifold geometry*, Lect. Notes in Math. **1353**, Springer-Verlag, Berlin 1988.

AMS Subject Classification: 53Cxx, 58E10.

Roberto GIAMBÒ, Dipartimento di Matematica e Informatica, Università degli Studi di Camerino, Via Madonna delle Carceri, 62032 Camerino (MC), ITALY
e-mail: `roberto.giambo@unicam.it`

Rend. Sem. Mat. Univ. Pol. Torino - Vol. 63, 4 (2005)
Control Theory and Stabil., I

E. Girejko*

ON GENERALIZED DIFFERENTIAL QUOTIENTS OF SET-VALUED MAPS

Abstract. There are many important maps that are not differentiable in the classical way. In this paper one of the concepts of generalized differential, called Generalized Differential Quotient (abbr. GDQ), is studied. GDQ may exist for functions not differentiable in the classical sense and for set-valued maps, but it is not unique. Existence of minimal GDQs has been proved.

1. Introduction

The axiomatic definition of generalized differentiation theory (abbr. GDT) has been introduced by Hector Sussmann in [5]. He described also a few theories satisfying the axioms of GDT (see [5]). One of them is the theory of Generalized Differential Quotients (abbr. GDQs). It is known that GDQ of a single-valued or set-valued map, unlike ordinary differential, is not unique. In order to develop further Sussmann's theory we have then to answer some questions concerning GDQs: Which of GDQs are the best mathematical tools? Is any intersection of GDQs also a GDQ? The answers to these questions extend the basic issues concerning GDQ. Since GDQs are generalization of Clarke generalized gradients we will start with the definition of these gradients.

DEFINITION 1. *Let $f : X \to \mathbb{R}$ be Lipschitz near $x \in X$, where X is a finite-dimensional real linear space, and let v be a vector in X. The generalized directional derivative of f at x in the direction v, denoted $f^\circ(x; v)$, is defined as follows:*

$$f^\circ(x; v) = \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{f(y + tv) - f(y)}{t},$$

where y is a vector in X and t is a positive scalar.

DEFINITION 2. *Let X^* denote the dual space of X, i.e. the space of all bounded linear functionals $\zeta : X \to \mathbb{R}$. Let $f : X \to \mathbb{R}$ and $x \in X$. Then the generalized gradient of f at x, denoted $\partial f(x)$, is the subset of X^* given by*

$$\{\zeta \in X^* : f^\circ(x; v) \geq \langle \zeta, v \rangle \text{ for all } v \text{ in } X\}.$$

Here are some basic properties of generalized gradients (we assume that f is Lipschitz near x):

*This work was supported in part by CTS and in part by KBN (Bialystok Technical University grant No. W/IMF/1/04)

(a) $\partial f(x)$ is a nonempty, convex, compact subset of X^*
(b) for every v in X, one has

$$f^\circ(x; v) = \max\{\langle \zeta, v\rangle : \zeta \in \partial f(x)\}.$$

EXAMPLE 1. Consider $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) = x^2 \sin \frac{1}{x}$. Then f is differentiable at 0, Lipschitz near 0 and $f'(0) = 0$. We will show that $\partial f(0) = [-1, 1]$. Let us take $v \geq 0$, then we can calculate:

$$\begin{aligned} f^\circ(0; v) &= \lim_{y\to 0}\sup_{t\downarrow 0} \frac{(y+tv)^2 \sin \frac{1}{y+tv} - y^2 \sin \frac{1}{y}}{t} \\ &= \lim_{y\to 0}\sup_{tv\downarrow 0} \frac{v((y+tv)^2 \sin \frac{1}{y+tv} - y^2 \sin \frac{1}{y})}{tv} \\ &= \lim_{y\to 0}\sup\,(2vy \sin \frac{1}{y} - v \cos \frac{1}{y}). \end{aligned}$$

Since $\lim_{y\to 0} y \sin \frac{1}{y} = 0$, we have

$$\lim_{y\to 0}\sup\,(2vy \sin \frac{1}{y} - v \cos \frac{1}{y}) = \lim_{y\to 0}\sup\,(-v \cos \frac{1}{y}) = v.$$

By analogy for $v < 0$ we will get $f^\circ(0; v) = -v$, thus

$$f^\circ(0; v) = \begin{cases} v & \text{for } v \geq 0 \\ -v & \text{for } v < 0 \end{cases}$$

i.e. $f^\circ(0; v) = | v |$. Thus $\partial f(0)$ consists of those ζ satisfying $| v | \geq \zeta v$ for all v; that is, $\partial f(0) = [-1, 1]$.

2. Generalized differential quotients (GDQs)

If X, Y are metric spaces and $SVM(X, Y)$ denotes the set of all set-valued maps from X to Y, then $SVM_{comp}(X, Y)$ will denote the subset of $SVM(X, Y)$ whose members are the set-valued maps from X to Y that have a compact graph. We say that a sequence $\{F_j\}_{j\in\mathbb{N}}$ of members of $SVM_{comp}(X, Y)$ *inward graph-converges* to an $F \in SVM_{comp}(X, Y)$ - and write $F_j \xrightarrow{igr} F$ - if for every open subset Ω of $X \times Y$ such that $Gr(F) \subseteq \Omega$ there exists a $j_\Omega \in \mathbb{N}$ such that $Gr(F_j) \subseteq \Omega$ whenever $j \geq j_\Omega$ ($Gr(F)$ denotes *the graph* of F, where $Gr(F) = \{(x, y) : y \in F(x)\}$).

DEFINITION 3. *Assume that X, Y are metric spaces. A* regular set-valued map *from X to Y is a set-valued map F such that for every compact subset K of X, the* *restriction $F\lceil K$ of F to K belongs to $SVM_{comp}(K,Y)$ and is a limit - in the sense of inward graph-convergence - of a sequence of continuous single-valued maps from K to Y.*

We use $REG(K,M)$ to denote the set of all regular set-valued maps from K to M.

When $f : X \to Y$ is a single-valued map, then f belongs to $REG(X,Y)$ if and only if f is continuous.

An important class of examples of regular maps is provided by the following results.

THEOREM 1. *[5] Assume that K is a compact metric space, Y is a normed space, and C is a convex subset of Y. Let $F \in SVM(K,C)$ be a set-valued map such that the graph of F is compact and the value $F(x)$ is a nonempty convex set for every $x \in K$. Then F is regular as a map from K to C.*

Thus compactness of the graph of F, in principal, guarantees regularity of F. We can relax this assumption imposing compactness of the values of F and adding upper semicontinuity of F.

THEOREM 2. *[5] Assume that X is a metric space, Y is a normed space, and C is a convex subset of Y. Let $F \in SVM(X,C)$ be an upper semicontinuous set-valued map with nonempty compact convex values. Then $F \in REG(X,C)$.*

THEOREM 3. *[5] Assume that X, Y, Z are metric spaces. Let $F \in SVM(X,Y)$, $G \in SVM(Y,Z)$. Then the composite map $G \circ F$ belongs to $REG(X;Z)$.*

DEFINITION 4. *Let $m,n \in \mathbb{Z}_+$, $F : \mathbb{R}^m \twoheadrightarrow \mathbb{R}^n$ be a set-valued map, $\overline{x} \in \mathbb{R}^m$, $\overline{y} \in \mathbb{R}^n$ ($\overline{y} \in F(\overline{x})$) and let Λ be a nonempty compact subset of $\mathbb{R}^{n\times m}$. Then an element of Λ is a $n\times m$ matrix. Let S be a subset of $\mathbb{R}^m$. We say that Λ is a generalized differential quotient (abbr. "GDQ") of F at ($\overline{x}$,$\overline{y}$) in the direction of S, and write $\Lambda \in GDQ(F;\overline{x},\overline{y};S)$ if for every positive real number δ there exist U, G such that*
1. U is a compact neighborhood of 0 in $\mathbb{R}^m$ and $U \cap S$ is compact;
2. G is a regular set-valued map from $\overline{x} + U \cap S$ to the δ-neighborhood Λ^δ of Λ in $\mathbb{R}^{n\times m}$;
3. $G(x)\cdot(x-\overline{x}) \subseteq F(x) - \overline{y}$ for every $x - \overline{x} \in U \cap S$

The above definition of GDQ can be extended to the one defined for manifolds (see [5]).

GDQs are generalizations of Clarke generalized gradients and, like other generalized differentials introduced by Sussmann, are not unique. Every compact overset $\tilde{\Lambda}$ of $\Lambda \in GDQ(F;\overline{x},\overline{y};S)$ also belongs to $GDQ(F;\overline{x},\overline{y};S)$.

Let X, Y be finite-dimensional real linear spaces. A GDQ of $F : X \twoheadrightarrow Y$ at a point $(\overline{x},\overline{y})$ can be identified with a set of nonempty compact linear multimaps from

X to Y, where a linear multimap from X to Y is a subset of $Lin(X, Y)$.
Note that *a linear multimap* is a set of linear maps whilst *a set-valued map* is a map with values that are sets.

EXAMPLE 2. Let $f(x) = x \sin \frac{1}{x}$ and $f(0) = 0$. This function is not differentiable in the classical way. We will show that $[-1, 1] \in GDQ(f; 0, 0; \mathbb{R})$, i.e. that for every $\delta > 0$ there exist $U,\ G$ such that
1. U is a compact neighborhood of 0 in $\mathbb{R}$
2. G is a regular set-valued map from U to the δ-neighborhood Λ^δ of Λ
3. $G(x) \cdot x = f(x)$ for every $x \in U$.
Let $U = [-1, 1]$ and $G(x) = \begin{cases} \sin \frac{1}{x} & \text{if } x \neq 0 \\ [-1, 1] & \text{if } x = 0 \end{cases}$
Then points 1. and 3. are satisfied. One can show using Theorem 2 that G is a regular set-valued map. Hence the point 2. is also satisfied.

EXAMPLE 3. Let us consider the function $f : \mathbb{R} \to \mathbb{R}$ defined as follows $f(x) = x^2 \sin(\frac{1}{x})$ and $f(0) = 0$. Then f is Lipschitz and differentiable in the classical sense. One can show that $[-1, 1] \in GDQ(f; 0, 0; \mathbb{R})$ and $[-1, 1]$ is also the Clarke generalized gradient of f at 0. But for Lipschitz maps usually exist generalized differentials smaller then Clarke's generalized gradient, like in our case: $\{f'(0)\} \in GDQ(f; 0, 0; \mathbb{R})$

3. Minimal GDQs

In Example 3 we have seen that there are many GDQs: there exist Clarke generalized gradient and classical derivative and both of them are GDQs for the same function in the same point and in the same direction. This leads us to the following question: which GDQ one should deal with? We are interested in minimal GDQs (in the sense of inclusion of sets). They are the simplest. We are going to show that such minimal GDQs exist. In [3] we studied minimal multidifferentials of set-valued maps. Multidifferential is another concept of a generalized differential introduced by Sussmann.

LEMMA 1. *[2] Let U be an open subset of a topological space X. If the family $\{K_s\}_{s\in S}$ of closed subsets of the space X includes at least one compact set and $\bigcap\limits_{s\in S} K_s \subset U$ then there exists such a finite set $\{s_1, ..., s_n\} \subset S$ such that $K_{s_1} \cap K_{s_2} \cap ... \cap K_{s_n} \subset U$.*

The following lemma is easy to prove.

LEMMA 2. *Let $\Lambda_k,\ k \in \mathbb{N}$ be a sequence of descending compact sets*

$$\Lambda_1 \supset \Lambda_2 \supset ...$$

and let $\bigcap\limits_k \Lambda_k = \Lambda$ (then Λ is compact and $\Lambda \neq \emptyset$). Then for every $\delta > 0$ there exist δ'

and $k \in \mathbb{N}$ such that

$$\Lambda_k \subset \Lambda_k^{\delta'} \subset \Lambda^{\delta}$$

THEOREM 4. *Let $F \in SVM(\mathbb{R}^n, \mathbb{R}^m)$ and $\Lambda_k \in GDQ(F; \overline{x}, \overline{y}; \mathbb{R}^n)$ for $k = 1, 2, \ldots$. Let us assume that $\Lambda_1 \supset \Lambda_2 \supset \ldots$ and $\bigcap_k \Lambda_k = \Lambda$. ($\Lambda \neq \emptyset$ and Λ is compact as implied by definition of Λ_k). Then*

$$\Lambda \in GDQ(F; \overline{x}, \overline{y}; \mathbb{R}^n)$$

Proof. Without loss of generality we can assume that $\overline{x} = 0$ and $\overline{y} = 0$. We have to show that for every $\delta > 0$ there exist U_δ and G_δ such that conditions of definition of GDQ are satisfied for Λ.

From the assumption we have that for every $\delta > 0$ and every $k \in \mathbb{N}$ there exist $U_{\delta,k}$ and $G_{\delta,k}$ such that $U_{\delta,k}$ is a compact neighborhood of 0 in $\mathbb{R}^n$, $G_{\delta,k} \in REG(U_{\delta,k}, \Lambda_k^{\delta})$ where Λ_k^{δ} is a δ-neighborhood of Λ_k in $\mathbb{R}^{n\times m}$ and $G_{\delta,k}(x) \cdot x \subseteq F(x)$ for every $x \in U_{\delta,k}$ and $k \in \mathbb{N}$.

It is sufficient to show that for every $\delta > 0$ there exist $\delta' < 0$ and Λ_k with its neighborhood $\Lambda_k^{\delta'}$ contained in δ-neighborhood of Λ. But this fact follows from Lemma 2. So if we put $U_\delta = U_{\delta',k}$ and $G_\delta = G_{\delta',k}$ then from the arbitrary of choice of δ we get what we wanted to show. □

COROLLARY 1. *If the set of GDQs of a set-valued map F at $(\overline{x}, \overline{y})$ is not empty then there exists in this set at least one minimal GDQ at this point in the sense of inclusion of sets.*

Proof. From Kuratowski-Zorn Lemma, a family of sets with the property that descending sequences have a lower bound, possesses a minimal element. In our case such a sequence (D_k) of GDQs has a lower bound – their intersection D. Thus there exists for the family of GDQ a minimal element D, which we call a minimal GDQ .

□

COROLLARY 2. *An intersection of every descending sequence of $GDQs$ of F at $(\overline{x}, \overline{y})$ in the direction of S is again a GDQ of F at the same point in the same direction.*

Proof. Immediately from the definition of GDQ and from Theorem 4. □

COROLLARY 3. *Every element Λ of $GDQ(F; \overline{x}, \overline{y}; S)$ contains a minimal element of $GDQ(F; \overline{x}, \overline{y}; S)$ in the sense of inclusion of sets.*

Proof. Assume that Λ does not contain such a minimal element. Let us construct a descending sequence of element of $GDQ(F; \overline{x}, \overline{y}; S)$. Then from Corollary 2 we have that intersection of this sequence is a minimal GDQ and we have a contradiction.

□

EXAMPLE 4. Consider the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) = |x|$. Then one can show that $[-1, 1] \in GDQ(f; 0, 0; \mathbb{R})$ and that this is the minimal GDQ. This interval is also Clarke generalized gradient of f at 0. However for $f(x) = x^2 \sin \frac{1}{x}$ and $f(0) = 0$ the same interval is again Clarke generalized gradient of f at 0, while the minimal GDQ is just the ordinary derivative at 0 equal 0.

References

[1] CLARKE F., *Optimization and nonsmooth analysis*, SIAM, Philadelphia 1990.

[2] ENGELKING R., *General topology*, Polish Scientific Publ., Warsaw 1977.

[3] GIREJKO E., *On multidifferentials of multifunctions*, Zeszyty Naukowe Politechniki Bialostockiej, Matematyka-Fizyka-Chemia **20** (2001), 23–35.

[4] SUSSMANN H.J., *New theories of set-valued differentials and new version of the maximum principle of optimal control theory*, in: "Nonlinear Control in the Year 2000", (Eds. Isidori A., Lamnabhi-Lagarrigue F. and Respondek W.), Springer-Verlag, London 2000, 487–526.

[5] SUSSMANN H.J., *Warga derivate containers and other generalized differentials*, in: "Proceedings of the 41stIEEE 2002 Conference on Decision and Control", Las Vegas 2002 **1**; IEEE Publications, New York 2002, 1101–1106.

[6] SUSSMANN H.J., *Path-integral generalized differentials*, in: "Proceedings of the 41stIEEE 2002 Conference on Decision and Control", Las Vegas 2002 **4**; IEEE Publications, New York 2002, 4728–4732.

AMS Subject Classification: 54C60, 49J52.

Ewa GIREJKO, Institute of Mathematics and Physics, Bialystok Technical University, Wiejska 45A,
Bialystok, 15-351, POLAND
e-mail: `egirejko@pb.bialystok.pl`

Rend. Sem. Mat. Univ. Pol. Torino - Vol. 63, 4 (2005)
Control Theory and Stabil., I

M. Guerra*

DISCONTINUOUS HAMILTONIAN FLOWS FOR NONLINEAR CONTROL SYSTEMS

Abstract. It is known that, under appropriate commutativity assumptions, smooth control systems that are affine with respect to controls can be extended into classes of generalized controls that contain impulses. A version of Pontryagin's maximum principle that applies in such cases as been obtained but it operates on a reduced system. We show how to construct generalized Hamiltonian trajectories for the original system in a way consistent with that version of Pontryagin's maximum principle. By lifting both the continuous and the discontinuous components of candidate optimal trajectories into the cotangent bundle this construction allows for an attractive geometric description of extremal trajectories.

1. Introduction and statement of the problem

Consider a nonlinear control system affine with respect to inputs

$$\dot{x}(t) = Y(x(t)) + \sum_{i=1}^{k} X_i(x(t))\, u_i(t), \tag{1}$$

where the vector fields $Y(x)$, $X_i(x)$, $i = 1, 2, ..., k$ are smooth. The state space is $\mathbb{R}^n$ and the set of admissible controls is $L^k_{\infty,loc}(\mathbb{R})$ (the space of measurable locally essentially bounded functions with domain $\mathbb{R}$ and range $\mathbb{R}^k$). For a control system of this class, the trajectory corresponding to a given control $u \in L^k_{\infty,loc}(\mathbb{R})$ and a given initial condition, $x(0) = \overline{x}$, is uniquely defined in some open interval containing the point $t = 0$. Let $x_{u,\overline{x}}$ denote this trajectory, defined in the maximal interval. For fixed $\overline{x} \in \mathbb{R}^n$, consider the set of points which are accessible from $\overline{x}$ through trajectories of system (1) in a given time $T > 0$

$$\mathcal{A}(\overline{x}, T) = \left\{ x_{u,\overline{x}}(T),\, u \in L^k_\infty[0, T] \right\}.$$

We wish to characterize the set $\partial\mathcal{A}(\overline{x}, T)$ (the boundary of $\mathcal{A}(\overline{x}, T)$, with $\overline{x}$ and T fixed). Pontryagin's maximum principle is a powerful tool to characterize the set $\mathcal{A}(\overline{x}, T) \cap \partial\mathcal{A}(\overline{x}, T)$: this must be contained in the set of endpoints of extremal trajectories with initial state $x(0) = \overline{x}$. However, it has some important shortcomings:

1. For an affine system (1), the maximum condition does not yield immediately a unique control in the form of a feedback function of state and adjoint vector;
2. The maximum principles applies to trajectories whose end point lies in the set $\partial\mathcal{A}(\overline{x}, T)$. Hence it can't be used to search for points in $\partial\mathcal{A}(\overline{x}, T) \setminus \mathcal{A}(\overline{x}, T)$. However, $\mathcal{A}(\overline{x}, T)$ is not, in general, closed.

*This article was partially supported by POCTI/MAT/41683/01 of FCT

The first of these difficulties can be partially overcome by the technique of so-called *Dirac's constraints* [2]. Andrei Sarychev [4] proved that, under suitable commutativity assumptions, system (1) can be extended by continuity into a space of generalized controls that includes impulsive controls. Typically, at least some points in $\partial\mathcal{A}(\overline{x},T)\setminus\mathcal{A}(\overline{x},T)$ become accessible when we consider such generalized controls. The fact that the extension of the system is obtained by continuity means that if we know a generalized control that steers $\overline{x}$ to $\overline{\overline{x}}\in\partial\mathcal{A}(\overline{x},T)\setminus\mathcal{A}(\overline{x},T)$ then we "know" which L^k_∞-controls steer $\overline{x}$ to points close to $\overline{\overline{x}}$. A. Sarychev also provided a version of the maximum principle that applies to generalized controls [4][1]. These results are an important improvement with respect to the second difficulty indicated above.

In this paper we combine the technique of Dirac's constraints with Sarychev's generalized controls to obtain some useful geometric properties of Sarychev extremals. We prove that every classical extremal that satisfies $x_{u,\overline{x}}(T)\in\partial\mathcal{A}(\overline{x},T)$ is also an extremal in the sense of Sarychev, but a classical extremal such that $x_{u,\overline{x}}(T)\notin\partial\mathcal{A}(\overline{x},T)$ may fail to be an extremal in the sense of Sarychev. We give a characterization of the classical extremals that are also Sarychev extremals. Our results also allow to compute the Sarychev extremals and/or to characterize their qualitative structure using computations that are simpler then those involved in the direct application of Sarychev's version of the maximum principle. In the end of the paper we provide an example in which these tools are used to describe the qualitative structure of the extremals in an optimal control problem.

This paper contains a partial generalization of results published in [3] concerning the case of linear-quadratic optimal control problems of arbitrary order of singularity. The generalization is only partial because here we deal essentially with the nonlinear analogous to L-Q problems whose order of singularity equals one. Research is being carried in order to obtain similar results for other types of singularity.

The results presented in this paper can be readily extended to systems where the fields Y, X_i, are time variant and/or systems whose state space is a connected smooth manifold of finite dimension. However, for the sake of simplicity, we only present the autonomous case with state space $\mathbb{R}^n$.

2. Notation, basic definitions and assumptions

In this paper, $\phi : L^k_{1,loc}(\mathbb{R}) \mapsto L^k_{1,loc}(\mathbb{R})$, denotes the "primitivation" operator, i.e., $\phi u(t) = \int_0^t u(\tau)\,d\tau$, $\forall t\in\mathbb{R}$, $\forall u\in L^k_{1,loc}(\mathbb{R})$.

A smooth vector field, F, is identified with the operator $F : C_\infty \mapsto C_\infty$ defined by $(Fg)(x) = Dg(x)\cdot F(x)$. A (local) diffeomorphism, $P:\mathbb{R}^n\mapsto\mathbb{R}^n$ generates an operator, AdP, acting in the space of smooth vector fields and defined as $(AdPF)\,g = F\left(g\circ P^{-1}\right)\circ P$, for all $g\in C_\infty$. In any system of local coordinates, the vector field $AdPF$ can be represented as $(AdPF)(x) = (DP(x))^{-1}F(P(x))$. We also use the Lie bracket, defined in the usual way: $[F,G]\,h = F(Gh) - G(Fh)$. In local coordinates this is $[F,G](x) = DG(x)\cdot F(x) - DF(x)\cdot G(x)$.

For brevity we will indicate the sum $\sum_{i=1}^{k} X_i u_i$ as Xu. Thus system (1) is indicated by $\dot{x} = Y + Xu$, being always understood that u is k-dimensional.

Let F_t denote a time-variant vector field in $\mathbb{R}^n$. $\Phi_t^{F_\tau \, d\tau}$ denotes the (local) flow generated by the time-variant vector field F_t. i.e., the map $(t, \overline{x}) \mapsto \Phi_t^{F_\tau \, d\tau}\overline{x}$ solves the differential equation

$$\text{(2)} \qquad \frac{d}{dt}\Phi_t^{F_\tau \, d\tau}\overline{x} = F_t\left(\Phi_t^{F_\tau \, d\tau}\overline{x}\right), \qquad \Phi_0^{F_\tau \, d\tau}\overline{x} = \overline{x}.$$

This should not be confused with the (local) flow $\Phi_t^{F_\theta \, d\tau}$, which is the unique solution of the autonomous differential equation

$$\text{(3)} \qquad \frac{d}{dt}\Phi_t^{F_\theta \, d\tau}\overline{x} = F_\theta\left(\Phi_t^{F_\theta \, d\tau}\overline{x}\right), \qquad \Phi_0^{F_\theta \, d\tau}\overline{x} = \overline{x},$$

where θ acts as a fixed parameter, independent of the "time" variable. If these flows are well defined at least locally, then, for each sufficiently small $t \in \mathbb{R}$, the map $\overline{x} \mapsto \Phi_t^{F_\theta \, d\tau}\overline{x}$ is a local diffeomorphism.

A (possibly time-variant) vector field, F_t, generates an Hamiltonian function with domain in the cotangent bundle, $T^*\mathbb{R}^n \sim \mathbb{R}^{2n}$, defined by $h_{F_t}(x, \xi) = \xi F_t(x)$. An Hamiltonian function, h_{F_t}, defines an Hamiltonian vector field, $\overrightarrow{h}_{F_t}$, represented in local coordinates by $\overrightarrow{h}_{F_t}(x, \xi) = \left(\frac{\partial h_{F_t}}{\partial \xi}(x, \xi), -\frac{\partial h_{F_t}}{\partial x}(x, \xi)\right)$. We use these definitions also in the case when F depends on a control, i.e., a control system $\dot{x} = F_t(x, u)$ defines an Hamiltonian function, $h_{F_t(\cdot, u)}(x, \xi) = \xi F_t(x, u)$, that depends jointly on the state, adjoint vector, time and control. The Hamiltonian system generated by a control system is, from this point of view, a new control system,

$$\text{(4)} \qquad \dot{x}(t) = \frac{\partial h_{F_t(\cdot, u(t))}}{\partial \xi}(x(t), \xi(t)), \qquad \dot{\xi}(t) = -\frac{\partial h_{F_t(\cdot, u(t))}}{\partial x}(x(t), \xi(t)),$$

whose trajectories lie in the cotangent bundle, $T^*\mathbb{R}^n \sim \mathbb{R}^{2n}$. In this perspective, the maximum principle states that the trajectories whose end-points lie in $\partial\mathcal{A}(\overline{x}, T)$ are to be found among the projections into the state space of the trajectories of the Hamiltonian system (4) that satisfy the maximum condition

$$\text{(5)} \qquad h_{F_t(\cdot, u(t))}(x(t), \xi(t)) = \max_{v} h_{F_t(\cdot, v)}(x(t), \xi(t)), \qquad a.e.\, t \in [0, T].$$

We will show how this point of view can be useful to deal with generalized controls and generalized trajectories.

Through all our paper we will assume that the following assumptions hold:

A1: The fields $Y, X_1, X_2, ..., X_k$ are complete, i.e., $\Phi_t^{Y \, d\tau}x$, $\Phi_t^{X_i \, d\tau}x$, $i = 1, 2, ..., k$ are uniquely defined for all $t \in \mathbb{R}$, $x \in \mathbb{R}^n$.

A2: The fields X_i, $i = 1, 2, ..., k$ commute, i.e., $\left[X_i, X_j\right] \equiv 0$, $\forall i, j \in \{1, 2, ..., k\}$.

3. Dirac's constraints

In the case of an affine control system (1), the maximum condition (5) reduces to

(6) $$\xi(t) X_i(x(t)) = 0, \qquad \forall t \in [0,T], \ i = 1,2,...,k.$$

These are what Dirac [2] called the *primary constraints* of the system. These constraints imply

(7) $$\frac{d^j}{dt^j}\xi(t) X_i(x(t)) = 0, \qquad \forall t \in [0,T], \ i = 1,2,...,k, \ j \in \mathbb{N}.$$

These are called *Dirac's secondary constraints of j^{th} order*.

The trajectories which satisfy the Pontryagin maximum principle are projections into the state space of trajectories of the Hamiltonian system (4) that satisfy *all* Dirac's constraints. Since we will need to distinguish these trajectories from "generalized trajectories" that satisfy a special version of the maximum principle, we introduce the following definition:

DEFINITION 1. *A trajectory of system* (1) *is a classical extremal if it is the projection into the state space of some trajectory of the Hamiltonian system* (4) *which satisfies all Dirac's constraints* (6, 7).

In favorable cases, the set of primary and secondary constraints can be reduced to a set of equations of the type $\Psi_j(x,\xi) = 0, \ j = 1,2,...,m, \ u = \omega(x,\xi)$. Hence the maximum principle can be used in much the same way as in the cases where it gives immediately the control as a unique feedback function of the state and adjoint vector. The main difference is that in system (1), the Hamiltonian flow is restricted to the Dirac set:

$$\mathcal{D} = \left\{(x,\xi) \in \mathbb{R}^{2n} : \Psi_j(x,\xi) = 0, \ j = 1,2,...,m\right\}.$$

The first and second-order secondary constraints are specially important for our results. It can be easily checked that, due to assumption **A2**, they are, respectively

(8) $$\xi(t)[Y, X_i](x(t)) = 0, \qquad \forall t \in [0,T], \ i = 1,2,...,k;$$

(9) $$\xi(t)[Y,[Y,X_i]](x(t)) + \sum_{l=1}^{k} \xi(t)[X_l,[Y,X_i]](x(t)) u_l(t) = 0,$$
$$a.e.\ t \in [0,T], \ i = 1,2,...,k.$$

4. Generalized controls

Here we provide a short outline of Sarychev's construction of the spaces of generalized controls and generalized trajectories. For details, see [4]. The following Theorem is fundamental

THEOREM 1. *Let $Y_t(x)$ denote a time-variant vector field, smooth with respect to x, absolutely continuous with respect to t, and let $X_t(x) = \left(X_{1,t}(x), X_{2,t}(x), ..., X_{k,t}(x)\right)$ denote an array of k time-variant complete vector fields, smooth with respect to x, absolutely continuous with respect to t. Then, for each $u = (u_1, u_2, ..., u_k) \in L^k_{1,loc}$, and every sufficiently small t, we have*

$$\Phi_t^{Y_\tau + X_\tau u(\tau)\, d\tau} = $$
$$= \Phi_1^{X_t \phi u(t)\, d\tau}\, \Phi_t^{Y_\tau + \int_0^1 Ad\Phi_{\theta_1}^{X_\tau \phi u(\tau)\, d\theta_2}\left([X_\tau \phi u(\tau), Y_\tau + \theta_1 X_\tau u(\tau)] - \dot{X}_\tau \phi u(\tau)\right)\, d\theta_1\, d\tau}.$$

Now, fix an arbitrary $T > 0$ and consider the space $L^k_\infty[0, T]$ provided with the norm $\|u\|_1 = \|\phi u\|_{L^k_1[0,T]} + |\phi u(T)|$. This norm is weaker then the usual L_1-norm. For each $\alpha \in]0, +\infty[$, let U_α denote the topological completion of the set $\left\{u \in L^k_\infty[0, T] : \|\phi u\|_{L^k_\infty[0,T]} \le \alpha\right\}$, with respect to the norm $\|\cdot\|_1$. Under assumptions **A1** and **A2**, Theorem 1 has the following Corollary (see Theorem 4.1 in [4]).

COROLLARY 1. *For each $u = (u_1, u_2, ..., u_k) \in L^k_1[0, T]$, and every $t \in [0, T]$, the corresponding flow of system (1) can be represented as*

$$\Phi_t^{Y + Xu(\tau)\, d\tau} = \Phi_1^{X\phi u(t)\, d\tau}\, \Phi_t^{Y - \int_0^1 Ad\Phi_{\theta_1}^{X\phi u(\tau)\, d\theta_2}[Y, X\phi u(\tau)]\, d\theta_1\, d\tau}. \tag{10}$$

For every $\overline{x} \in \mathbb{R}^n$ and every $\alpha \in]0, +\infty[$, the transformation $u \mapsto \Phi_{(\cdot)}^{Y+Xu(\tau)\, d\tau}\overline{x}$ is a uniformly continuous map from U_α into $L^n_1[0, T]$.

For brevity, let G denote the controlled vector field $G(x, v) = -\int_0^1 Ad\Phi_{\theta_1}^{Xv\, d\theta_2}[Y, Xv](x)\, d\theta_1,\ v \in \mathbb{R}^k$. Corollary 1 shows that there exists one unique extension of the map $u \mapsto \Phi_{(\cdot)}^{Y+Xu(\tau)\, d\tau}\overline{x}$ into the space $U = \bigcup_{\alpha \in]0,+\infty[} U_\alpha$ that is continuous in U_α, for every sufficiently large $\alpha > 0$. Thus, we have the following Definition:

DEFINITION 2. *Any element of U is called a generalized control. The generalized trajectory corresponding to a given generalized control, $u \in U$, with a given initial condition, $x(0) = \overline{x}$, is the function*

$$x_{u,\overline{x}}(t) = \Phi_1^{X\phi u(t)\, d\tau}\, \Phi_t^{Y + G(\cdot, \phi u(\tau))\, d\tau}\overline{x}.$$

It can be shown that U is a subspace of the Sobolev space $W_{-1,\infty}$. Notice that, for $u \in U \backslash L^k_1[0, T]$, $x_{u,\overline{x}}(t)$ is defined only as an L_1-function (i.e., for almost every $t \in [0, T]$) but the value of $\Phi_T^{Y+Xu(\tau)\, d\tau}\overline{x}$ is uniquely defined because the norm $\|\cdot\|_1$ distinguishes different values of $\phi u(T)$. Let $\widetilde{\mathcal{A}}(\overline{x}, T) = \left\{\Phi_T^{Y+Xu(\tau)\, d\tau}\overline{x} : u \in U\right\}$, denote the set of points which are accessible from the point $\overline{x}$ in time T, through generalized trajectories. Since every ordinary trajectory is also a generalized trajectory, it is clear that $\mathcal{A}(\overline{x}, T) \subset \widetilde{\mathcal{A}}(\overline{x}, T)$. The continuity of the map $u \mapsto \Phi_{(\cdot)}^{Y+Xu(\tau)\, d\tau}\overline{x}$ implies that $\widetilde{\mathcal{A}}(\overline{x}, T) \subset \overline{\mathcal{A}(\overline{x}, T)}$.

5. Maximum principle for generalized controls

Sarychev proved a version of the maximum principle that applies to the time-optimal generalized trajectories of an affine time-variant control system [4],[1]. Below we present the corresponding version for the accessibility problem. The proof in [4] can easily be adapted to our case.

THEOREM 2. *Let $\widehat{u} \in U$, such that $x_{\widehat{u},\overline{x}}(T) \in \partial\widetilde{\mathcal{A}}(\overline{x}, T)$. Then, there exists an adjoint vector $\overline{\xi} \in \mathbb{R}^n$ such that the trajectory of the Hamiltonian system $\left(\dot{x}, \dot{\xi}\right) = \overrightarrow{h}_{Y+G(\cdot,\phi\widehat{u})}(x, \xi)$, $(x(0), \xi(0)) = \left(\overline{x}, \overline{\xi}\right)$ satisfies*

1. $h_{Y+G(\cdot,\phi\widehat{u}(t))}(x(t), \xi(t)) = \max_{\upsilon\in\mathbb{R}^k} h_{Y+G(\cdot,\upsilon)}(x(t), \xi(t))$, *a.e.* $t \in [0, T]$;
2. $\xi(T) X(x(T)) = 0$.

Sketch of the proof. Let $\widehat{u} \in U$, such that $x_{\widehat{u},\overline{x}}(T) \in \partial\widetilde{\mathcal{A}}(\overline{x}, T)$. From Corollary 1, we have $x_{\widehat{u},\overline{x}}(T) = \Phi_1^{X\phi\widehat{u}(T)} \Phi_T^{Y+G(\cdot,\phi\widehat{u}(\tau))\,d\tau}\overline{x}$. Since $x \mapsto \Phi_1^{X\phi\widehat{u}(T)}x$ is a diffeomorphism, it follows that $\Phi_T^{Y+G(\cdot,\phi\widehat{u}(\tau))\,d\tau}\overline{x}$ must lie in the boundary of the accessible set of the system

$$\dot{x} = Y(x) + G(x, \upsilon), \qquad x(0) = \overline{x}, \qquad \upsilon \in L_\infty^k[0, T]. \tag{11}$$

Let K denote the Pontryagin cone for system (11) at the point $\Phi_T^{Y+G(\cdot,\phi\widehat{u}(\tau))\,d\tau}\overline{x}$, and let S denote the tangent space to the integral manifold of X at the point $\Phi_T^{Y+G(\cdot,\phi\widehat{u}(\tau))\,d\tau}\overline{x}$. Now, $\widehat{u} \in U$, implies that $\phi\widehat{u}$ is defined only as a function of class $L_\infty^k[0, T]$, i.e., only "almost every where". This means that $\phi\widehat{u}(T)$ (the value of $\phi\widehat{u}$ at the particular point $t = T$) can be chosen independently of $\phi\widehat{u}$ as an element of $L_\infty^k[0, T]$. It follows that $K + S$ is an approximation of $\widetilde{\mathcal{A}}(\overline{x}, T)$ in the neighborhood of $x_{\widehat{u},\overline{x}}(T)$. The proof follows by a procedure analogous to the classical proof of the Pontryagin maximum principle, by using $K + S$ instead of the Pontryagin cone of system (1), which may fail to exist at the point $x_{\widehat{u},\overline{x}}(T)$. □

This version of the maximum principle gives a second definition of extremal trajectory.

DEFINITION 3. *A generalized trajectory, $x \in L_1^n[0, T]$ is a Sarychev extremal with initial condition $x(0) = \overline{x}$ if it can be represented in the form*

$$x(t) = \Phi_1^{X\phi u(t)\,d\tau} \Phi_t^{Y+G(\cdot,\phi u(\tau))\,d\tau}\overline{x},$$

for some $u \in U$ which satisfies the conditions (1) *and* (2) *of Theorem 2.*

We will show that Definition 1 includes curves that are not Sarychev extremals. Hence, Theorem 2 is a necessary condition that is stronger than the classical maximum principle.

6. Generalized Hamiltonian flows

Theorem 2 states that we should look for extremals of the reduced system, $\dot{x} = Y(x) + G(x, v)$ and then recover the extremals for the original system (1) by using equality (10). Now we show that the extension of system (1) into the class of generalized controls induces naturally an extension of the Hamiltonian flow (understood as the flow of the Hamiltonian system (4)) into the same class of generalized controls. This gives useful insights into the structure of Sarychev extremals. It also gives a method to compute the Sarychev extremals without having to compute explicitly the field G.

Any smooth map, $\varphi : \mathbb{R}^m \mapsto \mathbb{R}^n$, generates a map, $T\varphi$, that assigns to each vector $V \in T_x\mathbb{R}^n$, a new vector, $T\varphi V \in T_{\varphi(x)}\mathbb{R}^n$, defined by $(T\varphi V) f = V(f \circ \varphi)$, $\forall f \in C_\infty$. If $\varphi : \mathbb{R}^n \mapsto \mathbb{R}^n$ is a (local) diffeomorphism, then it also defines a map, φ_*, which assigns to each adjoint vector $\xi \in T_x^*\mathbb{R}^n$, a new adjoint vector $\varphi_*\xi \in T^*_{\varphi(x)}\mathbb{R}^n$, defined by $(\varphi_*\xi)\, V = \xi T\left(\varphi^{-1}\right) V$, $\forall V \in T_{\varphi(x)}\mathbb{R}^n$. $\varphi_*\xi$ is called the *pushforward* of ξ by the map φ.

LEMMA 1. *Let $F_t(x)$ denote a time-variant vector field, smooth with respect to x, absolutely continuous with respect to t. Then,*

$$\Phi_t^{\overrightarrow{h}_{F_\tau}\, d\tau} = \left(\Phi_t^{F_\tau\, d\tau}\right)_* .$$

Proof. Consider a fixed $\overline{x} \in \mathbb{R}^n$, $\overline{\xi} \in T_{\overline{x}}\mathbb{R}^n$. In local coordinates, we have $\left(\Phi_t^{F_\tau\, d\tau}\right)_* \overline{\xi} = \overline{\xi} D_x \Phi_t^{F_\tau\, d\tau}\overline{x}$. Use the Theorem of differentiability of the solution of an ODE with respect to initial conditions to prove that the pair $\left(\Phi_t^{F_\tau\, d\tau}\overline{x}, \overline{\xi} D_x \Phi_t^{F_\tau\, d\tau}\overline{x}\right)$ is the unique solution of the Hamiltonian system $\left(\dot{x}, \dot{\xi}\right) = \overrightarrow{h}_{F_t}(x, \xi)$, with initial point $(x(0), \xi(0)) = \left(\overline{x}, \overline{\xi}\right)$. $\square$

Using Corollary 1, we obtain a representation of the Hamiltonian flow in the form of a composition of flows that depends only on the primitives of the controls.

LEMMA 2. *For any $u \in L_1^k[0, T]$, we have*

$$\Phi_t^{\overrightarrow{h}_{Y+Xu(\tau)}\, d\tau} = \Phi_1^{\overrightarrow{h}_{X\phi u(t)}\, d\tau}\, \Phi_t^{\overrightarrow{h}_{Y+G(\cdot,\phi u(\tau))}\, d\tau}.$$

For any fixed $\left(\overline{x}, \overline{\xi}\right) \in \mathbb{R}^{2n}$ and any $\alpha \in\,]0, +\infty[$, the map $u \mapsto \Phi_t^{\overrightarrow{h}_{Y+Xu(\tau)}\, d\tau}\left(\overline{x}, \overline{\xi}\right)$ is uniformly continuous from U_α into $L_1^{2n}[0, T]$.

Proof. Lemma 1 states that $\Phi_t^{\overrightarrow{h}_{Y+Xu(\tau)}\, d\tau} = \left(\Phi_t^{Y+Xu(\tau)\, d\tau}\right)_*$. By Corollary 1, this is $\Phi_t^{\overrightarrow{h}_{Y+Xu(\tau)}\, d\tau} = \left(\Phi_1^{X\phi u(t)\, d\tau}\, \Phi_t^{Y+G(\cdot,\phi u(\tau))\, d\tau}\right)_*$. Since $(\varphi \circ \psi)_* = \varphi_* \circ \psi_*$, this reduces to $\Phi_t^{\overrightarrow{h}_{Y+Xu(\tau)}\, d\tau} = \left(\Phi_1^{X\phi u(t)\, d\tau}\right)_* \left(\Phi_t^{Y+G(\cdot,\phi u(\tau))\, d\tau}\right)_*$, and the equality follows by using again Lemma 1. The continuity of the map follows immediately from Corollary 1. □

It follows from Lemma 2 that the map $u \mapsto \Phi_{(\cdot)}^{\overrightarrow{h}_{Y+Xu(\tau)}\,d\tau}\overline{x}$ has one unique extension into the space U. Thus, the generalized Hamiltonian flow corresponding to a given generalized control, $u \in U$, is defined in the same way as the generalized flow introduced in Definition 2.

We will now show that Lemma 2 provides a link between Theorem 2 and Dirac's constraints. We start with the following Proposition which shows that the primary constraints can be represented using the flow of the reduced Hamiltonian system, $\left(\dot{x}, \dot{\xi}\right) = \overrightarrow{h}_{Y+G(\cdot,\phi\widehat{u})}(x,\xi)$.

PROPOSITION 1. *Let $u \in U$. The Dirac's constraints are satisfied at every point of the trajectory $t \mapsto \Phi_t^{\overrightarrow{h}_{Y+Xu}\,d\tau}\left(\overline{x},\overline{\xi}\right)$ if and only if they are satisfied at every point of the trajectory $t \mapsto \Phi_t^{\overrightarrow{h}_{Y+G(\cdot,\phi u)}\,d\tau}\left(\overline{x},\overline{\xi}\right)$.*

Proof. Using Lemma 1, we have

$$\xi(t)\,X_i\,(x(t)) = \left(\left(\Phi_t^{Y+Xu(\tau)\,d\tau}\right)_* \overline{\xi}\right) X_i \left(\Phi_t^{Y+Xu(\tau)\,d\tau}\overline{x}\right).$$

Using Lemma 2, this reduces to

$$\begin{aligned}
&\xi(t)\,X_i\,(x(t)) = \\
&= \left(\left(\Phi_1^{X\phi u(t)\,d\tau}\right)_* \left(\Phi_t^{Y+G(\cdot,\phi u(\tau))\,d\tau}\right)_* \overline{\xi}\right) X_i \left(\Phi_1^{X\phi u(t)\,d\tau}\,\Phi_t^{Y+G(\cdot,\phi u(\tau))\,d\tau}\overline{x}\right) = \\
&= \left(\left(\Phi_t^{Y+G(\cdot,\phi u(\tau))\,d\tau}\right)_* \overline{\xi}\right) \left(Ad\Phi_1^{X\phi u(t)\,d\tau} X_i\right) \left(\Phi_t^{Y+G(\cdot,\phi u(\tau))\,d\tau}\overline{x}\right).
\end{aligned}$$

Since assumption **A2** implies $Ad\Phi_1^{X\phi u(t)\,d\tau} X_i = X_i$, this proves the Proposition. □

Proposition 1 has the following Corollary:

COROLLARY 2. *Consider fixed $\widehat{u} \in U$, $\left(\overline{x},\overline{\xi}\right) \in \mathbb{R}^{2n}$. Dirac's primary constraints are satisfied at every point of the trajectory of the Hamiltonian system*

$$\left(\dot{x}, \dot{\xi}\right) = \overrightarrow{h}_{Y+X\widehat{u}}(x,\xi), \qquad (x(0), \xi(0)) = \left(\overline{x},\overline{\xi}\right) \tag{12}$$

if and only if all the following conditions are satisfied:

1. $\overline{\xi}X_i(\overline{x}) = 0, \qquad i = 1, 2, ..., k;$
2. $\xi(t)\left[Y + G(\cdot, \phi\widehat{u}(t)), X_i\right](x(t)) = 0$, $i = 1, 2, ..., k$ *hold at almost every $t \in [0,T]$, along the trajectory of the reduced Hamiltonian system*

$$\left(\dot{x}, \dot{\xi}\right) = \overrightarrow{h}_{Y+G(\cdot,\phi\widehat{u})}(x,\xi), \qquad (x(0), \xi(0)) = \left(\overline{x},\overline{\xi}\right). \tag{13}$$

Proof. Proposition 1 states that a trajectory of system (12) satisfies Dirac's primary constraints if and only if the corresponding trajectory of system (13) satisfies them. Since

$$\begin{aligned}
&\tfrac{d}{dt}\left(\left(\Phi_t^{Y+G(\cdot,\phi u(\tau))\,d\tau}\right)_*\overline{\xi}\right) X_i \left(\Phi_t^{Y+G(\cdot,\phi u(\tau))\,d\tau}\overline{x}\right) = \\
&= \left(\left(\Phi_t^{Y+G(\cdot,\phi u(\tau))\,d\tau}\right)_*\overline{\xi}\right) \left[Y + G\left(\cdot, \phi u\left(t\right)\right), X_i\right] \left(\Phi_t^{Y+G(\cdot,\phi u(\tau))\,d\tau}\overline{x}\right),
\end{aligned}$$

we proved the Corollary. □

The next Lemma, together with Corollary 2, shows that classical extremals correspond to critical points of the reduced Hamiltonian function, $h_{Y+G(\cdot,v)}$. These may happen to be minima, saddle points or just local maxima instead of global maxima as required in Theorem 2 (see example in next Section).

LEMMA 3. $\frac{\partial}{\partial v_i} G\left(x, v\right) = \left[X_i, Y + G\left(\cdot, v\right)\right]\left(x\right)$, *for every* $\left(x, v\right) \in \mathbb{R}^{n+k}$, $i = 1, 2, ..., k$.

Proof. Assumption **A2** implies that

$$\Phi_{\theta_1}^{X\cdot(v+\delta v)\ d\theta_2} = \Phi_{\theta_1}^{X\cdot v\ d\theta_2}\Phi_{\theta_1}^{X\cdot\delta v\ d\theta_2}.$$

It follows that,

$$\begin{aligned}
G\left(x, v+\delta v\right) &= -\int_0^1 Ad\left(\Phi_{\theta_1}^{X\cdot v\ d\theta_2}\Phi_{\theta_1}^{X\cdot\delta v\ d\theta_2}\right)\left[Y, X\cdot\left(v+\delta v\right)\right]\left(x\right)d\theta_1 = \\
&= -\int_0^1 Ad\Phi_{\theta_1}^{X\cdot\delta v\ d\theta_2}\left(Ad\Phi_{\theta_1}^{X\cdot v\ d\theta_2}\left[Y, X\cdot\left(v+\delta v\right)\right]\right)\left(x\right)d\theta_1 = \\
&= -\int_0^1 \left(Id - \theta_1 D\left(X\cdot\delta v\right)\left(x\right) + o\left(\delta v\right)\right)\cdot \\
&\qquad \left(Ad\Phi_{\theta_1}^{X\cdot v\ d\theta_2}\left[Y, X\cdot\left(v+\delta v\right)\right]\right)\left(x + \theta_1 X\left(x\right)\cdot\delta v + o\left(\delta v\right)\right)\ d\theta_1 = \\
&= -\int_0^1 Ad\Phi_{\theta_1}^{X\cdot v\ d\theta_2}\left[Y, X\cdot\left(v+\delta v\right)\right]\left(x\right)d\theta_1 - \\
&\quad -\int_0^1 \theta_1 D\left(Ad\Phi_{\theta_1}^{X\cdot v\ d\theta_2}\left[Y, X\cdot v\right]\right)\cdot X\left(x\right)\cdot\delta v d\theta_1 + \\
&\quad +\int_0^1 \theta_1 D\left(X\cdot\delta v\right)\left(x\right)\cdot Ad\Phi_{\theta_1}^{X\cdot v\ d\theta_2}\left[Y, X\cdot v\right]\left(x\right)\ d\theta_1 + o\left(\delta v\right) =
\end{aligned}$$

$$
\begin{aligned}
=&G(x,v)-\int_0^1 Ad\Phi_{\theta_1}^{X\cdot v\; d\theta_2}[Y,X\cdot\delta v](x)\;d\theta_1+\\
&+\int_0^1\theta_1\left[Ad\Phi_{\theta_1}^{X\cdot v d\theta_2}[Y,X\cdot v],X\cdot\delta v\right](x)d\theta_1+o(\delta v)=\\
&=G(x,v)-\int_0^1\left[Ad\Phi_{\theta_1}^{X\cdot v\; d\theta_2}Y,Ad\Phi_{\theta_1}^{X\cdot v\; d\theta_2}(X\cdot\delta v)\right](x)\;d\theta_1+\\
&+\int_0^1\theta_1\left[Ad\Phi_{\theta_1}^{X\cdot v\; d\theta_2}[Y,X\cdot v],X\cdot\delta v\right](x)\,d\theta_1+o(\delta v)=\\
=&G(x,v)-\int_0^1\left[Ad\Phi_{\theta_1}^{X\cdot v\; d\theta_2}Y,X\cdot\delta v\right](x)\;d\theta_1+\\
&+\int_0^1\theta_1\left[Ad\Phi_{\theta_1}^{X\cdot v d\theta_2}[Y,X\cdot v],X\cdot\delta v\right](x)\;d\theta_1+o(\delta v)=\\
=&G(x,v)+\left[X\cdot\delta v,\int_0^1 Ad\Phi_{\theta_1}^{X\cdot v\; d\theta_2}Y\;d\theta_1\right](x)-\\
&-\left[X\cdot\delta v,\int_0^1\theta_1 Ad\Phi_{\theta_1}^{X\cdot v\; d\theta_2}[Y,X\cdot v]\;d\theta_1\right](x)+o(\delta v)=\\
=&G(x,v)+\left[X\cdot\delta v,Y+\int_0^1(1-\theta_1)\,Ad\Phi_{\theta_1}^{X\cdot v\; d\theta_2}[X\cdot v,Y]\,d\theta_1\right](x)-\\
&-\left[X\cdot\delta v,\int_0^1\theta_1 Ad\Phi_{\theta_1}^{X\cdot v d\theta_2}[Y,X\cdot v]\;d\theta_1\right](x)+o(\delta v)=\\
=&G(x,v)+[X\cdot\delta v,Y+G(\cdot,v)](x)+o(\delta v).
\end{aligned}
$$

□

To explore further the relationship between Dirac's constraints and Sarychev extremals, we will use the following Proposition that states an important invariance property of the integral manifolds of the distribution $\overrightarrow{h}_{X_i}$, $i=1,2,\dots,k$.

PROPOSITION 2. *For every $(x,\xi)\in\mathbb{R}^{2n}$, $v,w\in\mathbb{R}^k$, $j\in\mathbb{N}$, $i_1,i_2,\dots,i_j\in\{1,2,\dots,k\}$, we have:*

$$
\begin{aligned}
&h_{Y+G(\cdot,v+w)}(x,\xi)=h_{Y+G(\cdot,v)}\circ\Phi_1^{\overrightarrow{h}_{Xw}\,d\tau}(x,\xi);\\
&h_{\left[X_{i_j},\dots\left[X_{i_2},\left[X_{i_1},Y+G(\cdot,v+w)\right]\right]\right]}(x,\xi)=\\
&\qquad\qquad=h_{\left[X_{i_j},\dots\left[X_{i_2},\left[X_{i_1},Y+G(\cdot,v)\right]\right]\right]}\circ\Phi_1^{\overrightarrow{h}_{Xw}\,d\tau}(x,\xi).
\end{aligned}
$$

Proof. Fix arbitrary $(x,\xi)\in\mathbb{R}^{2n}$, $v,w\in\mathbb{R}^k$, and consider the function

$$
f(\varepsilon)=h_{\left[X_{i_j},\dots\left[X_{i_1},Y+G(\cdot,v+\varepsilon w)\right]\right]}\circ\Phi_1^{\overrightarrow{h}_{X(1-\varepsilon)w}\,d\tau}(x,\xi).
$$

Then,

$$f'(\varepsilon) = \frac{\partial h_{\left[X_{i_j},\ldots\left[X_{i_1},Y+G(\cdot,v+\varepsilon w)\right]\right]}}{\partial \varepsilon}\left(\Phi_1^{\overrightarrow{h}_{X(1-\varepsilon)w}\,d\tau}(x,\xi)\right)+$$

$$+\frac{\partial h_{\left[X_{i_j},\ldots\left[X_{i_1},Y+G(\cdot,v+\varepsilon w)\right]\right]}}{\partial (x,\xi)}\left(\Phi_1^{\overrightarrow{h}_{X(1-\varepsilon)w}\,d\tau}(x,\xi)\right)\cdot\frac{\partial \Phi_1^{\overrightarrow{h}_{X(1-\varepsilon)w}\,d\tau}}{\partial\varepsilon}(x,\xi)=$$

$$= h_{\left[X_{i_j},\ldots\left[X_{i_1},\frac{\partial G}{\partial\varepsilon}(\cdot,v+\varepsilon w)\right]\right]}\left(\Phi_1^{\overrightarrow{h}_{X(1-\varepsilon)w}\,d\tau}(x,\xi)\right)-$$

$$-\frac{\partial h_{\left[X_{i_j},\ldots\left[X_{i_1},Y+G(\cdot,v+\varepsilon w)\right]\right]}}{\partial (x,\xi)}\left(\Phi_1^{\overrightarrow{h}_{X(1-\varepsilon)w}\,d\tau}(x,\xi)\right)\cdot$$

$$\cdot\frac{\partial \Phi_1^{\overrightarrow{h}_{X(1-\varepsilon)w}\,d\tau}}{\partial (x,\xi)}(x,\xi)\cdot\overrightarrow{h}_{Xw}(x,\xi)=$$

$$= h_{\left[X_{i_j},\ldots\left[X_{i_1},\frac{\partial G}{\partial\varepsilon}(\cdot,v+\varepsilon w)\right]\right]}\left(\Phi_1^{\overrightarrow{h}_{X(1-\varepsilon)w}\,d\tau}(x,\xi)\right)-$$

$$-\left(\frac{\partial h_{\left[X_{i_j},\ldots\left[X_{i_1},Y+G(\cdot,v+\varepsilon w)\right]\right]}}{\partial (x,\xi)}\cdot Ad\Phi_1^{\overrightarrow{h}_{X(\varepsilon-1)w}\,d\tau}\overrightarrow{h}_{Xw}\right)\left(\Phi_1^{\overrightarrow{h}_{X(1-\varepsilon)w}\,d\tau}(x,\xi)\right).$$

Using Lemma 3 and assumption **A2**, this reduces to

$$f'(\varepsilon) = h_{\left[X_{i_j},\ldots\left[X_{i_1},\left[Xw,Y+G(\cdot,v+\varepsilon w)\right]\right]\right]}\left(\Phi_1^{\overrightarrow{h}_{X(1-\varepsilon)w}\,d\tau}(x,\xi)\right)-$$

$$-\left(\frac{\partial h_{\left[X_{i_j},\ldots\left[X_{i_1},Y+G(\cdot,v+\varepsilon w)\right]\right]}}{\partial (x,\xi)}\cdot\overrightarrow{h}_{Xw}\right)\left(\Phi_1^{\overrightarrow{h}_{X(1-\varepsilon)w}\,d\tau}(x,\xi)\right).$$

Taking into account the fact that $h_{[F_1,F_2]} = \frac{\partial h_{F_2}}{\partial(x,\xi)}\cdot\overrightarrow{h}_{F_1}$, this is

$$f'(\varepsilon) = h_{\left[X_{i_j},\ldots\left[X_{i_1},\left[Xw,Y+G(\cdot,v+\varepsilon w)\right]\right]\right]}\left(\Phi_1^{\overrightarrow{h}_{X(1-\varepsilon)w}\,d\tau}(x,\xi)\right)-$$

$$-h_{\left[Xw,\left[X_{i_j},\ldots\left[X_{i_1},Y+G(\cdot,v+\varepsilon w)\right]\right]\right]}\left(\Phi_1^{\overrightarrow{h}_{X(1-\varepsilon)w}\,d\tau}(x,\xi)\right).$$

Assumption **A2** implies

$$\left[X_{i_j},\ldots\left[X_{i_1},\left[Xw,Y+G(\cdot,v+\varepsilon w)\right]\right]\right]=$$
$$\left[Xw,\left[X_{i_j},\ldots\left[X_{i_1},Y+G(\cdot,v+\varepsilon w)\right]\right]\right],$$

and hence f is constant. Since $f(0) = h_{\left[X_{i_j},\ldots\left[X_{i_1},Y+G(\cdot,v)\right]\right]}\left(\Phi_1^{\overrightarrow{h}_{Xw}\,d\tau}(x,\xi)\right)$, and $f(1) = h_{\left[X_{i_j},\ldots\left[X_{i_1},Y+G(\cdot,v+w)\right]\right]}(x,\xi)$, this proves the Proposition. □

Now consider a generalized control, $\widehat{u} \in U$, and a point, $\left(\overline{x}, \overline{\xi}\right) \in \mathbb{R}^{2n}$. Let $\left(x^1(t), \xi^1(t)\right)$ denote the trajectory of the reduced Hamiltonian system,

$$\left(\dot{x}, \dot{\xi}\right) = \overrightarrow{h}_{Y+G(\cdot,\phi\widehat{u})}(x, \xi), \qquad (x(0), \xi(0)) = \left(\overline{x}, \overline{\xi}\right),$$

and let $\left(x^0(t), \xi^0(t)\right)$ denote the generalized trajectory of the original Hamiltonian system,

$$\left(\dot{x}, \dot{\xi}\right) = \overrightarrow{h}_{Y+X\widehat{u}}(x, \xi), \qquad (x(0), \xi(0)) = \left(\overline{x}, \overline{\xi}\right).$$

Suppose that $\widehat{u}$ satisfies Sarychev's version of the maximum principle (Theorem 2). Then Lemma 3 states that

$$h_{[X_i, Y+G(\cdot,\phi\widehat{u}(t))]}\left(x^1(t), \xi^1(t)\right) = 0,$$

at almost every $t \in [0, T]$. On the other hand, Proposition 2 and Lemma 2 imply that

$$\begin{aligned} h_{[X_i, Y+G(\cdot,\phi\widehat{u}(t))]}\left(x^1(t), \xi^1(t)\right) &= h_{[X_i, Y+G(\cdot,0)]} \circ \Phi_1^{\overrightarrow{h}_{X\phi\widehat{u}(t)}}\left(x^1(t), \xi^1(t)\right) = \\ &= h_{[X_i, Y]}\left(x^0(t), \xi^0(t)\right) = \\ &= \xi^0(t)\,[X_i, Y]\left(x^0(t)\right). \end{aligned}$$

This shows that Sarychev extremals must satisfy the first order secondary constraint but only at "almost every $t \in [0, T]$", the exception of a null but nonempty set being possible. This differs from the primary constraints, which must be satisfied at *every point* of the time interval.

At almost every $t \in [0, T]$, a Sarychev extremal must satisfy

$$\left.\frac{\partial^2 h_{Y+G(\cdot,\upsilon)}}{\partial \upsilon^2}\right|_{\upsilon=\phi\widehat{u}(t)}\left(x^1(t), \xi^1(t)\right) \leq 0.$$

By Lemma 3, we have

$$\frac{\partial^2 h_{Y+G(\cdot,\upsilon)}}{\partial \upsilon_i \partial \upsilon_j} = h_{[X_i[X_j, Y+G(\cdot,\upsilon)]]}.$$

Hence Proposition 2 implies that the matrix with entries

$$L_{ij} = h_{[X_i[X_j,Y]]}\left(x^0(t), \xi^0(t)\right) = \xi^0(t)\left[X_i, \left[X_j, Y\right]\right]\left(x^0(t)\right)$$

must be semidefinite negative at almost every $t \in [0, T]$. This is a version for generalized controls of the well known generalized Legendre-Clebsch condition. If the generalized Hamiltonian trajectory satisfies the strong Legendre-Clebsch condition (i.e., the matrix with entries $L_{ij} = \xi^0(t)\left[X_i, \left[X_j, Y\right]\right]\left(x^0(t)\right)$ is strictly negative at almost every $t \in [0, T]$), then it corresponds to local maxima of the reduced Hamiltonian function, $h_{Y+G(\cdot,\upsilon)}$, but it may fail to correspond to global maxima (see example below).

Hence the strong Legendre-Clebsch condition is not sufficient in order for a classical extremal to be a Sarychev extremal.

Now, let $\Phi^{\overrightarrow{h}_X}(x,\xi))$ denote the integral manifold of $\overrightarrow{h}_X$ through the point (x,ξ). By the same arguments as above, the maximum condition, $h_{Y+G(\cdot,\phi\widehat{u}(t))}\left(x^1(t),\xi^1(t)\right) = \max_{\upsilon\in\mathbb{R}^k} h_{Y+G(\cdot,\upsilon)}\left(x^1(t),\xi^1(t)\right)$, reduces to $h_Y\left(x^0(t),\xi^0(t)\right) = \max\left\{h_Y(x,\xi) : (x,\xi)\in\Phi^{\overrightarrow{h}_X}\left(x^0(t),\xi^0(t)\right)\right\}$. Hence, an absolutely continuous arc, $(x(t),\xi(t))$, of a generalized Hamiltonian trajectory, satisfies the maximum condition in Theorem 2 if and only if the Dirac's constraints hold along the trajectory and

$$h_Y(x(t),\xi(t)) = \max\left\{h_Y(x,\xi) : (x,\xi)\in\Phi^{\overrightarrow{h}_X}(x(t),\xi(t))\right\} \tag{14}$$

at almost every t. Notice that, if the existence of the right handside of (14) is granted, we only have to check that

$$h_Y(x(t),\xi(t)) = \max\left\{h_Y(x,\xi) : (x,\xi)\in\mathcal{D}\cap\Phi^{\overrightarrow{h}_X}(x(t),\xi(t))\right\}$$

for almost every t.

The results above allow for a geometric description of Sarychev's extremals. In order to see this, consider an affine system (1) such that every secondary constraint of order higher than 2 is redundant (i.e., the set of Dirac's constraints reduces to (6), (8) and (9)). Define the set

$$\mathcal{W} = \left\{(x,\xi)\in\mathbb{R}^{2n} : \xi X_i(x) = 0,\ i = 1,2,...,k\right\}.$$

Then the Dirac set is

$$\mathcal{D} = \{(x,\xi)\in\mathcal{W} : \xi[Y,X_i](x) = 0,\ i = 1,2,...,k\}.$$

Suppose that, for each pair $(x,\xi)\in\mathcal{W}$, the set $\left\{\upsilon : \upsilon = \arg\max h_{Y+G(\cdot,\upsilon)}(x,\xi)\right\}$ is nonempty and finite. We will describe the generalized Hamiltonian trajectories whose projection into the state space coincide with the Sarychev extremals with initial condition $x(0) = \overline{x}$ ($\overline{x}$ fixed but arbitrary).
We can choose for initial adjoint vector any $\overline{\xi}$ such that $\left(\overline{x},\overline{\xi}\right)\in\mathcal{W}$. By choosing different $\overline{\xi}$, we are choosing different sets $\left\{\upsilon : \upsilon = \arg\max h_{Y+G(\cdot,\upsilon)}\left(\overline{x},\overline{\xi}\right)\right\}$. For each fixed $\overline{\xi}$, the choice of a $\upsilon\in\left\{\upsilon : \upsilon = \arg\max h_{Y+G(\cdot,\upsilon)}\left(\overline{x},\overline{\xi}\right)\right\}$ amounts to the choice of a point $\left(x\left(0^+\right),\xi\left(0^+\right)\right)\in\mathcal{D}\cap\Phi^{\overrightarrow{h}_X}\left(\overline{x},\overline{\xi}\right)$, which maximizes $h_Y(x,\xi) = \xi Y(x)$ over the set $\mathcal{D}\cap\Phi^{\overrightarrow{h}_X}\left(\overline{x},\overline{\xi}\right)$. If in the $\mathcal{D}$-neighborhood of $\left(x\left(0^+\right),\xi\left(0^+\right)\right)$ there exists a feedback, $u = u(x,\xi)$ which satisfies the second-order secondary constraints (9) along the curve $\Phi_t^{\overrightarrow{h}_{Y+Xu}}\left(x\left(0^+\right),\xi\left(0^+\right)\right)$ and, for every sufficiently small $t > 0$, $h_Y\circ\Phi_t^{\overrightarrow{h}_{Y+Xu}}\left(x\left(0^+\right),\xi\left(0^+\right)\right)$ is maximal among the set

$$\mathcal{D}\cap\Phi^{\overrightarrow{h}_X}\Phi_t^{\overrightarrow{h}_{Y+Xu}}\left(x\left(0^+\right),\xi\left(0^+\right)\right),$$

then we may follow the curve $\Phi_t^{\overrightarrow{h}_{Y+Xu}}\left(x\left(0^+\right),\xi\left(0^+\right)\right)$ during at least a small interval of time. If $[0,t_1]$ is the maximum interval of time where the feedback u satisfies the stated conditions, then a new jump must occur at time $t = t_1$, this time to a point $\left(x\left(t_1^+\right),\xi\left(t_1^+\right)\right) \in \mathcal{D} \cap \Phi^{\overrightarrow{h}_X}\left(x\left(t_1\right),\xi\left(t_1\right)\right)$, which maximizes $h_Y(x,\xi)$ over the set $\mathcal{D} \cap \Phi^{\overrightarrow{h}_X}\left(x\left(t_1\right),\xi\left(t_1\right)\right)$, and so on. If at some point of the generalized trajectory, $\left(x\left(\widehat{t}\right),\xi\left(\widehat{t}\right)\right)$ there exists more than one $\upsilon = \arg\max h_{Y+G(\cdot,\upsilon)}\left(x\left(\widehat{t}\right),\xi\left(\widehat{t}\right)\right)$ (i.e., there exists more than one (x,ξ) which maximizes $h_Y(x,\xi)$ over the set $\mathcal{D} \cap \Phi^{\overrightarrow{h}_X}\left(x\left(\widehat{t}\right),\xi\left(\widehat{t}\right)\right)$) then we may have alternative jumps at time $t = \widehat{t}$, corresponding to different generalized Hamiltonian trajectories. At the final time, $t = T$, we are free to choose as end-point of the trajectory any point in $\Phi^{\overrightarrow{h}_X}\left(x\left(T^-\right),\xi\left(T^-\right)\right)$ (not necessarily in $\mathcal{D}$).

7. An Example

Consider the optimal control problem

$$
\begin{aligned}
&\int_0^T x_1x_2\,dt \to \min;\\
&\dot{x}_1 = x_1 + u, \qquad \dot{x}_2 = -x_1 + x_2 + \left((x_1+1)^2+1\right)u,\\
&x_1 \in AC[0,T], \qquad x_2 \in AC[0,T], \qquad u \in L_\infty[0,T],
\end{aligned}
\tag{15}
$$

with fixed boundary conditions, $x_1(0) = \overline{x}_1,\ x_2(0) = \overline{x}_2,\ x_1(T) = \overline{\overline{x}}_1,\ x_2(T) = \overline{\overline{x}}_2$, and fixed $T > 0$. We can consider the augmented system,

$$\dot{x}_1 = x_1 + u, \qquad \dot{x}_2 = -x_1 + x_2 + \left((x_1+1)^2+1\right)u, \qquad \dot{x}_3 = x_1x_2.$$

This can be represented in the form $\dot{x} = Y(x) + X(x)u$, with $Y(x) = (x_1,\ -x_1+x_2,\ x_1x_2)$, $X(x) = \left(1,\ (x_1+1)^2+1,\ 0\right)$. Since the problem has one single input, assumption **A2** holds trivially. The first Lie brackets are

$$
\begin{aligned}
[Y,X] &= \left(-1,\ x_1^2-1, -x_1^3-2x_1^2-2x_1-x_2\right),\\
[Y[Y,X]] &= \left(1,\ x_1^2,\ -4x_1^2(x_1+1)\right),\\
[X[Y,X]] &= \left(0,\ 4x_1+2,\ -2\left(2x_1^2+3x_1+2\right)\right).
\end{aligned}
$$

The primary constraint, first-order secondary constraint and second-order secondary constraint reduce to $\xi_1 + \xi_2\left((x_1+1)^2+1\right) = 0$, $-\xi_1 + \xi_2\left(x_1^2-1\right) - \xi_3\left(x_1^3+2x_1^2+2x_1+x_2\right) = 0$ and $\xi_1 + \xi_2x_1^2 - 4\xi_3x_1^2(x_1+1) + \left(\xi_2(4x_1+2) - 2\xi_3\left(2x_1^2+3x_1+2\right)\right)u = 0$, respectively. This three constraints give u in the feedback form

$$u = -\frac{(x1+1)\left(4x_1^4+5x_1^3+4x_1^2+2x_1+x_2\right)}{2x_1^4+5x_1^3+6x_1^2+5x_1+2-(2x_1+1)x_2}. \tag{16}$$

Thus, any further constraint is redundant and we have

$$\mathcal{W} = \left\{(x_1, x_2, x_3, \xi_1, \xi_2, \xi_3) : \xi_1 = -\xi_2\left((x_1+1)^2+1\right)\right\},$$
$$\mathcal{D} = \left\{(x_1, x_2, x_3, \xi_1, \xi_2, \xi_3) : \xi_1 = -\xi_3 \frac{(x_1^2+2x_1+2)\left(x_1^3+2x_1^2+2x_1+x_2\right)}{2x_1^2+2x_1+1}, \right.$$
$$\left. \xi_2 = \xi_3 \frac{x_1^3+2x_1^2+2x_1+x_2}{2x_1^2+2x_1+1}\right\}.$$

Since the constraints are homogeneous with respect to (ξ_1, ξ_2, ξ_3), it follows that no abnormal extremal exists for this problem. Hence we can set $\xi_3 = -1$ and analyze the generalized Hamiltonian flow in the 4-dimensional space with coordinates (x_1, x_2, ξ_1, ξ_2). Since any generalized trajectory must lie in $\mathcal{W}$ and the coordinate ξ_1 for a point in $\mathcal{W}$ is uniquely defined by its coordinates (x_1, ξ_2), we can reduce further the dimension of the phase space and analyze the generalized Hamiltonian flow in the 3-dimensional space with coordinates (x_1, x_2, ξ_2). Thus, we can use the identifications $\mathcal{W} \sim \mathbb{R}^3$, $\mathcal{D} \sim \left\{(x_1, x_2, \xi_2) : \xi_2 = -\frac{x_1^3+2x_1^2+2x_1+x_2}{2x_1^2+2x_1+1}\right\}$. On the other hand,

$$\Phi_1^{\overrightarrow{h}\,X_\upsilon}(x,\xi) = \begin{pmatrix} x_1+\upsilon \\ x_2 + \left(x_1^2+2x_1+2\right)\upsilon + (x_1+1)\,\upsilon^2 + \frac{\upsilon^3}{3} \\ x_3 \\ \xi_1 - \xi_2\upsilon\,(\upsilon+2x_1+2) \\ \xi_2 \\ \xi_3 \end{pmatrix}.$$

Hence any jump changes only the coordinates (x_1, x_2, ξ_1), leaving the coordinates (x_3, ξ_2, ξ_3) unchanged. In particular, since the coordinate x_3 represents the running cost, in this example jumps have zero cost. This is a particular case: in general the cost of one particular jump can be any real number.

For $(x, \xi) \in \mathcal{W}$, $\upsilon \in \mathbb{R}$, $h_Y \circ \Phi_1^{\overrightarrow{h}\,X_\upsilon}(x,\xi)$ is a 4^{th} degree polynomial with leading term $-\frac{\upsilon^4}{3}$, hence it has a maximum. It follows that there exist extremals starting at every point $(x_1, x_2) \in \mathbb{R}^2$.

Different extremal trajectories with initial point $(x_1(0), x_2(0)) = (\overline{x}_1, \overline{x}_2)$ can be selected by choosing different initial adjoint states, $\overline{\xi}_2$. An initial jump will transfer the (fixed) initial point, $\left(\overline{x}_1, \overline{x}_2, \overline{\xi}_2\right)$, to a point $\left(x_1\left(0^+\right), x_2\left(0^+\right), \xi_2\left(0^+\right)\right)$ located in the intersection of the surface

$$\xi_2 = -\frac{x_1^3+2x_1^2+2x_1+x_2}{2x_1^2+2x_1+1}$$

with the curve

$$\mathcal{J}_{(\overline{x}_1,\overline{x}_2,\overline{\xi}_2)} = \left\{\left(\overline{x}_1+\upsilon,\ \overline{x}_2 + \left(\overline{x}_1^2+2\overline{x}_1+2\right)\upsilon + (\overline{x}_1+1)\,\upsilon^2 + \frac{\upsilon^3}{3},\ \overline{\xi}_2\right),\ \upsilon \in \mathbb{R}\right\}.$$

The intersection of these two submanifolds is given by the zeros of the polynomial

$$P(\upsilon) = \begin{array}{l} \frac{4}{3}\upsilon^3 + \left(2\overline{\xi}_2 + 4\overline{x}_1 + 3\right)\upsilon^2 + \left(4\overline{x}_1^2 + 6\overline{x}_1 + 4 + (4\overline{x}_1 + 2)\,\overline{\xi}_2\right)\upsilon + \\ +\overline{x}_2 + \overline{x}_1^3 + 2\overline{x}_1^2 + 2\overline{x}_1 + \left(2\overline{x}_1^2 + 2\overline{x}_1 + 1\right)\overline{\xi}_2. \end{array}$$

It follows that for each $\left(\overline{x}_1, \overline{x}_2, \overline{\xi}_2\right) \in \mathbb{R}^3$, the set $\mathcal{J}_{\left(\overline{x}_1, \overline{x}_2, \overline{\xi}_2\right)} \cap \mathcal{D}$ is nonempty and has at most three different points.

Consider $\left(\overline{x}_1, \overline{x}_2, \overline{\xi}_2\right) \in \mathcal{D}$. Then, the polynomial P reduces to

$$P(\upsilon) = \upsilon^3 + \frac{3}{4}\frac{6\overline{x}_1^3+10\overline{x}_1^2+6\overline{x}_1+3-2\overline{x}_2}{2\overline{x}_1^2+2\overline{x}_1+1}\upsilon^2 + \frac{3}{2}\frac{2\overline{x}_1^4+5\overline{x}_1^3+6\overline{x}_1^2+5\overline{x}_1+2-(2\overline{x}_1+1)\overline{x}_2}{2\overline{x}_1^2+2\overline{x}_1+1}\upsilon.$$

$\upsilon = 0$ is the unique zero of P if and only if $(\overline{x}_1, \overline{x}_2)$ lies in the region of the plane between the curves $x_2 = \gamma_1(x_1) = \frac{-14x_1^3-18x_1^2-14x_1+1+4\left(2x_1^2+2x_1+1\right)\sqrt{(2x_1+1)^2+6}}{6}$, $x_2 = \gamma_2(x_1) = \frac{-14x_1^3-18x_1^2-14x_1+1-4\left(2x_1^2+2x_1+1\right)\sqrt{(2x_1+1)^2+6}}{6}$. If $\overline{x}_2 \geq \gamma_1(\overline{x}_1)$ (resp., $\overline{x}_2 \leq \gamma_2(\overline{x}_1)$), then $\mathcal{J}_{\left(\overline{x}_1, \overline{x}_2, \overline{\xi}_2\right)} \cap \mathcal{D}$ contains either two or three points. If it contains only two points, $\left(\overline{x}_1, \overline{x}_2, \overline{\xi}_2\right)$ and $\left(\widetilde{x}_1, \widetilde{x}_2, \widetilde{\xi}_2\right)$, then either $\overline{x}_2 = \gamma_1(\overline{x}_1)$ (resp., $\overline{x}_2 = \gamma_2(\overline{x}_1)$) or $\widetilde{x}_2 = \gamma_1(\widetilde{x}_1)$ (resp., $\widetilde{x}_2 = \gamma_2(\widetilde{x}_1)$). In the case when $\mathcal{J}_{\left(\overline{x}_1, \overline{x}_2, \overline{\xi}_2\right)} \cap \mathcal{D}$ contains more than one point, then the jump must proceed to the point that maximizes x h_Y, i.e., we must choose υ that maximizes

$$h_Y \circ \Phi_1^{\overrightarrow{h}\,X_\upsilon}\left(\overline{x}, \overline{\xi}\right) = \frac{-1}{3}\upsilon^4 - \frac{6\overline{x}_1^3 + 10\overline{x}_1^2 + 6\overline{x}_1 + 3 - 2\overline{x}_2}{3\left(2\overline{x}_1^2 + 2\overline{x}_1 + 1\right)}\upsilon^3 -$$
$$\frac{2\overline{x}_1^4+5\overline{x}_1^3+6\overline{x}_1^2+5\overline{x}_1+2-(2\overline{x}_1+1)\overline{x}_2}{2\overline{x}_1^2+2\overline{x}_1+1}\upsilon^2 + \frac{\overline{x}_1^6+4\overline{x}_1^5+9\overline{x}_1^4-5\overline{x}_1^3+6\overline{x}_1^2+5\overline{x}_1+2-\overline{x}_2^2-2\left(\overline{x}_1^3+\overline{x}_1^2\right)\overline{x}_2}{2\overline{x}_1^2+2\overline{x}_1+1}.$$

One may check that $\upsilon = 0$ is the maximizer of $h_Y \circ \Phi_1^{\overrightarrow{h}\,X_\upsilon}\left(\overline{x}, \overline{\xi}\right)$ if and only if the polynomial

$$Q(\upsilon) = \frac{-1}{3}\upsilon^2 - \frac{6\overline{x}_1^3 + 10\overline{x}_1^2 + 6\overline{x}_1 + 3 - 2\overline{x}_2}{3\left(2\overline{x}_1^2 + 2\overline{x}_1 + 1\right)}\upsilon - \frac{2\overline{x}_1^4+5\overline{x}_1^3+6\overline{x}_1^2+5\overline{x}_1+2-(2\overline{x}_1+1)\overline{x}_2}{2\overline{x}_1^2+2\overline{x}_1+1}$$

has at most one zero. This happens if and only if $(\overline{x}_1, \overline{x}_2)$ lies on or between the curves

$$x_2 = \gamma_3(x_1) = \frac{-6x_1^3 - 8x_1^2 - 6x_1 + \sqrt{3}\left(2x_1^2 + 2x_1 + 1\right)\sqrt{(2x_1 + 1)^2 + 4}}{2}$$

and

$$x_2 = \gamma_4(x_1) = \frac{-6x_1^3 - 8x_1^2 - 6x_1 - \sqrt{3}\left(2x_1^2 + 2x_1 + 1\right)\sqrt{(2x_1 + 1)^2 + 4}}{2}.$$

Thus, the state space can be divided in 3 regions: $A_1 = \{(x_1, x_2) : x_2 > \gamma_3(x_1)\}$, $A_2 = \{(x_1, x_2) : x_2 < \gamma_4(x_1)\}$, $A_3 = \{(x_1, x_2) : \gamma_4(x_1) \leq x_2 \leq \gamma_3(x_1)\}$. If $(x_1(t), x_2(t))$

is a Sarychev extremal, then the set $\{t > 0 : (x_1(t), x_2(t)) \in A_1 \cup A_2\}$ must be a set of zero measure. If $(\overline{x}_1, \overline{x}_2) \in A_1 \cup A_2$, then every Sarychev extremal must start by a jump that transfers the state into A_3.

Now, let's consider the case when $\left(\overline{x}_1, \overline{x}_2, \overline{\xi}_2\right) \in \mathcal{D}$, $(\overline{x}_1, \overline{x}_2) \in int\,(A_3)$. For such initial points, a classical extremal exists and coincides with the Sarychev extremal, at least for small time intervals. Hence the extremal control can be determined through Dirac's constraints (16). By substituting this control in system (15), we obtain the differential equations

$$\begin{aligned} \dot{x}_1 &= -\frac{\left(2x_1^2+2x_1+1\right)\left(x_1^3+x_1^2+x_2\right)}{2x_1^4+5x_1^3+6x_1^2+5x_1+2-(2x_1+1)x_2}, \\ \dot{x}_2 &= -x_1 + x_2 - \frac{(x_1+1)\left(x_1^2+2x_1+2\right)\left(4x_1^4+5x_1^3+4x_1^2+2x_1+x_2\right)}{2x_1^4+5x_1^3+6x_1^2+5x_1+2-(2x_1+1)x_2}. \end{aligned} \tag{17}$$

This field has one unique critical point in A_3 which is $(x_1, x_2) = (0, 0)$. This is a saddle and its stable and unstable manifolds divide A_3 in four regions, with one branch of the unstable manifold passing through the narrow space between A_1 and A_2, and one branch of the stable manifold intersecting the boundary of A_2 (see figure). Along the curve $x_2 = \gamma_3(x_1)$, the field (17) points outward from A_1 at every point to the left of the point $x_1 = \frac{\sqrt{2}-1}{2}$ and points inward to A_1 at every point to the right of $x_1 = \frac{\sqrt{2}-1}{2}$. Hence, A_1 is attractive for every extremal trajectory lying in A_3 that passes sufficiently close to the curve $x_2 = \gamma_3(x_1)$, $x_1 > \frac{\sqrt{2}-1}{2}$ and is repulsive for every extremal trajectory lying in A_3 that passes sufficiently close to the curve $x_2 = \gamma_3(x_1)$, $x_1 < \frac{\sqrt{2}-1}{2}$. Similarly, A_2 is attractive for every extremal trajectory lying in A_3 that passes sufficiently close to the curve $x_2 = \gamma_4(x_1)$, $x_1 < \frac{-\sqrt{2}-1}{2}$ and is repulsive for every extremal trajectory lying in A_3 that passes sufficiently close to the curve $x_2 = \gamma_4(x_1)$, $x_1 > \frac{-\sqrt{2}-1}{2}$.

From the above arguments it follows that the structure of Sarychev extremals with initial point $(\overline{x}_1, \overline{x}_2) \in \mathbb{R}^2$ is as follows. At time $t = 0$, execute a jump to any point in $\Phi^X(\overline{x}_1, \overline{x}_2) \cap A_3$. This jump is selected by choosing the appropriated lift $(\overline{x}_1, \overline{x}_2) \mapsto \left(\overline{x}_1, \overline{x}_2, \overline{\xi}_2\right)$. If $(\overline{x}_1, \overline{x}_2) \in A_3$, a jump with zero length is possible, otherwise the jump must transfer the initial state to A_3 and hence it must have positive length. In the interval $]0, T[$, the generalized extremal must be of one of the following types:

1. The point $\left(x_1\left(0^+\right), x_2\left(0^+\right)\right)$ lies in the region to the left of the stable and the unstable manifolds of (17) and to the left of A_2, but not close to the curve $x_2 = \gamma_4(x_1)$, $x_1 < \frac{-\sqrt{2}-1}{2}$. In this case the generalized extremal is an upward moving continuous trajectory in the interval $]0, T[$, for every $T > 0$.

2. The point $\left(x_1\left(0^+\right), x_2\left(0^+\right)\right)$ lies in the region to the left of the stable and the unstable manifolds of (17), to the left of A_2 and sufficiently close to the curve $x_2 = \gamma_4(x_1)$, $x_1 < \frac{-\sqrt{2}-1}{2}$. In this case, for sufficiently large $T > 0$, the

generalized extremal has one unique discontinuity in the interval $]0,T[$. This discontinuity transfers the state $\left(x_1\left(t_1^-\right),x_2\left(t_1^-\right)\right)$ which lies in the curve $x_2=\gamma_4\left(x_1\right),\ x_1<\frac{-\sqrt{2}-1}{2}$ to a point $\left(x_1\left(t_1^+\right),x_2\left(t_1^+\right)\right)$, lying in the curve $x_2=\gamma_4\left(x_1\right),\ x_1>\frac{-\sqrt{2}-1}{2}$. Two subcases can occur:

(a) If the point $\left(x_1\left(t_1^+\right),x_2\left(t_1^+\right)\right)$ lies to the left of the stable manifold of (17), then the generalized extremal in the interval $]t_1,T[$ is a trajectory moving upwards to the left;

(b) If the point $\left(x_1\left(t_1^+\right),x_2\left(t_1^+\right)\right)$ lies to the right of the stable manifold of (17), then the generalized extremal in the interval $]t_1,T[$ is a trajectory moving downwards to the right.

3. The point $\left(x_1\left(0^+\right),x_2\left(0^+\right)\right)$ lies in the region above the stable and the unstable manifolds of (17) and below A_1, but not close to the curve $x_2=\gamma_3\left(x_1\right),\ x_1>\frac{\sqrt{2}-1}{2}$. In this case the generalized extremal is a continuous trajectory in the interval $]0,T[$, for every $T>0$. It evolves in the region between the stable and unstable manifolds of (17) and A_1, from right to left.

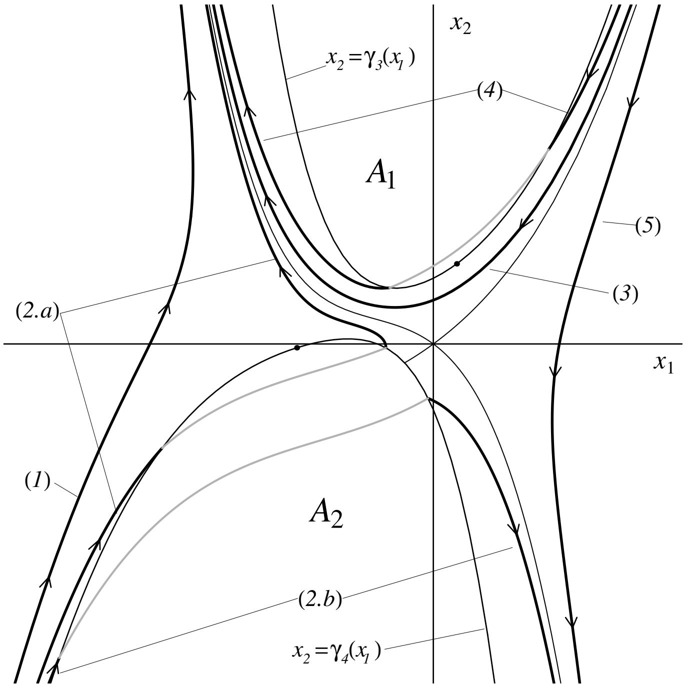
4. The point $\left(x_1\left(0^+\right),x_2\left(0^+\right)\right)$ lies in the region above the stable and the unstable manifolds of (17), below A_1 and close to the curve $x_2=\gamma_3\left(x_1\right),\ x_1>\frac{\sqrt{2}-1}{2}$. In this case, for sufficiently large $T>0$, the generalized extremal has one unique discontinuity in the interval $]0,T[$. This discontinuity transfers the state $\left(x_1\left(t_1^-\right),x_2\left(t_1^-\right)\right)$ which lies in the curve $x_2=\gamma_3\left(x_1\right),\ x_1>\frac{\sqrt{2}-1}{2}$ to a point $\left(x_1\left(t_1^+\right),x_2\left(t_1^+\right)\right)$, lying in the curve $x_2=\gamma_3\left(x_1\right),\ x_1<\frac{\sqrt{2}-1}{2}$. In the intervals $]0,t_1[$, $]t_1,T[$ the generalized extremal is a continuous trajectory lying in the region between the stable and unstable manifolds of (17) and A_1, moving from right to left.

5. The point $\left(x_1\left(0^+\right),x_2\left(0^+\right)\right)$ lies in the region to the right of the stable and the unstable manifolds of (17). In this case the generalized extremal is a downward moving continuous trajectory in the interval $]0,T[$, for every $T>0$.

6. If the point $\left(x_1\left(0^+\right),x_2\left(0^+\right)\right)$ lies in the region below the stable and the unstable manifolds of (17) and to the right of A_2, Then the generalized extremal is a continuous curve in the interval $]0,T[$, for every $T>0$, of the same type as the arc described by an extremal of the type $(2.b)$ in the interval $]t_1,T[$.

At time $t=T$, we may choose a final jump that transfers the state $\left(x_1\left(T^-\right),x_2\left(T^-\right)\right)$ to any point in $\Phi^X\left(x_1\left(T^-\right),x_2\left(T^-\right)\right)$ (including points in $A_1\cup A_2$).

Arcs of extremals of types (1) to (5) are shown in the figure, identified by the corresponding numbers. The continuous arcs of the generalized extremals are drawn in black, while the segments of the integral manifolds Φ^X that describe jumps during the interval $]0,T[$ are drawn in grey (initial and final jumps are not represented).



It should be noticed that the denominator in feedback (16) has zeroes in $\mathbb{R}^2$. However, the only zeroes that lie in A_3 are the points

$$\left(\frac{-\sqrt{2}-1}{2}, \frac{9-7\sqrt{2}}{8}\right) \text{ and } \left(\frac{\sqrt{2}-1}{2}, \frac{9+7\sqrt{2}}{8}\right),$$

which are, respectively, the points where the curves $x_2 = \gamma_4(x_1)$, and $x_2 = \gamma_3(x_1)$ are neither repulsive nor attractive for extremal trajectories lying in A_3 (black dots in the figure). The direction of the field is well defined in both these points and it is tangent to the boundary of A_3, pointing from the attractive segment towards the repulsive segment. The extremals that pass through these points attain infinite speed when they pass over the point but lie entirely on A_3 and are continuous.

Also, notice that every point $(\overline{x}_1, \overline{x}_2) \in A_1 \cup A_2$ which is not a zero of the denominator of (16), admits a classical extremal with initial condition $(x_1(0), x_2(0)) = (\overline{x}_1, \overline{x}_2)$, but none of these classical extremals is a Sarychev extremal. However, every classical extremal arc that lies in one of the regions, between the curves

$$x_2 = \frac{(x_1+1)^2\left(2x_1^2+x_1+2\right)}{2x_1+1},\ x_1 > -\frac{1}{2} \text{ and } x_2 = \gamma_3(x_1),$$

or between the curves

$$x_2 = \frac{(x_1+1)^2\left(2x_1^2+x_1+2\right)}{2x_1+1},\ x_1 < -\frac{1}{2} \text{ and } x_2 = \gamma_4\left(x_1\right),$$

satisfies the strong Legendre-Clebsch condition at every point.

References

[1] AGRACHEV A.A. AND SARYCHEV A.V., *On reduction of a smooth system linear in the control*, Math USSR Sbornik **58** (1) (1987), 15–30.

[2] DIRAC P.A.M., *Lectures on quantum mechanics*, Belfer Graduate School of Science, 1964.

[3] GUERRA M. *Distribution-like Hamiltonian flows and generalized optimal controls*, J. Math Sci. **120** (1) (2004), 895-918.

[4] SARYCHEV A.V. *Nonlinear Systems with Impulsive and Generalized Function Controls*, in: "Proceedings of a IIASA Workshop, Sopron 1989" (Eds. Byrnes C.I. and Kurzhansky A.) Birkhäuser, Boston 1991, 244–257.

AMS Subject Classification: 93C15, 93B50, 49K15.

Manuel GUERRA, Department of Mathematics, ISEG, Technical University of Lisbon, R. do Quelhas 6, 1200 Lisboa, PORTUGAL
e-mail: `mguerra@iseg.utl.pt`

Rend. Sem. Mat. Univ. Pol. Torino - Vol. 63, 4 (2005)
Control Theory and Stabil., I

N. Martins – V. Neves*

NONSTANDARD DISCRETE DERIVATIVES AND EXISTENCE THEOREMS FOR ODE

Abstract. We present nonstandard generalizations of Peano's and Carathéodory's Existence Theorems, which avoid Ascoli's Theorem as well as Lebesgue's Dominated Convergence Theorem.

1. Introduction

Nonstandard Analysis is a mathematical theory discovered by Abraham Robinson in the early 1960's ([9]) which among other things provides a logical foundation for the concept of infinitesimal number.

We begin with a brief and informal presentation of the main tools necessary for the understanding of this paper. Background on foundations of Nonstandard Analysis may be found in either [2], [4] or [11] and nonstandard integration theory is treated in [5], [6], [10] or [1].

${}^\star\mathbb{R}$ is a proper ordered field extension of $\mathbb{R}$, the set of real numbers. The elements of ${}^\star\mathbb{R}$ are said **hyperreals** numbers. A hyperreal number x is

1. **infinitesimal** if $|x| < \frac{1}{n}$ for all $n \in \mathbb{N}$;
2. **finite** if $|x| < n$ for some $n \in \mathbb{N}$;
3. **infinite** if it is not finite.

We denote by ${}^\star\mathbb{R}_b$ the set of all finite hyperreal numbers and by ${}^\star\mathbb{R}_\infty$ the set of all infinite hyperreal numbers. For $x, y \in {}^\star\mathbb{R}$, $x \approx y$ means that $x - y$ is infinitesimal or, in other words, x is **infinitely close** to y.

THEOREM 1. **(Standard Part Theorem)** *If $x \in {}^\star\mathbb{R}_b$, there exists one and only one real number $r \in \mathbb{R}$, called the **standard part** of x and denoted by* $\mathbf{st(x)}$ *or* $^\circ\mathbf{x}$*, such that $x \approx r$. Moreover, $st(x + y) = st(x) + st(y)$ and $st(xy) = st(x)st(y)$ whenever $x, y \in {}^\star\mathbb{R}_b$.*

The **nonstandard universe** consists of a pair of structures, $V(\mathbb{R})$ and $V({}^\star\mathbb{R})$, and a mapping

$$^\star(.) : V(\mathbb{R}) \to V({}^\star\mathbb{R})$$

*Work for this article was partially supported by the R&D unit Center for Research in Optimization and Control (CEOC) of the University of Aveiro and grant POCTI/MAT/41683/2001 of the Portuguese Foundation for Science and Technology (FCT) via FEDER.

which associates to each element $a \in V(\mathbb{R})$ its **nonstandard extension** ${}^\star a \in V({}^\star\mathbb{R})$.

All elements of $V(\mathbb{R})$ and their nonstandard extensions are called **standard**, that is, for all $a \in V(\mathbb{R})$, a and ${}^\star a$ are standard. Elements of standard sets will be called **internal**; in particular, ${}^\star a$ is also internal. All elements of $V({}^\star\mathbb{R})$ which are not internal, are called **external**.

The nonstandard extension of $\mathbb{N}$, ${}^\star\mathbb{N}$, is the set of **hypernatural** numbers and we denote by ${}^\star\mathbb{N}_\infty$ the set of infinite hypernatural numbers.

A family $\mathcal{C}$ of sets satisfies the **finite intersection property (f.i.p.)** if intersections of finite subfamilies of $\mathcal{C}$ are non empty. In the following, if E is a set, $\mathcal{P}(E)$ denotes the set of subsets of E, $\mathcal{P}_{fin}(E)$ the set of all finite subsets of E and $card(E)$ the cardinality of E. If $E \in V(\mathbb{R})$, the elements of ${}^\star\mathcal{P}_{fin}(E)$ are called **hyperfinite** subsets of ${}^\star E$.

A **bounded formula** is a first order formula that can be written in such a way that all quantifiers range over a fixed set. A **sentence** is a formula without free variables.

The mapping ${}^\star(.)$ satisfies the following principles.

THEOREM 2. **(Polysaturation Principle)** *Given a set $E \in V(\mathbb{R})$ and $\mathcal{C} \subseteq {}^\star\mathcal{P}(E)$, if $\mathcal{C}$ verifies the f.i.p. and $card(\mathcal{C}) < card(V(\mathbb{R}))$, then $\mathcal{C}$ has non empty intersection.*

THEOREM 3. **(Transfer Principle)** *Suppose $\varphi(a_1, \cdots, a_n)$ is a bounded sentence whose only constants are the a_i. Then $\varphi(a_1, \cdots, a_n)$ is true in $V(\mathbb{R})$ if and only if $\varphi({}^\star a_1, \cdots, {}^\star a_n)$ is true in $V({}^\star\mathbb{R})$.*

The Transfer Principle shows that hyperfinite sets have the same formal properties of finite sets; any hyperfinite set of hyperreals has a minimum and a maximum. The Transfer Principle also shows that a set $B \in V({}^\star\mathbb{R})$ is hyperfinite iff there exists $N \in {}^\star\mathbb{N}$ and an internal bijective map $f : B \to \{n \in {}^\star\mathbb{N} \mid n \leq N\}$. Another consequence of the Transfer Principle is that the map ${}^\star(.)$ respects boolean operations. Polysaturation Principle creates new nonstandard elements.

We say that a bounded formula φ is standard (resp. internal) if all of its constants denote standard (resp. internal) elements of $V({}^\star\mathbb{R})$.

THEOREM 4. *A set $b \in V({}^\star\mathbb{R})$ is standard (resp. internal) iff there exits a set $a \in V(\mathbb{R})$ and a bounded standard (resp. internal) formula φ such that $b = \{x \in {}^\star a | \varphi(x)\}$.*

THEOREM 5. *For any set $A \in V(\mathbb{R})$ there exists a hyperfinite set H such that*

$$A \subseteq H \subseteq {}^\star A.$$

A is infinite iff both inclusions are strict. In particular, the set A is infinite iff ${}^\star A$ contains nonstandard elements.

DEFINITION 1. *Let $Y \in {}^\star\mathcal{P}(\mathbb{R})$ and $F : Y \to {}^\star\mathbb{R}$ be an internal function. Then*

*F is said to be **S-continuous** if for all $x, y \in Y$ we have*

$$x \approx y \Rightarrow F(x) \approx F(y).$$

Some relations between this notion and the usual continuity are the following results.

THEOREM 6. *A function $f : \mathbb{R} \to \mathbb{R}$ is*

1. *continuous on $c \in \mathbb{R}$ iff for all $x \in {}^\star\mathbb{R}$ such that $x \approx c$ then ${}^\star f(x) \approx f(c)$;*
2. *continuous iff ${}^\star f$ is S-continuous in ${}^\star\mathbb{R}_b$;*
3. *uniformily continuous iff ${}^\star f$ is S-continuous in ${}^\star\mathbb{R}$.*

THEOREM 7. *If $[a, b] \subseteq \mathbb{R}$, $F : {}^\star[a, b] \to {}^\star\mathbb{R}$ is internal and S-continuous and there exists $z \in {}^\star[a, b]$ such that $F(z)$ is finite, then*

1. *$F(x)$ is finite for all $x \in {}^\star[a, b]$;*
2. *the standard function $f : [a, b] \to \mathbb{R}$, defined by $f(t) = {}^\circ F(t)$ is continuous and for all $x \in {}^\star[a, b]$, ${}^\star f(x) \approx F(x)$.*

DEFINITION 2. *Let $Y \in {}^\star\mathcal{P}(\mathbb{R})$ and $F : Y \to {}^\star\mathbb{R}$ an internal function. Then F is **S-absolutely continuous** if*

$$\sum_{i=1}^{N} |F(b_i) - F(a_i)| \approx 0$$

for every hyperfinite collection

$$\{[a_1, b_1[, [a_2, b_2[, \cdots, [a_N, b_N[\}$$

(where $[a, b[$ denotes the set $\{t \in {}^\star\mathbb{R} : a \leq t < b\} \cap Y$) of non overlapping subintervals of Y such that $\sum_{i=1}^{N}(b_i - a_i) \approx 0$.

THEOREM 8. *If $[a, b] \subseteq \mathbb{R}$, $F : {}^\star[a, b] \to {}^\star\mathbb{R}_b$ is internal and S-absolutely continuous function, then there exists a standard absolutely continuous function $f : [a, b] \to \mathbb{R}$ such that $f(t) = {}^\circ F(t)$.*

REMARK 1.

1. It is clear that each S-absolutely continuous function is S-continuous.
2. Theorems 7 and 8 remain true if we substitute ${}^\star[a, b]$ by a hyperfinite set $\mathbb{X}$ such that $st(\mathbb{X}) = [a, b]$ with $a, b \in \mathbb{R}$.
3. Often we avoid * on nonstandard extensions of functions.

2. Loeb integration theory

Loeb measures were discovered by Peter Loeb in 1975 ([8]). These measures are obtained from an internal measure in the following way.

Suppose that $(\Omega, \mathcal{A}, \mu)$ is an **internal measure space**, that is, Ω is an internal non empty set, $\mathcal{A}$ an internal algebra on Ω and $\mu : \mathcal{A} \to {}^\star\mathbb{R}$ an internal finitely additive measure. In general, this is not a measure space because $\mathcal{A}$ is not a σ-algebra except in the trivial case where $\mathcal{A}$ is finite. The Loeb measure generated by μ will be denoted by μ_L and is a measure defined in a family of subsets of Ω that contains the internal algebra $\mathcal{A}$ and that coincide with ${}^\circ\mu = st(\mu)$ on $\mathcal{A}$.

DEFINITION 3. *Let $B \subseteq \Omega$ (B not necessarily internal). We say that*

1. *B is a **Loeb null set** if for each real $\epsilon > 0$ there exists an internal set $A \in \mathcal{A}$ such that $B \subseteq A$ and $\mu(A) < \epsilon$;*
2. *B is **Loeb measurable** if there exists a set $A \in \mathcal{A}$ such that $A \triangle B := (A \setminus B) \cup (B \setminus A)$ is Loeb null. Denote the collection of all Loeb measurable sets by $L(\mathcal{A})$;*
3. *For $B \in L(\mathcal{A})$ define*
$$\mu_L(B) = {}^\circ\mu(A)$$
*for all $A \in \mathcal{A}$ such that $A \triangle B$ is Loeb null (where ${}^\circ x = st(x) = +\infty$ if $0 < x \in {}^\star\mathbb{R}_\infty$); $\mu_L(B)$ is called the **Loeb measure** of B.*

Note that $\mu_L : L(\mathcal{A}) \to [0, +\infty]$ and

$$\forall A \in \mathcal{A} \ \ \mu_L(A) = {}^\circ\mu(A).$$

THEOREM 9. *$L(\mathcal{A})$ is a σ-algebra, called **Loeb σ-algebra**, and μ_L is a complete σ-additive measure on $L(\mathcal{A})$.*

$(\Omega, L(\mathcal{A}), \mu_L)$ is a measure space, called the **Loeb space**, generated by $(\Omega, \mathcal{A}, \mu)$. Note that μ_L acts on sets which may not be standard.

An important example of a Loeb space is the Loeb counting measure space. Fix $N \in {}^\star\mathbb{N}_\infty$, define $\triangle = \frac{1}{N}$ and make

$$\mathbb{T} = \{k\triangle : k = 0, 1, 2, \cdots, N-1\} = \{0, \frac{1}{N}, \frac{2}{N}, \cdots, 1 - \frac{1}{N}\}. \tag{1}$$

$\mathbb{T}$ is usually called **hyperfinite time line** with increment $\triangle$. Denoting the set of all internal subsets of $\mathbb{T}$ by $\mathcal{A}$ and defining $\nu : \mathcal{A} \to {}^\star[0, 1]$ by

$$\nu(A) = \frac{card(A)}{card(\mathbb{T})} = \frac{card(A)}{N}$$

we obtain an internal measure space $(\mathbb{T}, \mathcal{A}, \nu)$ called **internal counting measure space**. The Loeb space $(\mathbb{T}, L(\mathcal{A}), \nu_L)$ generated by $(\mathbb{T}, \mathcal{A}, \nu)$ is called the **Loeb counting measure space**. This hyperfinite space can be used to represent the Lebesgue space $([0, 1], \mathcal{L}, \lambda)$:

THEOREM 10. *Let* $(\mathbb{T}, L(\mathcal{A}), \nu_L)$ *be the Loeb counting measure space. A set* $A \subseteq [0,1]$ *is Lebesgue measurable iff*

$$st_{\mathbb{T}}^{-1}(A) = \{t \in \mathbb{T} : \ {}^{\circ}t \in A\}$$

is Loeb measurable and

$$\lambda(A) = \nu_L(st_{\mathbb{T}}^{-1}(A)).$$

We deal now with measurable functions.

DEFINITION 4. *A function* $f : \Omega \to \mathbb{R}$ *is* ***Loeb measurable*** *if* f *is* μ_L*-measurable in the conventional sense, that is, for every open set* $B \subseteq \mathbb{R}$, $f^{-1}(B) \in L(\mathcal{A})$.

DEFINITION 5. *An internal function* $F : \Omega \to {}^{\star}\mathbb{R}$ *is* ***$\star$-measurable*** *if* $F^{-1}(A) \in \mathcal{A}$*, for any* **open set* $A \subseteq {}^{\star}\mathbb{R}$.

Some connections between these notions are given in the following theorem.

THEOREM 11. *If* $F : \Omega \to {}^{\star}\mathbb{R}$ *is internal and* $\star$*-measurable, then* ${}^{\circ}F$ *is Loeb measurable.*

DEFINITION 6. *Let* $(X, \mathcal{L}, \lambda)$ *be a standard measure space. An internal* $\star$*-measurable function* $F : {}^{\star}X \to {}^{\star}\mathbb{R}$ *is a (two legged)* ***lifting*** *of* $f : X \to \mathbb{R}$ *if*

$${}^{\circ}F(x) = f({}^{\circ}x) \qquad {}^{\star}\lambda_L\text{- a.a. } x \in {}^{\star}X$$

(a.a. means almost all).

THEOREM 12. **(Anderson's Theorem)** *Let* $(X, \mathcal{L}, \lambda)$ *be a Lebesgue measure space,* (Y, Γ) *a Hausdorff space with a countable base of open sets and* $f : X \to Y$ *a Lebesgue measurable function. Then* ${}^{\star}f$ *is a lifting of* f.

This may be considered the main lemma for Carathéodory's Existence Theorem 19, the basic ideas of its proof being that nearness in ${}^{\star}Y$ is "measured countably" — $u \approx y \in Y$ if and only if $u \in {}^{\star}B_n$, for whatever basic open set B_n such that $y \in B_n$ $(n \in \mathbb{N})$ —, countable unions of null sets are also null as well as sets of the form ${}^{\star}C \Delta st^{-1}(C)$ ([5, page 158]).

REMARK 2. Anderson proves this result in the case where $(X, \mathcal{L}, \lambda)$ is a complete Radon space. A proof of a version of Anderson's Theorem is given in [5, page 167].

Loeb measures are classical measures over σ-algebras (with possibly nonstandard elements), thus Loeb integration theory is simply the classical theory of integration with respect to Loeb measure: in particular, a Loeb measurable function $f : \Omega \to$

$\mathbb{R}$ is Loeb integrable if f is integrable in the classical sense with respect to the Loeb measure μ_L, in which case the Loeb integral $\int_\Omega f d\mu_L$ is a real number.

The $\star$**-integral** or **internal integral** of a $\star$-measurable function $F : \Omega \to {}^\star\mathbb{R}$ is obtained by Transfer of the definition of the standard integral.

Although Theorem 11 says that if $F : \Omega \to {}^\star\mathbb{R}$ is internal and $\star$-measurable then ${}^\circ F$ is Loeb measurable and for all $x \in \Omega$

$$F(x) \approx {}^\circ F(x),$$

the equation

$$(2) \qquad {}^\circ\left(\int_\Omega F d\mu\right) = \int_\Omega {}^\circ F d\mu_L$$

is, in general, false:

EXAMPLE 1. Let $(\mathbb{T}, L(\mathcal{A}), \nu_L)$ be the Loeb counting measure space. Define the internal $\star$-measurable function

$$F(\tau) = \begin{cases} N^2 & \text{if} \quad \tau = 0 \\ 0 & \text{if} \quad \tau \in \mathbb{T} \setminus \{0\} \end{cases}$$

where $N \in {}^\star\mathbb{N}_\infty$ is the same used in the construction of $\mathbb{T}$. Then $\int_\mathbb{T} {}^\circ F d\nu_L = 0$ (since ${}^\circ F(\tau) = 0$ for ν_L- almost all $\tau \in \mathbb{T}$) and

$$\int_\mathbb{T} F d\nu = \sum_{\tau \in \mathbb{T}} F(\tau) \frac{1}{N} = N.$$

To obtain equality of ${}^\circ\left(\int_\Omega F d\mu\right)$ and $\int_\Omega {}^\circ F d\mu_L$ we must restrict the class of $\star$-integrable functions.

DEFINITION 7. *An internal $\star$-measurable function $F : \Omega \to {}^\star\mathbb{R}$ is* ***S-integrable*** *if*

1. *$\int_\Omega \mid F \mid d\mu$ is finite;*
2. *for all $A \in \mathcal{A}$ such that $\mu(A) \approx 0$, then $\int_A \mid F \mid d\mu \approx 0$;*
3. *if $A \in \mathcal{A}$ and $F \approx 0$ on A, then $\int_A \mid F \mid d\mu \approx 0$.*

Condition 1 is necessary to guarantee that all S-integrable function are $\star$-integrable. Condition 2 is needed for equality (2), because $\int_A {}^\circ \mid F \mid d\mu_L = 0$, so $\int_A \mid F \mid d\mu$ must be infinitesimal. The last condition is also necessary to obtain equality (2) because in this case, $\int_A {}^\circ \mid F \mid d\mu_L = 0$.

Note that if the internal measure μ is finite, the last condition is always satisfied, since $F \approx 0$ on A means that for every $\epsilon \in \mathbb{R}^+$, $\int_A \mid F \mid d\mu \leq \epsilon\mu(A)$.

For $f : [0,1] \to \mathbb{R}$ we define $\widehat{f} : \mathbb{T} \to \mathbb{R}$ by

$$\widehat{f}(\tau) = f({}^{\circ}\tau).$$

THEOREM 13. *Let $(\mathbb{T}, L(\mathcal{A}), \nu_L)$ be the Loeb counting measure space and $([0,1], \mathcal{L}, \lambda)$ the Lebesgue measure space on $[0,1]$. The following conditions are equivalent:*

1. *$f : [0,1] \to \mathbb{R}$ is Lebesgue integrable;*
2. *$\widehat{f} : \mathbb{T} \to \mathbb{R}$ is Loeb integrable;*
3. *there exists an internal S-integrable function $F : \mathbb{T} \to {}^{\star}\mathbb{R}$ that is a lifting of f.*

In this case

$$\int_{[0,1]} f(t)d\lambda(t) = \int_{\mathbb{T}} \widehat{f}(\tau)d\nu_L(\tau) = {}^{\circ}\left(\int_{\mathbb{T}} F d\nu\right) = {}^{\circ}\left(\sum_{\tau\in\mathbb{T}} F(\tau)\frac{1}{N}\right)$$

REMARK 3. Note that the last theorem defines the Lebesgue integral on $[0,1]$ as the standard part of some hyperfinite sum. This is also true for the Lebesgue integral on $\mathbb{R}$ (see [10] for details).

The next theorem characterizes nonstandard extensions of Lebesgue integrable functions.

THEOREM 14. *Let $(Z, \mathcal{L}, \lambda)$ be a Lebesgue measure space and suppose that $f : Z \to \mathbb{R}$ is Lebesgue integrable. Then ${}^{\star}f : {}^{\star}Z \to {}^{\star}\mathbb{R}$ is S-integrable.*

3. Nonstandard discrete derivative

Let $\mathbb{T}$ be the hyperfinite time line with respect to the increment $\triangle = \frac{1}{N}$ and $N \in \mathbb{N}_\infty$ (see (1)). The **nonstandard discrete derivative** ([12]) of an internal function $X : \mathbb{T} \to {}^{\star}\mathbb{R}$ is the function $X' : \mathbb{T} \setminus \{1-\triangle\} \to {}^{\star}\mathbb{R}$ defined by

$$X'(t) := \frac{X(t+\triangle) - X(t)}{\triangle}.$$

THEOREM 15. *Let $(\mathbb{T}, \mathcal{A}, \nu)$ be the internal counting measure space and suppose $X : \mathbb{T} \to {}^{\star}\mathbb{R}_b$ is an internal function. The following conditions are equivalent:*

1. *X is S-absolutely continuous;*
2. *X' is S-integrable;*
3. *$\int_A \mid X' \mid d\nu = \sum_{\tau\in A} \mid X'(\tau) \mid \triangle \approx 0$ for all $A \in \mathcal{A}$ such that $\nu(A) \approx 0$.*

4. Nonstandard Peano's Existence Theorem

THEOREM 16. **(Nonstandard Peano's Existence Theorem)** * *Suppose* $F : \mathbb{T} \times {}^\star\mathbb{R} \to {}^\star\mathbb{R}_b$ *is internal and* $\alpha \in {}^\star\mathbb{R}$. *Then there exists one and only one internal S-absolutely continuous function* $X : \mathbb{T} \to {}^\star\mathbb{R}$ *such that*

$$\begin{cases} X(0) &= \alpha \\ X'(t) &= F(t, X(t)) \quad (t \in \mathbb{T} \setminus \{1 - \triangle\}) \end{cases} \tag{3}$$

Moreover, if α *is finite, then* $X(\mathbb{T}) \subseteq {}^\star\mathbb{R}_b$.

Proof. Define $X : \mathbb{T} \to {}^\star\mathbb{R}$ recursively by

$$\begin{aligned} X(0) &= \alpha \\ X(t + \triangle) &:= X(t) + F(t, X(t))\triangle \qquad (t \in \mathbb{T} \setminus \{1 - \triangle\}). \end{aligned}$$

X is internal, by construction

$$X(t) = \alpha + \sum_{i=0}^{k-1} F(i\triangle, X(i\triangle))\triangle \quad (t = k\triangle \in \mathbb{T}) \tag{4}$$

and

$$X'(t) = \frac{X(t + \triangle) - X(t)}{\triangle} = F(t, X(t)) \quad (t \in \mathbb{T} \setminus \{1 - \triangle\}). \tag{5}$$

X is actually S-Lipschitz, that is,

$$|X(t) - X(s)| \leq M|t - s| \qquad (s, t \in \mathbb{T}) \tag{6}$$

where $M \in \mathbb{R}$ is such that $|F(t, z)| \leq M$ for all $(t, z) \in \mathbb{T} \times {}^\star\mathbb{R}$:

Suppose that $s = k_1\triangle < k_2\triangle = t$, for certain $k_1, k_2 \in \{0, 1, \cdots, N - 1\}$. Then

$$\begin{aligned} |X(t) - X(s)| &= \Big|\sum_{i=k_1}^{k_2-1} F(i\triangle, X(i\triangle))\triangle\Big| \\ &\leq \sum_{i=k_1}^{k_2-1} |F(i\triangle, X(i\triangle))|\triangle \\ &\leq \sum_{i=k_1}^{k_2-1} M\triangle \\ &= M(t - s). \end{aligned}$$

*The reader might wish to consult chapter 8 of [7], where this subject is also treated from another viewpoint.

Then X satisfies (6) and is S-absolutely continuous.

Using the definition of the discrete derivative, it is clear that there exists only one internal function $X : \mathbb{T} \to {}^\star\mathbb{R}$ such that (3) holds.

Using (4) we can prove that, for each $k = 0, 1, \cdots, N-1$

$$|X(k\triangle) - \alpha| \leq \sum_{i=0}^{k-1} |F(i\triangle, X(i\triangle))|\triangle \leq \sum_{i=0}^{k-1} M\triangle \leq M$$

hence, $X(\mathbb{T}) \subseteq [\alpha - M, \alpha + M]$. If α is finite, we conclude that $X(\mathbb{T}) \subseteq {}^\star\mathbb{R}_b$. □

5. Peano's Existence Theorem

Using Nonstandard Peano's Existence Theorem we can prove

THEOREM 17. **(Peano's Existence Theorem)** *Suppose* $f : [0, 1] \times \mathbb{R} \to \mathbb{R}$ *is bounded and continuous and* $x_0 \in \mathbb{R}$. *Then there exists* $x : [0, 1] \to \mathbb{R}$ *such that*

$$\begin{cases} x(0) &= x_0 \\ x'(t) &= f(t, x(t)) \end{cases}$$

Proof. Suppose $F = {}^\star f_{|\mathbb{T}\times{}^\star\mathbb{R}} : \mathbb{T} \times {}^\star\mathbb{R} \to {}^\star\mathbb{R}$. F is internal, $F(\mathbb{T} \times {}^\star\mathbb{R}) \subseteq {}^\star\mathbb{R}_b$ and for each $\tau \in \mathbb{T}$ and $y \in {}^\star\mathbb{R}_b$,

$${}^\circ F(\tau, y) = f({}^\circ\tau, {}^\circ y);$$

note that

$$F(\tau, y) = {}^\star f(\tau, y) \approx f({}^\circ\tau, {}^\circ y)$$

because $\tau \approx {}^\circ\tau \in [0, 1]$, $y \approx {}^\circ y \in \mathbb{R}$ and f is continuous.

By Nonstandard Peano's Existence Theorem, there exists an internal S-absolutely continuous function

$$X : \mathbb{T} \to {}^\star\mathbb{R}_b$$

such that

$$\begin{cases} X(0) &= x_0 \\ X'(\tau) &= F(\tau, X(\tau)) \quad (\tau \in \mathbb{T} \setminus \{1 - \triangle\}) \end{cases}$$

Theorem 8 says there exists a standard absolutely continuous function $x : [0, 1] \to \mathbb{R}$ such that

$$x({}^\circ\tau) = {}^\circ X(\tau)$$

for all $\tau \in \mathbb{T}$. Hence

$$x(0) = {}^\circ X(0) = x_0$$

so that x satisfies the initial condition.

Using the definition and continuity of x we have that

$$X(\tau) \approx x({}^\circ\tau) \approx x(\tau) \quad (\tau \in \mathbb{T})$$

and since f is continuous,

$${}^\star f(\tau, X(\tau)) \approx {}^\star f(\tau, x(\tau)) \approx f({}^\circ\tau, x({}^\circ\tau)) \quad (\tau \in \mathbb{T}).$$

Hence $G : \mathbb{T} \to {}^\star\mathbb{R}$ such that $G(\tau) = {}^\star f(\tau, X(\tau))$ is a lifting of the Lebesgue integrable function $g : [0,1] \to \mathbb{R}$, $g(t) = f(t, x(t))$.

Moreover, G is S-integrable since for all $A \in \mathcal{A}$

$$\int_A |\, G \,|\, d\nu \leq \int_A M d\nu = M\nu(A)$$

where $M \in \mathbb{R}$ is an upper bound of f, and then

$$\int_{\mathbb{T}} |\, G \,|\, d\nu \leq M$$

and

$$\int_A |\, G \,|\, d\nu \approx 0$$

whenever $\nu(A) \approx 0$.

Next, we will prove that x is a solution to the initial value problem.

Fix $z \in [0,1]$ and $\tau = k\triangle \in \mathbb{T}$ such that $\tau \approx z$. Observe that

$$\begin{array}{rcll} x(z) & = & {}^\circ X(\tau) & \\ & = & x_0 + {}^\circ\left(\sum_{i=0}^{k-1} F(i\triangle, X(i\triangle))\triangle\right) & \\ & = & x_0 + {}^\circ\left(\sum_{i=0}^{k-1} G(i\triangle)\triangle\right) & \\ & = & x_0 + \int_{[0,z]} f(t, x(t)) d\lambda(t) & \text{(Theorem 13)} \end{array} \tag{7}$$

$\square$

6. Nonstandard Carathéodory's Existence Theorem

THEOREM 18. **(Nonstandard Carathéodory's Existence Theorem)** *Let $F : \mathbb{T} \times {}^\star\mathbb{R} \to {}^\star\mathbb{R}$ be an internal $\star$-measurable function. Suppose there exists an internal S-integrable function $M : \mathbb{T} \to {}^\star\mathbb{R}$ such that*

$$\forall (\tau, x) \in \mathbb{T} \times^\star \mathbb{R} \;\; |\, F(\tau, x) \,| \leq M(\tau).$$

Then, for each $\alpha \in {}^\star\mathbb{R}$ there exists one and only one internal S-absolutely continuous function $X : \mathbb{T} \to {}^\star\mathbb{R}$ such that

$$\left\{ \begin{array}{rcl} X(0) & = & \alpha \\ X'(\tau) & = & F(\tau, X(\tau)) \quad (\tau \in \mathbb{T} \setminus \{1 - \triangle\}) \end{array} \right. \tag{8}$$

If α is finite, then $X(\mathbb{T}) \subseteq {}^\star\mathbb{R}_b$.

Proof. Define $X : \mathbb{T} \to {}^\star\mathbb{R}$ as in the proof of Nonstandard Peano's Existence Theorem. In this case, for each $\tau = k\triangle \in \mathbb{T}$ we have

$$| X(\tau) - \alpha | = \left| \sum_{i=0}^{k-1} F(i\triangle, X(i\triangle))\triangle \right| \leq \sum_{i=0}^{k-1} M(i\triangle)\triangle \leq \int_{\mathbb{T}} M d\nu$$

and $\int_{\mathbb{T}} M d\nu$ is finite since M is S-integrable. Hence, if α is finite, $X(\mathbb{T}) \subseteq {}^\star\mathbb{R}_b$. It remains to be proven that X is S-absolutely continuous. We will use Theorem 15. Take A an internal subset of $\mathbb{T}$ such that $\nu(A) \approx 0$. Note that

$$\sum_{\tau\in A} | X'(\tau) | \triangle = \sum_{\tau\in A} | F(\tau, X(\tau)) | \triangle \leq \sum_{\tau\in A} M(\tau)\triangle = \int_A M d\nu.$$

Since M is S-integrable, $\int_A M d\nu \approx 0$ and therefore $\int_A | X' | d\nu \approx 0$ which proves that X is S-absolutely continuous. □

7. Carathéodory's Existence Theorem

Using Nonstandard Carathéodory's Existence Theorem we can prove

THEOREM 19. **(Carathéodory's Existence Theorem)** *Suppose that the function $f : [0, 1] \times \mathbb{R} \to \mathbb{R}$ is Lebesgue measurable, continuous in the second variable and let $x_0 \in \mathbb{R}$. If there exists a Lebesgue integrable function $m : [0, 1] \to \mathbb{R}$ such that*

$$\forall (t, x) \in [0, 1] \times \mathbb{R} \quad | f(t, x) | \leq m(t)$$

then there exists a solution x to the problem

$$\left\{ \begin{array}{lcl} x(0) & = & x_0 \\ x'(t) & = & f(t, x(t)) \quad a.a.\ t \in [0, 1] \end{array} \right. \tag{9}$$

Proof. Since $f : [0, 1]\times\mathbb{R} \to \mathbb{R}$ is a Lebesgue measurable function then $F = {}^\star f\,|_{\mathbb{T}\times{}^\star\mathbb{R}} : \mathbb{T} \times {}^\star\mathbb{R} \to {}^\star\mathbb{R}$ is $\star$-measurable. Theorem 14 says ${}^\star m : {}^\star[0, 1] \to {}^\star\mathbb{R}$ is S-integrable and therefore $M = {}^\star m\,|_{\mathbb{T}}$ is also S-integrable. Using the Transfer Principle we conclude that

$$\forall (t, x) \in {}^\star[0, 1] \times {}^\star\mathbb{R} \quad | {}^\star f(t, x) | \leq {}^\star m(t)$$

and then

$$\forall (\tau, x) \in \mathbb{T} \times {}^\star\mathbb{R} \quad | F(\tau, x) | \leq M(\tau).$$

Nonstandard Carathéodory's Existence Theorem shows that there exists an internal S-absolutely continuous $X : \mathbb{T} \to {}^\star\mathbb{R}_b$ such that

$$\left\{ \begin{array}{lcl} X(0) & = & x_0 \\ X'(\tau) & = & F(\tau, X(\tau)) \quad (\tau \in \mathbb{T} \setminus \{1 - \triangle\}) \end{array} \right.$$

and for all $\tau = k\triangle \in \mathbb{T}$

$$X(\tau) = x_0 + \sum_{i=0}^{k-1} F(i\triangle, X(i\triangle))\triangle \in {}^\star\mathbb{R}_b.$$

Since $X(\mathbb{T}) \subseteq {}^\star\mathbb{R}_b$, we can choose $r \in \mathbb{R}^+$ such that

$$\forall \tau \in \mathbb{T} \ \mid X(\tau) \mid \leq r.$$

Defining $x : [0, 1] \to \mathbb{R}$ by

$$x(^\circ\tau) = {}^\circ X(\tau) \qquad (\tau \in \mathbb{T})$$

we conclude, by Theorem 8, that x is absolutely continuous. We will prove that this function is a solution to problem (9).

By hypothesis f is Lebesgue measurable, then the function

$$\widetilde{f} : [0, 1] \to \mathbf{C}([-r, r])$$

where $\mathbf{C}([-r, r])$ denotes the Banach space of real continuous functions on $[-r, r]$, defined by

$$\widetilde{f}(t)(z) = f(t, z) \qquad ((t, z) \in [0, 1] \times [-r, r])$$

is also Lebesgue measurable. Taking the uniform topology in $\mathbf{C}([-r, r])$ and using Anderson's Theorem we can conclude that

$$^\star\widetilde{f} : {}^\star[0, 1] \to {}^\star\mathbf{C}([-r, r])$$

is a lifting of $\widetilde{f}$ we respect of the Loeb measure $^\star\lambda_L$, that is

$$^\star\widetilde{f}(\tau) \approx \widetilde{f}(^\circ\tau) \quad {}^\star\lambda_L - \text{a.a. } \tau \in {}^\star[0, 1].$$

Using the definition of the uniform norm in $\mathbf{C}([-r, r])$ we conclude that

$$\left(\forall z \in {}^\star[-r, r]\ {}^\star f(\tau, z) \approx {}^\star f(^\circ\tau, z)\right) \quad {}^\star\lambda_L - \text{a.a. } \tau \in {}^\star[0, 1].$$

Since f is continuous in the second variable, we obtain that

$$\left(\forall z \in {}^\star[-r, r]\ {}^\star f(\tau, z) \approx f(^\circ\tau, {}^\circ z)\right) \quad {}^\star\lambda_L - \text{a.a. } \tau \in {}^\star[0, 1].$$

Therefore

$$^\star f(\tau, X(\tau)) \approx f(^\circ\tau, {}^\circ(X(\tau))) = f(^\circ\tau, x(^\circ\tau)) \quad \nu_L - \text{a.a. } \tau \in \mathbb{T}$$

because $\nu_L(\mathbb{T}) = 1$.

Finally we may now prove that for all $t \in [0, 1]$,

$$x(t) = x_0 + \int_{[0,t]} f(s, x(s))d\lambda(s)$$

as we did in the proof of Peano's Existence Theorem. □

References

[1] ANDERSON R., *Star-finite representations of measures spaces*, Trans. Amer. Math. Soc. **271** (1982), 667–687.

[2] ARKERYD O., CUTLAND J. AND HENSON C. (EDS.), *Nonstandard analysis, Theory and applications*, Kluwer 1997.

[3] CODDINGTON E. AND LEVINSON N., *Theory of ordinary differential equations*, McGraw-Hill 1955.

[4] CUTLAND N. (ED.), *Nonstandard analysis and its applications*, CUP 1988.

[5] CUTLAND N., NEVES V., OLIVEIRA F. AND PINTO J. (EDS.) *Developments in nonstandard mathematics*, Longman 1995.

[6] CUTLAND N., *Loeb measures in practice: recent advances*, Springer 2000.

[7] DIENER F. AND REEB G., *Analyse non standard*, Hermann 1989.

[8] LOEB P., *Conversion from nonstandard to standard measures spaces and applications in probability theory*, Trans. Amer. Math. Soc. **211** (1975).

[9] ROBINSON A., *Non-standard analysis*, Proc. Roy. Acad. Amsterdam Ser A, **64** (1961), 432–440.

[10] STROYAN K. AND BAYOD J., *Foundations of infinitesimal stochastic analysis*, North-Holland 1986.

[11] STROYAN K. AND LUXEMBURG W., *Introduction to the theory of infinitesimals*, Academic Press 1976.

[12] TUCKEY C., *Nonstandard methods in the calculus of variations*, Longman Scientific & Technical 1993.

AMS Subject Classification: 26E35, 28E05, 34A12.

Natália MARTINS, Vítor NEVES, Department of Mathematics, University of Aveiro, Campus
Universitário de Santiago, 3810-193 Aveiro, PORTUGAL
e-mail: `nataliam@mat.ua.pt, vneves@mat.ua.pt`

Rend. Sem. Mat. Univ. Pol. Torino - Vol. 63, 4 (2005)
Control Theory and Stabil., I

B. Picasso – A. Bicchi

CONTROL SYNTHESIS FOR PRACTICAL STABILIZATION OF QUANTIZED LINEAR SYSTEMS*

Abstract. In this work we face the stability problem for quantized control systems (QCS). A discrete–time single–input linear model is considered and, motivated by technological applications, we assume that a uniform quantization of the control set is a priori fixed. As it is well known, for QCS only practical stability properties can be achieved, therefore we focus on the existence and construction of quantized controllers capable of steering a system to within invariant neighborhoods of the equilibrium.
The main contribution of the paper consists in a theorem which provides a condition for the practical stabilization in a fixed number of steps: not only the result is interesting in itself, but also it enables to construct a family of stabilizing controllers by means of Model Predictive Control (MPC) techniques.
In the last part of the paper some results on the characterization of controlled–invariant sets are reviewed and a lower bound on the size of invariant sets is provided. The bound is attained by an explicitly constructed element.

1. Introduction

The interest of the control community for quantized control systems (QCS) has been considerably raising in the past twenty years. Situations in which quantization may arise and can not be neglected are varied: a popular example is that of "networked control systems", i.e., systems interconnected through communication channels capable of transmitting only a finite amount of information between the plant and the controller.

Special attention has been devoted to the stabilization problem for QCS (see for instance [5, 6, 7, 9, 10, 11, 13, 14, 19, 21]): in [6] the author clarifies that asymptotic stability is a too strong requirement for QCS, hence practical stability concepts have been considered.

Unlike most of the existing literature (where quantization is considered as a parameter to be designed), our work is inspired by the belief that another kind of question is as much important: the stability problem for systems whose quantized resources (i.e., discrete input and output sets) are *fixed a priori*. Such analysis is helpful because it allows to decide in advance whether a desired control objective can be achieved by using a *given* technology (actuators, sensors, communication and computational means). Moreover, issues of this kind may also represent the basis to solve more general stabilization problems (remarkable examples are presented in [9]).

This paper is focused on the stabilization of single–input discrete–time linear systems where a uniformly quantized control set is given.

*This work was supported by European Commission through the IST RECSYS project and by MIUR PRIN 095297_002-2002 "Embedded control of dynamical systems with limited resources for computation and communication".

Let Ω and X_0 be two neighborhoods of the origin with $\Omega \subseteq X_0$: the aim is to design (X_0, Ω)–stabilizing controllers, that is feedback control laws rendering both Ω and X_0 invariant and such that all the states of X_0 are initial points of trajectories which enter Ω in a finite time.
Since the control set is given, the problem is not feasible for all pairs (X_0, Ω). Previous results on the construction of invariant neighborhoods obtained in [15, 14] are useful in this context and briefly reported in Section 3.1. In particular, a continuous family of invariant sets that includes a minimal element called $Q_n(\epsilon)$ is constructed. In the same section we review necessary and sufficient conditions on the control set diameter ensuring that the system is $\big(X_0, Q_n(\epsilon)\big)$–stabilizable.

Section 3.2 is the core of this work and contains the main original contribution: the $\big(X_0, Q_n(\epsilon)\big)$–stabilizability problem enforcing a bound on the number of steps to converge within $Q_n(\epsilon)$ is addressed. A sufficient (and in some cases necessary) condition is provided on the diameter of the control set ensuring the desired stability property. This result is interesting in itself and it is also a useful tool for establishing feasibility of optimal control problems. This leads to the construction of a family of stabilizing controllers by application of Model Predictive Control (MPC) techniques.
Since the goal is the stabilization of the system near the origin, we are interested in confining trajectories within small controlled–invariant neighborhoods of 0. Hence, in Section 4, we state some minimality properties of $Q_n(\epsilon)$ and prove the main result on the subject.
Notation: $Q_n(\Lambda) := \big[-\frac{\Lambda}{2}; \frac{\Lambda}{2}\big]^n = \big\{x \in \mathbb{R}^n \mid \|x\|_\infty \le \frac{\Lambda}{2}\big\}$ is the hypercube of edge length Λ whilst $Q_n^o(\Lambda) := \big[-\frac{\Lambda}{2}; \frac{\Lambda}{2}\big)^n$ is the semi–open hypercube. $\lfloor x \rfloor := \max\{z \in \mathbb{Z} \mid z \le x\}$ and $\lceil x \rceil := \min\{z \in \mathbb{Z} \mid z \ge x\}$ are the floor and the ceil function. ${}^cE$ denotes the complementary of E, $-E := \{x \in \mathbb{R}^n \mid -x \in E\}$, $\mathrm{diam}(E) := \sup\big\{\|x-y\|_2 \mid (x,y) \in E \times E\big\}$ is the diameter of E, $\mathrm{Pr}_i\, x := x_i$ is the projection on the i^{th} coordinate axis and $\mathrm{diam}_i\, \Omega := \mathrm{diam}\big(\mathrm{Pr}_i\, \Omega\big)$. $|A|$ is the matrix defined by $|A|_{i,j} := |A_{i,j}|$, x' denotes the transpose of the vector x, $Q = Q' > 0$ means that Q is a symmetric positive definite matrix. $x^+(t)$ denotes $x(t+1)$: the dependance on t will be often omitted.

2. Preliminaries

We deal with a single–input discrete time–invariant linear system subject to a fixed uniformly quantized control set, more precisely:

$$\text{(1)} \qquad \begin{cases} x^+(t) = Ax(t) + bu(t) \\ x \in \mathbb{R}^n, \quad u \in \mathcal{U} \subseteq \epsilon\,\mathbb{Z} \quad (\epsilon > 0) \\ A \in \mathbb{R}^{n\times n}, \quad b \in \mathbb{R}^n, \end{cases}$$

We suppose that the pair (A, b) is reachable. In this case, changing the coordinates in the state space, we can assume

H1) the pair (A, b) is in *controller form*:

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \alpha_1 & \alpha_2 & \alpha_3 & \cdots & \alpha_n \end{pmatrix}, \; b = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix},$$

where $s^n - \alpha_n s^{n-1} - \cdots - \alpha_2 s - \alpha_1$ is the characteristic polynomial of A.
If $\sum_{i=1}^n |\alpha_i| < 1$, the system is asymptotically stable, we hence assume
H2) $\sum_{i=1}^n |\alpha_i| \geq 1$.
Let us recall the basic definitions about invariant sets [4]:

DEFINITION 1. *The set* $\Omega \subseteq \mathbb{R}^n$ *is said to be* positively invariant *for a closed–loop system* $x^+ = f(x)$ *iff* $\forall x \in \Omega$, $x^+ \in \Omega$*;*

DEFINITION 2. *The set* $\Omega \subseteq \mathbb{R}^n$ *is said to be* controlled–invariant *for system (1) iff* $\forall x \in \Omega \, \exists u \in \mathcal{U}$ *such that* $x^+ = Ax + bu \in \Omega$.

The weak (practical) stability notion we will use is the (X_0, Ω)*–stability* (see also [9]) :

DEFINITION 3. *Let* $0 \in \Omega \subseteq X_0 \subseteq \mathbb{R}^n$ *with* Ω *being a neighborhood of 0 ; a feedback law* $u : \mathbb{R}^n \to \mathcal{U}$ *is said to be* (X_0, Ω)–stabilizing *iff* Ω *and* X_0 *are positively invariant for the closed–loop system* $x^+ = Ax + bu(x)$ *and* $\forall x(0) \in X_0 \, \exists t_{x(0)} \in \mathbb{N}$ *such that* $x(t_{x(0)}) \in \Omega$.
If moreover $\forall x(0) \in X_0 \; t_{x(0)} \leq H_p$*, then the feedback is said to be* (X_0, Ω)-stabilizing in H_p steps.
System (1) is said to be (X_0, Ω)-stabilizable (in H_p steps) *iff there exists an* (X_0, Ω)-*stabilizing (in* H_p *steps) feedback law.*

3. Invariant sets and stabilizing control laws

3.1. Review

We briefly review some basic results concerning the practical stabilization problem: for a more detailed treatment we refer to [14].

THEOREM 1. *If* $\mathcal{U} = \epsilon\,\mathbb{Z}$*, then* $\forall H_p \geq n$ *and* $\forall \Delta \geq \epsilon$*, system (1) is* $\big(Q_n(\Delta),\, Q_n(\epsilon)\big)$*–stabilizable in* H_p *steps.*

Theorem 1 holds for arbitrarily large Δ 's because the control set is unbounded. In the finite control set case we have analyzed the invariance and stabilizability properties for control sets of the type $\mathcal{U}_k := \{-k\epsilon, \ldots, 0, \ldots, +k\epsilon\}$, $\forall k \in \mathbb{N}$.
Let us introduce the *saturated* quantized deadbeat controllers:

DEFINITION 4. *Let* $k \in \mathbb{N}$ *and*

$$w(x) := \begin{cases} -k\,\epsilon & \textit{if} \quad \left\lfloor -\frac{\sum_{i=1}^{n}\alpha_i x_i}{\epsilon} + \frac{1}{2} \right\rfloor < -k \\ \left\lfloor -\frac{\sum_{i=1}^{n}\alpha_i x_i}{\epsilon} + \frac{1}{2} \right\rfloor \cdot \epsilon & \textit{otherwise}\,. \end{cases}$$

The feedback law $u : \mathbb{R}^n \to \mathcal{U}_k$ *defined by*

$$\begin{cases} u(x) = w(x) & \textit{if } \sum_{i=1}^{n}\alpha_i x_i \geq 0 \\ u(x) = -w(-x) & \textit{otherwise}\,. \end{cases} \tag{2}$$

is called the k–levels saturated *quantized deadbeat controller* $\big([k]$*qdb–controller*$\big)$. *Denote by* Ξ *the region where the controller saturates, namely* $\Xi = \Xi_1 \cup (-\Xi_1)$, *where*
$\Xi_1 := \left\{ x \in \mathbb{R}^n \,\middle|\, \left\lfloor -\frac{\sum_{i=1}^{n}\alpha_i x_i}{\epsilon} + \frac{1}{2} \right\rfloor < -k \right\} = \left\{ x \in \mathbb{R}^n \,\middle|\, \sum_{i=1}^{n}\alpha_i x_i > k\epsilon + \frac{\epsilon}{2} \right\}.$

LEMMA 1. *Let* $k \in \mathbb{N}$, *consider the closed–loop dynamics induced by the* $[k]$*qdb–controller, then*

$$|x_n^+| \leq \tfrac{\epsilon}{2} \iff x \notin \Xi \iff \left| \sum_{i=1}^{n}\alpha_i x_i \right| \leq \tfrac{\epsilon}{2} + k\epsilon\,.$$

Consider system (1), assume that $\Delta \geq \epsilon$ and let

$$k(\Delta) := \left\lceil \frac{1}{2}\frac{\Delta}{\epsilon}\Big(\sum_{i=1}^{n} |\alpha_i| - 1 \Big) \right\rceil. \tag{3}$$

PROPOSITION 1. *Assume that* $\mathcal{U} = \mathcal{U}_k$ *and* $\Delta \geq \epsilon$, *the following properties are equivalent:*
ı) $Q_n(\Delta)$ *is controlled–invariant;*
ıı) $k \geq k(\Delta)$*;*
ııı) $Q_n(\Delta)$ *is positively invariant for the closed–loop system* $x^+ = Ax + b\,u(x)$, *where* $u(x)$ *is the* $[k]$*qdb–controller.*

The basic result concerning the stabilizability analysis is

PROPOSITION 2. *Let* $\Delta > \epsilon$, *consider* $k(\Delta)$ *as in Equation (3) and*

$$\bar{k} := \begin{cases} k(\Delta) & \textit{if } \frac{1}{2}\frac{\Delta}{\epsilon}\big(\sum_{i=1}^{n} |\alpha_i| - 1 \big) \notin \mathbb{N} \\ k(\Delta) + 1 & \textit{otherwise}\,. \end{cases}$$

The $[\,\bar{k}\,]$*qdb–controller is* $\big(Q_n(\Delta)\,,\, Q_n(\epsilon)\big)$*–stabilizing.*

One could expect that, in order to achieve the $\big(Q_n(\Delta)\,,\, Q_n(\epsilon)\big)$–stability, it is necessary a control set diameter larger than the one ensuring the invariance of $Q_n(\Delta)$. On the contrary Proposition 2 shows that, generically, such a diameter is also sufficient for the $\big(Q_n(\Delta)\,,\, Q_n(\epsilon)\big)$–stabilizability. This phenomenon is referable to the control quantization: in fact, the minimal diameter of the control set ensuring the invariance of $Q_n(\Delta)$ is larger than the one necessary in the continuous case (because of the ceil function) so that the controller has enough authority to achieve also the convergence towards $Q_n(\epsilon)$.

3.2. Synthesis of stabilizing control laws: Model Predictive Control

The rather strong property of the control set described in Proposition 2, to be both necessary and generically sufficient for $\big(Q_n(\Delta)\,,\ Q_n(\epsilon)\big)$–stabilizability, has as a counterpart a relative weakness, in that no bound on the number of steps necessary to reach $Q_n(\epsilon)$ can be enforced. On the other hand, it is natural to expect that a larger control set would ensure better performance in terms of convergence time. We are hence interested in looking for conditions on the control set diameter for the $\big(Q_n(\Delta)\,,\ Q_n(\epsilon)\big)$–stabilizability in a fixed number of steps.

THEOREM 2. *Let* $\Delta > \epsilon > 0$ *and* $\mathcal{U} \subseteq \epsilon\,\mathbb{Z}$. *Fix* $H_p \geq n:\ H_p = n + p - 1$ *with* $p \geq 1$. *A sufficient condition in order that the system is* $\big(Q_n(\Delta)\,,\ Q_n(\epsilon)\big)$*–stabilizable in* H_p *steps is that* $\mathcal{U}_{k_p} \subseteq \mathcal{U}$, *with*

$$k_p = \left\lceil \frac{1}{2}\frac{\Delta}{\epsilon}\Big(\sum_{i=1}^{n} |\alpha_i| - 1 \Big) + \frac{1}{\epsilon}\frac{\Delta - \epsilon}{2\psi_p} \right\rceil , \tag{4}$$

where the sequence $\{\psi_m\}_{m\in\mathbb{N}\setminus\{0\}}$ *is defined as follows:*

$$\begin{cases} \psi_1 := 1 \\ \psi_m := 1 + \sum_{i=1}^{m-1} |\alpha_{n-m+i+1}|\, \psi_i\,, \quad m \geq 2\,, \\ \qquad\ \text{where } \alpha_j := 0 \text{ if } j \leq 0\,. \end{cases} \tag{5}$$

Moreover, if $\alpha_i \geq 0\ \forall\, i = 1, \ldots, n$, *the bound* k_p *is strict, that is* $\mathcal{U}_{k_p - 1}$ *does not make the system* $\big(Q_n(\Delta)\,,\ Q_n(\epsilon)\big)$*–stabilizable in* H_p *steps.*

Proof. The proof of the Theorem will be given in Section 5 by showing that the $[k_p]$qdb–controller is $\big(Q_n(\Delta)\,,\ Q_n(\epsilon)\big)$–stabilizing in H_p steps. □

Theorem 2 is the main contribution of this paper. Actually, although the sufficiency of the bound is proved by exhibition of a controller achieving the desired performance, the result should be interpreted as the condition for the *existence* of a stabilizing feedback law. It is then interesting to look for other control laws different from the saturated quantized deadbeat. To this aim Theorem 2 is useful because, as it is explained below, it provides a condition for the applicability of *Model Predictive Control* techniques (MPC) which enable us to construct a family of $\big(Q_n(\Delta)\,,\ Q_n(\epsilon)\big)$–stabilizing feedback laws.

Let $\Delta > \epsilon > 0$, assume that $\mathcal{U} \subseteq \epsilon\,\mathbb{Z}$ is such that $Q_n(\epsilon)$ is controlled–invariant, hence define a feedback law

$$F_\epsilon : Q_n(\epsilon) \longrightarrow \mathcal{U}$$

rendering $Q_n(\epsilon)$ positively invariant. We use model predictive control techniques to define a controller in $Q_n(\Delta) \setminus Q_n(\epsilon)$ that steers the states to within $Q_n(\epsilon)$ in finite time, then switch to the feedback law F_ϵ .
To this aim, let $L(x,u) = I_{{}^c Q_n(\epsilon)}(x) \cdot \big(x'Qx + Ru^2\big)$ represent a cost function, where: $I_{{}^c Q_n(\epsilon)}$ is the characteristic function † of ${}^c Q_n(\epsilon)$, $Q \in \mathbb{R}^{n\times n}$ and $Q = Q' > 0$, $R \in \mathbb{R}$ and $R > 0$. For a fixed a number of steps $H_p > 0$, the model predictive controller is defined as

$$u(x) = U_0^*(x)\,,$$

where $U_0^*(x) \in \mathcal{U}$ is the first element of a minimizing sequence (if it exists) $U^*(x) = \big(U_0^*(x), U_1^*(x), \dots, U_{H_p-1}^*(x)\big) \in \mathcal{U}^{H_p}$ of the following optimization problem:

$$\text{(6a)} \qquad \min U \in \mathcal{U}^{H_p} \left\{ J(U,x) = \sum_{k=0}^{H_p-1} L\big(x(k), U_k\big) \right\}$$

subject to

$$\text{(6b)} \qquad \begin{cases} x(0) := x \\ x(k+1) := Ax(k) + bU_k\,, & k = 0,\dots,H_p-1 \\ x(k+1) \in Q_n(\Delta)\,, & k = 0,\dots,H_p-1 \\ x(H_p) \in Q_n(\epsilon)\,, \end{cases}$$

where $x(0)$ is the current state, $\big(x(1),\dots,x(H_p)\big)$ is the predicted trajectory for the future H_p steps when the control sequence $U = \big(U_0, U_1, \dots, U_{H_p-1}\big) \in \mathcal{U}^{H_p}$ is applied.
Assume that $\forall\, x \in Q_n(\Delta)$ the optimization problem (6) is solvable (i.e., the minimum is attained), then the feedback law

$$\text{(7)} \qquad \begin{gathered} F : Q_n(\Delta) \longrightarrow \mathcal{U} \\ F(x) := \begin{cases} F_\epsilon(x) & \text{if } x \in Q_n(\epsilon) \\ U_0^*(x) & \text{otherwise}\,, \end{cases} \end{gathered}$$

is well defined and will be referred to as quantized–MPC controller.

PROPOSITION 3. *The quantized–MPC controller is* $\big(Q_n(\Delta), Q_n(\epsilon)\big)$*–stabilizing.*

Proof. The proof of the Proposition follows the same arguments used for the classical dual–mode MPC scheme: see for instance [18]. □

In order to apply MPC techniques we must first guarantee that $\forall\, x \in Q_n(\Delta)$ the optimization problem (6) is solvable. To this aim it is sufficient to ensure the existence of a control sequence $U \in \mathcal{U}^{H_p}$ so that the constraints (6b) are satisfied. The quantized–MPC controller is well defined if and only if $\forall\, x \in Q_n(\Delta)$ there exists $U \in \mathcal{U}^{H_p}$

†That is $I_{{}^c Q_n(\epsilon)}(x) = \begin{cases} 1 & \text{if } \; x \in {}^c Q_n(\epsilon) \\ 0 & \text{otherwise}\,. \end{cases}$

such that the predicted trajectory lies within $Q_n(\Delta)$ and enters $Q_n(\epsilon)$ after H_p steps: this is equivalent to the requirement that the system is $\big(Q_n(\Delta)\,,\ Q_n(\epsilon)\big)$–stabilizable in H_p steps. By the way, notice that even if this condition is satisfied, the feedback law (7) is $\big(Q_n(\Delta)\,,\ Q_n(\epsilon)\big)$–stabilizing but not necessarily in H_p steps since at each time instant only the first element of the minimizing sequence is applied.
When $\mathcal{U} = \epsilon\,\mathbb{Z}$ the quantized–MPC controller is well defined $\forall\, H_p \geq n$ as a consequence of Theorem 1 . In the finite control set case the problem is more complicate: a sufficient condition ensuring that the quantized–MPC controller is well defined is provided by Theorem 2 .

For a fixed number of steps H_p , different quantized–MPC controllers achieving the $\big(Q_n(\Delta)\,,\ Q_n(\epsilon)\big)$–stability can be obtained by varying at discretion the matrices Q and R which enter in the definition of the model predictive controller. Also H_p is a parameter that can be varied provided the optimization problem remains solvable: an obvious necessary condition is $H_p \geq n$.
The quantized-MPC controller can be implemented by modelling system (1) as a *Mixed Logical Dynamical* system (see [2, 16]) : in this framework efficient algorithms to solve the optimization problem (6) are available.

For a more detailed treatment we refer to [16, 17] and to the literature about model predictive control (see in particular [12]) .

4. Geometric properties of invariant sets

Since the aim is the stabilization of the system near the origin, we are interested in confining the trajectories within small controlled–invariant neighborhoods of 0 . It is then proper to investigate the minimality properties of $Q_n(\epsilon)$. We state two theorems on minimality of $Q_n(\epsilon)$ that provide the so–called *weak* and *strong* minimality properties, then we give the proof of the *strong* minimality theorem.

Obviously, if A is a stable matrix, there exist invariant sets of arbitrarily small size: therefore we will be interested only in the case of unstable matrices.
Throughout this section we will assume without loss of generality that $\mathcal{U} \subseteq \mathbb{Z}$, that is $\epsilon = 1$.

THEOREM 3. *[Weak minimality] If Ω is a bounded controlled–invariant neighborhood of the origin and A is an unstable matrix, then $\forall\, i = 1, \ldots, n$, $diam_i\, \Omega \geq 1$.*

The general property for controlled–invariant sets stated in Theorem 3 provides also a minimality property for $Q_n(1)$: indeed such set has the minimum diameter in all the coordinate directions. In particular, even if controlled–invariant neighborhoods of the origin contained in $Q_n(1)$ can exist, they have the same size as $Q_n(1)$. The size is measured in terms of the diameters of the set along the directions of the coordinate axes.

EXAMPLE 1. Consider $Q_n^o(1)$: it holds that $\forall\, x \in Q_n^o(1)$ there exists a unique $u \in \mathbb{Z}$

such that $x^+ \in Q_n^o(1)$ (see [14]). It is hence univocally defined the mapping

$$T: \begin{array}{ccc} Q_n^o(1) & \to & Q_n^o(1) \\ x & \mapsto & x^+ \end{array}, \tag{8}$$

where $x^+ = Ax + b\,u(x)$ and $u(x) \in \mathbb{Z}$.
Assume that A is an unstable matrix such that $0 < |\det A| < 1$. $T\,Q_n^o(1)$ is obviously a controlled–invariant neighborhood of the origin. Moreover, $T\,Q_n^o(1)$ is strictly contained in $Q_n^o(1)$ because, denoted by λ the Lebesgue measure, from $|\det A| < 1$ it follows that $\lambda\big(T\,Q_n^o(1)\big) < \lambda\big(Q_n^o(1)\big)$.
Hence,

$$Q_n^o(1) \supset T\,Q_n^o(1) \supset \cdots \supset T^k\,Q_n^o(1) \supset \cdots$$

is a strictly decreasing sequence of controlled–invariant neighborhoods of the origin. The typical structure of one of the sets of the sequence (in the two dimensional case) is represented by the shaded region in Fig. 1.

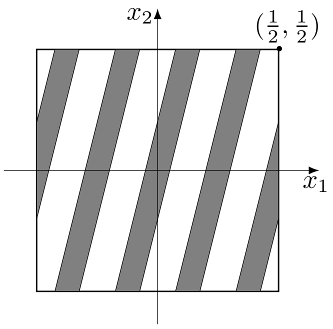


Figure 1: $T\,Q_n^o(1)$

THEOREM 4. *[Strong minimality] If* $|\alpha_1| > 1 + \sum_{i=2}^n |\alpha_i|$ *and* $\Omega \subseteq Q_n^o(1)$ *is a controlled–invariant neighborhood of the origin, then* $\Omega = Q_n^o(1)$.

Proof. The matrix A is invertible and

$$A^{-1} = \begin{pmatrix} -\frac{\alpha_2}{\alpha_1} & -\frac{\alpha_3}{\alpha_1} & \dots & -\frac{\alpha_n}{\alpha_1} & \frac{1}{\alpha_1} \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix}.$$

Hence,

$$(A^{-1}x)_j = \begin{cases} \frac{x_n - \sum_{i=1}^{n-1} \alpha_{i+1} x_i}{\alpha_1} & \text{if } j = 1 \\ x_{j-1} & \text{otherwise}. \end{cases} \tag{9}$$

Let $\theta := \frac{(1+\sum_{i=2}^{n}|\alpha_i|)}{|\alpha_1|}$, by the hypothesis $\theta < 1$. $\forall x \in \mathbb{R}^n$,

$$(10) \qquad \left|(A^{-1}x)_1\right| \le \frac{(1+\sum_{i=2}^{n}|\alpha_i|)}{|\alpha_1|} \cdot \|x\|_\infty = \theta \cdot \|x\|_\infty < \|x\|_\infty .$$

Equations (9) and (10) imply that $A^{-1}Q_n^o(1) \subset Q_n^o(1)$, thus

$$(11) \qquad A^{-h}Q_n^o(1) \subseteq A^{-h+1}Q_n^o(1) \subseteq \cdots \subseteq A^{-1}Q_n^o(1) \subset Q_n^o(1) \quad \forall h \in \mathbb{N}.$$

Moreover, $A^{-n}Q_n^o(1) \subseteq Q_n(\theta)$ in fact: by Equation (9) it holds that $\forall x \in Q_n^o(1)$ and $\forall i = 1,\ldots,n$, $\left(A^{-n}x\right)_i = \left(A^{i-n-1}x\right)_1$ and, by Equations (10) and (11), $\left|(A^{i-n-1}x)_1\right| \le \frac{\theta}{2}$. Similarly, $A^{-nk}Q_n^o(1) \subseteq Q_n(\theta^k)$ $\forall k \in \mathbb{N}$. Since $\lim_{k\to+\infty}\theta^k = 0$ and Ω is a neighborhood of the origin, $\exists k \in \mathbb{N}$ such that $Q_n(\theta^k) \subseteq \Omega$, therefore $A^{-nk}Q_n^o(1) \subseteq \Omega$.

Let $T : Q_n^o(1) \to Q_n^o(1)$ be the map defined in Equation (8): the controlled–invariance of Ω is equivalent to $T\,\Omega \subseteq \Omega$. We claim that $T^{nk}\left(A^{-nk}Q_n^o(1)\right) = Q_n^o(1)$, the claim implies the thesis because $Q_n^o(1) = T^{nk}\left(A^{-nk}Q_n^o(1)\right) \subseteq T^{nk}\,\Omega \subseteq \Omega$.

To prove the claim, thanks to Equation (11), it is sufficient to show that for every $x \in A^{-1}Q_n^o(1)$ the map T coincides with A: $Tx = Ax$ if and only if the unique control $u(x) \in \mathbb{Z}$ such that $x^+ \in Q_n^o(1)$ is $u(x) = 0$, which is the case $\forall x \in A^{-1}Q_n^o(1)$. □

It can be shown that the condition ensuring the strong minimality of $Q_n^o(1)$ is only sufficient, nevertheless the result is interesting because it shows that there are cases in which, among the minimal diameter sets (i.e., $\operatorname{diam}_i \Omega = 1 \quad \forall i = 1,\ldots,n$), the whole $Q_n^o(1)$ is actually the minimal one.

5. Proof of Theorem 2

Theorem 2 was first stated without proof in [15]. Although the proof appears to be complicate, it is instead based on simple ideas trickly exploiting the properties of the controller form coordinates. Hence, it is worth recalling that the control acts only on the n^{th} component while the others shift upward.

To prove the theorem we will take advantage of some lemmas: the hypotheses of Theorem 2 are implicitly assumed.

LEMMA 2. *The sequence* $\{\psi_m\}_{m\in\mathbb{N}\setminus\{0\}}$ *(see Equation (5)) is non–decreasing.*

Proof. We argue by induction: $\psi_1 = 1$, $\psi_2 = 1 + |\alpha_n| \ge \psi_1$.

Assume that $\psi_h \leq \psi_{h+1}\ \forall h < m$, then $\psi_m \leq \psi_{m+1}$. In fact:

$$\begin{aligned}\psi_{m+1} - \psi_m &= \sum_{i=1}^{m} |\alpha_{n-m+i}|\psi_i - \sum_{i=1}^{m-1} |\alpha_{n-m+i+1}|\psi_i \\ &= |\alpha_{n-m+1}|\psi_1 + \sum_{i=2}^{m} |\alpha_{n-m+i}|(\psi_i - \psi_{i-1}) \geq 0\end{aligned}$$

by the inductive hypothesis. □

We extend the sequence $\{\psi_m\}_{m\in\mathbb{N}\setminus\{0\}}$ defining $\psi_z = 0\ \ \forall z \in \mathbb{Z},\ z \leq 0$.

Let

$$\varphi := \frac{\Delta}{2}\Big(1 - \sum_{i=1}^{n} |\alpha_i|\Big) + k_p\epsilon\,. \tag{12}$$

By the definition of k_p $\big($see Equation (4)$\big)$ and the fact that $\lceil x \rceil = x + \theta\ \ (0 \leq \theta < 1)$, it is easy to check that $\varphi > 0$.

LEMMA 3. *Suppose A is such that $\alpha_i \geq 0\ \ \forall i = 1, \ldots, n$. Consider the sequence $\{x(t)\}_{t\in\mathbb{T}\subseteq\mathbb{N}}$ recursively defined by*

$$\begin{cases} x(0) := (\frac{\Delta}{2}, \ldots, \frac{\Delta}{2}) \\ \textit{while}\ \big(x_n(t) > \frac{\epsilon}{2}\big) \quad \textit{let} \quad x(t+1) := Ax(t) - b \cdot (k_p\epsilon)\,; \end{cases}$$

then $x_i(t) = \frac{\Delta}{2} - \varphi\ \psi_{t-n+i} \quad \forall i = 1, \ldots, n$ and $\forall t \in \mathbb{T}$.
Moreover, if $x_n(t) > \frac{\epsilon}{2}$ then $x_i(t) > 0 \quad \forall i = 1, \ldots, n$.

Proof. By induction: when $t = 0$ the statement is obvious.
Suppose that $x_n(t) > \frac{\epsilon}{2}$ and that $x_i(t) = \frac{\Delta}{2} - \varphi\ \psi_{t-n+i} \quad \forall i = 1, \ldots, n$, then $x_i(t+1) = \frac{\Delta}{2} - \varphi\ \psi_{t+1-n+i}$. In fact:
if $i < n$, since A is in controller form, $x_i(t+1) = x_{i+1}(t) = \frac{\Delta}{2} - \varphi\ \psi_{t-n+i+1}$.
When $i = n$, using respectively Equation (12), $\psi_z = 0$ for $z \leq 0$ and Equation (5), we have: $x_n(t+1) = \sum_{l=1}^{n} \alpha_l x_l(t) - k_p\epsilon = \sum_{l=1}^{n} \alpha_l(\frac{\Delta}{2} - \varphi\ \psi_{t-n+l}) + \frac{\Delta}{2} - \frac{\Delta}{2}\sum_{i=l}^{n} \alpha_l - \varphi = \frac{\Delta}{2} - \varphi(1 + \sum_{l=1}^{n} \alpha_l \psi_{t-n+l}) = \frac{\Delta}{2} - \varphi(1 + \sum_{l=n+1-t}^{n} \alpha_l \psi_{t-n+l}) = \frac{\Delta}{2} - \varphi(1 + \sum_{j=1}^{t} \alpha_{j-t+n}\psi_j) = \frac{\Delta}{2} - \varphi\ \psi_{t+1}$.
The statement $x_n(t) > \frac{\epsilon}{2} \Rightarrow x_i(t) > 0$ follows immediately by the definition of the sequence $\{x(t)\}_{t\in\mathbb{T}}$ and the controller form of A. □

Denote by $\Xi(|A|, k_p)$ the saturation region of the $[k_p]$qdb–controller for the system $(|A|, b)$. We say that $x \in \mathbb{R}^n$ satisfies the property (P_{k_p}) iff:

$$\begin{cases} x_i \geq 0 \quad \forall i = 1, \ldots, n \\ x \notin \Xi(|A|, k_p)\,. \end{cases} \tag{P_{k_p}}$$

In the proof of Theorem 2 we shall make extensive use of the following

LEMMA 4. *If x satisfies the property (P_{k_p}) and y is such that $|y_i| \le x_i \ \forall i = 1, \dots, n$, then the closed–loop dynamics induced by the $[k_p]$qdb–controller for system (A, b) is such that $|y_n^+| \le \frac{\epsilon}{2}$.*

Proof. We show that $\left|\sum_{i=1}^n \alpha_i y_i\right| \le \frac{\epsilon}{2} + k_p\epsilon$ and conclude applying Lemma 1: $\left|\sum_{i=1}^n \alpha_i y_i\right| \le \sum_{i=1}^n |\alpha_i| x_i \le \frac{\epsilon}{2} + k_p\epsilon$ because $x \notin \Xi(|A|, k_p)$ and Lemma 1. □

Proof of Theorem 2. We show that the $[k_p]$qdb–controller is $\big(Q_n(\Delta),\ Q_n(\epsilon)\big)$–stabilizing in $H_p = n + p - 1$ steps.

Let $k(\Delta)$ be as in Equation (3), since $k_p \ge k(\Delta) \ge k(\epsilon)$ the positive invariance of $Q_n(\Delta)$ and $Q_n(\epsilon)$ is ensured by Proposition 1. The proof of the convergence to $Q_n(\epsilon)$ in the desired number of steps is organized as follows: we first suppose that $\alpha_i \ge 0 \, \forall i = 1, \dots, n$ and prove that the property holds for the trajectory starting from $x(0) := (\frac{\Delta}{2}, \dots, \frac{\Delta}{2})$. This is obtained by showing that:

Statement I) Let $\tilde{t}$ be such that $|x_n(\tilde{t})| \le \frac{\epsilon}{2}$ and $|x_n(t)| > \frac{\epsilon}{2} \ \forall t < \tilde{t}$, then $|x_n(t)| \le \frac{\epsilon}{2} \ \forall t \ge \tilde{t}$;

Statement II) $\tilde{t} \le p$.

Statements I+II imply the assertion for $x(0)$ because $|x_n(p+h)| \le \frac{\epsilon}{2} \ \forall h \ge 0$ and A is in controller form.

The general case $\big(y(0) \in Q_n(\Delta)$ and arbitrary α_i's$\big)$ is proved by comparing the trajectories of the system with the one analyzed in the first part of the proof.

• *First case:* $\alpha_i \ge 0 \ \forall i = 1, \dots, n, \quad x(0) = (\frac{\Delta}{2}, \dots, \frac{\Delta}{2})$.

Proof of statement I) For $t < \tilde{t}$, the state evolves according to the sequence defined in Lemma 3, therefore

$$x_i(t) > 0 \qquad \forall t < \tilde{t} \ \text{and} \ \forall i = 1, \dots, n. \tag{13}$$

Equation (13) and $|x_n(\tilde{t})| \le \frac{\epsilon}{2}$ imply that $x(\tilde{t}-1)$ satisfies the property (P_{k_p}). By Lemma 3 we know that $x_i(\tilde{t}-1) = \frac{\Delta}{2} - \varphi\,\psi_{\tilde{t}-1-n+i} \ \ \forall i = 1, \dots, n$.

We show by induction that

$$\forall h \ge 0, \quad \begin{cases} |x_i(\tilde{t}+h)| \le x_i(\tilde{t}-1) & \forall i = 1, \dots, n \\ |x_n(\tilde{t}+h)| \le \dfrac{\epsilon}{2}; \end{cases}$$

this proves statement I.

Case $h = 0$:

if $i < n$ then $|x_i(\tilde{t})| = x_{i+1}(\tilde{t}-1) = \frac{\Delta}{2} - \varphi\,\psi_{\tilde{t}-n+i} \le \frac{\Delta}{2} - \varphi\,\psi_{\tilde{t}-n+i-1} = x_i(\tilde{t}-1)$, where the inequality follows by Lemma 2.

If $i = n$ then $|x_n(\tilde{t})| \le \frac{\epsilon}{2} < x_n(\tilde{t}-1)$.

Inductive step $h \Rightarrow h+1$:

if $i < n$ then $|x_i(\tilde{t}+h+1)| = |x_{i+1}(\tilde{t}+h)| \le x_{i+1}(\tilde{t}-1)$ by the inductive hypothesis; $x_{i+1}(\tilde{t}-1) \le x_i(\tilde{t}-1)$ as shown in case $h = 0$.

If $i = n$ we know by the inductive hypothesis that $|x_j(\tilde{t}+h)| \le x_j(\tilde{t}-1) \ \forall j = 1, \dots, n$, hence by Lemma 4 it follows that $|x_n(\tilde{t}+h+1)| \le \frac{\epsilon}{2} < x_n(\tilde{t}-1)$.

Proof of statement II) Because of statement I it is sufficient to show that $|x_n(p)| \leq \frac{\epsilon}{2}$. Suppose that $x_n(t) > \frac{\epsilon}{2} \quad \forall\, t < p$ $\big($we have dropped the modulus because of Equation (13)$\big)$, let $r := \sum_{i=1}^{n} \alpha_i x_i(p-1) - k_p\epsilon = \frac{\Delta}{2} - \varphi\,\psi_p$ by Lemma 3. By Lemma 1 $|x_n(p)| \leq \frac{\epsilon}{2}$ if and only if $r \leq \frac{\epsilon}{2}$, namely:

$$\frac{\Delta}{2} - \varphi\,\psi_p \leq \frac{\epsilon}{2} \;\Leftrightarrow\; \varphi \geq \frac{\Delta-\epsilon}{2\psi_p}\,.$$

By Equation (12) $\varphi = \frac{\Delta}{2} - \frac{\Delta}{2}\sum_{i=1}^{n}\alpha_i + k_p\epsilon$, it is then sufficient to show that

$$\min\left\{ m \in \mathbb{N} \,\Big|\, \frac{\Delta}{2} - \frac{\Delta}{2}\sum_{i=1}^{n}\alpha_i + m\epsilon \geq \frac{\Delta-\epsilon}{2\psi_p} \right\} = k_p\,. \tag{14}$$

Indeed, solving for $\mu \in \mathbb{R}$:

$$\frac{\Delta}{2} - \frac{\Delta}{2}\sum_{i=1}^{n}\alpha_i + \mu\epsilon \geq \frac{\Delta-\epsilon}{2\psi_p} \iff \mu \geq \frac{1}{2}\frac{\Delta}{\epsilon}\Big(\sum_{i=1}^{n}\alpha_i - 1\Big) + \frac{1}{\epsilon}\frac{\Delta-\epsilon}{2\psi_p} := \mu_{\min}\,,$$

hence the integer minimum in Equation (14) is $\lceil \mu_{\min} \rceil = k_p$. This concludes the first part of the proof.

From the discussion above it follows immediately that if $\mathcal{U} = \mathcal{U}_{k_p-1}$ and $\alpha_i \geq 0 \;\; \forall\, i = 1, \ldots, n$, then the system is not $\big(Q_n(\Delta), Q_n(\epsilon)\big)$–stabilizable in H_p steps.

• *General case: arbitrary α_i's and $y(0) \in Q_n(\Delta)$.*

The thesis is obtained by comparing the evolution of $y(0)$ according to the $[k_p]$qdb–controller and the evolution of $x(0) = (\frac{\Delta}{2}, \ldots, \frac{\Delta}{2})$ driven by system $(|A|, b)$ and the corresponding $[k_p]$qdb–controller.

From the first part of the proof we know that $\exists\, \tilde{t} \leq p$ such that $|x_n(\tilde{t})| \leq \frac{\epsilon}{2}$ and $x_n(t) > \frac{\epsilon}{2} \quad \forall\, t < \tilde{t}$. Moreover, property (P_{k_p}) holds for $x(\tilde{t}-1)$.
First we show by induction that $\forall\, h \leq \tilde{t}-1$ and $\forall\, i = 1, \ldots, n$, $|y_i(h)| \leq x_i(h)$:
the case $h = 0$ is obvious.
If $h < \tilde{t}-1$ let us show the inductive step $h \Rightarrow h+1$:
if $i < n$ then $|y_i(h+1)| = |y_{i+1}(h)| \leq x_{i+1}(h)$ because of the inductive hypothesis; also, $x_{i+1}(h) = x_i(h+1)$.
For $i = n$, if $|y_n(h+1)| \leq \frac{\epsilon}{2}$ then $|y_n(h+1)| \leq \frac{\epsilon}{2} < x_n(h+1)$ because $h+1 \leq \tilde{t}-1$. If instead $|y_n(h+1)| > \frac{\epsilon}{2}$, by Lemma 1 and the definition of the $[k_p]$qdb–controller, $y_n(h+1) = \sum_{j=1}^{n}\alpha_j y_j(h) \pm k_p\epsilon$ (with $+$ if $\sum_{j=1}^{n}\alpha_j y_j(h) < 0$ and vice versa). Since $h+1 \leq \tilde{t}-1$, then $x_n(h+1) = \sum_{j=1}^{n}|\alpha_j| x_j(h) - k_p\epsilon$. Let us suppose that $\sum_{j=1}^{n}\alpha_j y_j(h) > 0$ (the opposite case is analogue): $y_n(h+1) = \sum_{j=1}^{n}\alpha_j y_j(h) - k_p\epsilon > \frac{\epsilon}{2}$, thus $|y_n(h+1)| = y_n(h+1) \leq \sum_{j=1}^{n}|\alpha_j||y_j(h)| - k_p\epsilon \leq \sum_{j=1}^{n}|\alpha_j| x_j(h) - k_p\epsilon = x_n(h+1)$ where the last inequality follows by the inductive hypothesis.
In particular $|y_i(\tilde{t}-1)| \leq x_i(\tilde{t}-1) \;\; \forall\, i = 1, \ldots, n$: hence, as property (P_{k_p}) is satisfied by $x(\tilde{t}-1)$, by Lemma 4 it holds that $|y_n(\tilde{t})| \leq \frac{\epsilon}{2}$. Since $\tilde{t} \leq p$, to conclude the proof it is sufficient to show that

$$\forall\, h \geq 0\,, \quad \begin{cases} |y_i(\tilde{t}+h)| \leq x_i(\tilde{t}-1) & \forall\, i = 1,\ldots,n \\ |y_n(\tilde{t}+h)| \leq \dfrac{\epsilon}{2}\,. \end{cases}$$

We prove it by induction. Case $h = 0$:
if $i < n$ then $|y_i(\tilde{t})| = |y_{i+1}(\tilde{t}-1)| \leq x_{i+1}(\tilde{t}-1)$ as proved above. We already know (see the proof of statement I) that $x_{i+1}(\tilde{t}-1) \leq x_i(\tilde{t}-1)$.
If $i = n$ then $|y_n(\tilde{t})| \leq \frac{\epsilon}{2} < x_n(\tilde{t}-1)$.
The inductive step $h \Rightarrow h+1$ can be proved in the same way as the analogue property showed in the proof of statement I. □

6. Conclusion

We have considered the practical stabilization problem for discrete–time linear systems subject to a fixed uniformly quantized control set. Several results have been derived taking advantage of the controller form coordinates. In particular we have provided results on the feasibility of optimal control problems which allows the synthesis of stabilizing control laws in the framework of MPC. The approach is promising also to solve more general problems in the most important and challenging area where quantization is combined with limited communication bandwidth.

Acknowledgements Alberto Bemporad is gratefully acknowledged for useful discussions and Frédéric Gouaisbaut for his collaboration in the preliminary phase of this work.

References

[1] BAILLIEUL J., *Feedback coding for information–based control: operating near the data–rate limit*, in: "Proc. of the 41st IEEE Conference on Decision and Control" 2002, 3229–3236.

[2] BEMPORAD A. AND MORARI M., *Control of systems integrating logic, dynamics and constraints*, Automatica **35** (3) (1999), 407–427.

[3] BICCHI A., MARIGO A. AND PICCOLI B., *On the reachability of quantized control systems*, IEEE Trans. Autom. Control **47** (4) (2002), 546–563.

[4] BLANCHINI F., *Set invariance in control*, Automatica **35** (1999), 1747–1767.

[5] BROCKETT R. AND LIBERZON D., *Quantized feedback stabilization of linear systems*, IEEE Trans. Autom. Control **45** (7) (2000), 1279–1289.

[6] DELCHAMPS D.F., *Stabilizing a linear system with quantized state Feedback*, IEEE Trans. Autom. Control **35** (8) (1990), 916–924.

[7] ELIA N. AND MITTER S., *Stabilization of linear systems with limited information*, IEEE Trans. Autom. Control **46** (9) (2001), 1384–1400.

[8] FAGNANI F. AND ZAMPIERI S., *Stability analysis and synthesis for scalar linear systems with a quantized feedback*, IEEE Trans. Autom. Control **48** (9) (2003), 1569–1584.

[9] FAGNANI F. AND ZAMPIERI S., *Quantized stabilization of linear systems: complexity versus performance*, to appear on IEEE Trans. Autom. Control, special issue on "Networked Control Systems" (2004).

[10] ISHII H. AND FRANCIS B.A., *Stabilizing a linear systems by switching control with duel time*, IEEE Trans. Autom. Control **47** (12) (2002), 1962–1973.

[11] LIBERZON D., *Hybrid feedback stabilization of systems with quantized signals*, Automatica **39** (2003), 1543–1554.

[12] MAYNE D.Q., RAWLINGS J.B., RAO C.V. AND SCOKAERT P.M.O., *Constrained model predictive control*, Automatica **36** (6) (2000), 789–814.

[13] NAIR G.N. AND EVANS R.J., *Exponential stabilisability of finite–dimensional linear systems with limited data rates*, Automatica **39** (2003), 585–593.

[14] PICASSO B. AND BICCHI A., *On the stabilization of linear systems under assigned I/O quantization*, submitted.

[15] PICASSO B., GOUAISBAUT F. AND BICCHI A., *Construction of invariant and attractive sets for quantized–input linear systems*, in: "Proc. of the 41st IEEE Conference on Decision and Control" 2002, 824–829.

[16] PICASSO B., PANCANTI S., BEMPORAD A. AND BICCHI A., *Receding–horizon control of LTI systems with quantized inputs*, in: "Proc. of the IFAC Conference on Analysis and Design of Hybrid Systems" 2003, 295–300.

[17] PICASSO B., *Stabilization of quantized–input systems with optimal control techniques*, Degree Thesis: Dipartimento di Matematica L.Tonelli, Università di Pisa – Italia. Available writing at: picasso@piaggio.ccii.unipi.it, 2002.

[18] SCOKAERT P.O.M., MAYNE D.Q. AND RAWLINGS J.B., *Suboptimal model predictive control (feasibility implies stability)*, IEEE Trans. Autom. Control **44** (3) (1999), 648–654.

[19] TATIKONDA S.C., *Control under communication constraints*, PhD Thesis: Massachusetts Institute of Technology, 2000.

[20] WONG W. AND BROCKETT R., *Systems with finite communication bandwidth constraints - part I: State estimation problems*, IEEE Trans. Autom. Control **42** (1997), 1294–1299.

[21] WONG W. AND BROCKETT R., *Systems with finite communication bandwidth constraints - part II: Stabilization with limited information feedback*, IEEE Trans. Autom. Control **44** (5) (1999), 1049–1053.

AMS Subject Classification: 93C55, 93D99, 49N35.

Bruno PICASSO, Science, mathematics, Scuola Normale Superiore, Piazza dei Cavalieri 7, 56126 Pisa, ITALY
e-mail: b.picasso@sns.it

Antonio BICCHI, Centro Interdipartimentale di Ricerca "E. Piaggio", Università di Pisa, Via Diotisalvi 2, 56100 Pisa, ITALY
e-mail: bicchi@ing.unipi.it