ENSAIOS MATEMÁTICOS

SOCIEDADE BRASILEIRA DE MATEMÁTICA

Introduction to the Theory of Systems

I. KUPKA

University of Toronto

VOLUME 4 1991

Ensaios Matemáticos

Esta série objetiva publicar textos sobre temas atuais da Matemática, permitindo ao leitor, através de exposições acessíveis, adquirir ampla perspectiva do assunto em questão, inclusive no que se refere a seus aspectos em aberto. Os trabalhos devem ser enviados a um dos editores.

This series is intended as a vehicle for survey papers in all areas of Mathematics. The aim of the series is to offer readers an accessible overview of current topics in Mathematics. The Brazilian Mathematical Society invites authors to submit papers for inclusion in the series by sending their work to one of the editors.

Full instructions on the preparation of copy are available from the editors upon request.

How to order Ensaios Matemáticos

Ensaios Matemáticos are available directly from the Brazilian Mathematical Society by writing to the following address:

Ensaios Matemáticos
Sociedade Brasileira de Matemática
Estrada Dona Castorina, 110
22460 Rio de Janeiro RJ
Brazil

Financed by the Programa de Apoio a Publicações Científicas CNPq/FINEP.
Partial support from Instituto de Matemática Pura e Aplicada.

CONTENTS

PHILOSOPHICAL INTRODUCTION

What is system theory? System theory is the study of systems. What, then, is a system? A system is best defined by what it does. It is a device or plant that receives inputs and transforms them into outputs which it gives out. Inputs and outputs form a very disparate class of objects. The best is to give examples.

EX. 1: A car: it receives orders from the driver in the form of signals (the inputs) and it transforms them into a ride (the output).

EX. 2: A wireless set or a TV: it receives signals in the form of waves and transforms them into sound waves or images.

EX. 3: A production machine: it receives both allotments of raw materials and signals from the operator and transforms them into manufactured objects.

EX. 4: The economy of a country: it receives raw materials, capitals from investors, work from the labour force, directives from the managers or planners and transforms them into a variety of products.

EX. 5: A person walking: he receives messages from the brain and transforms them into an intricate movement of the body, especially of the legs.

We could go on and on, giving examples. In fact we shall present some more, later on. For the moment, we will formalize what has been done into a definition.

DEFINITION 0: A system is a device represented by a mapping which associates to elements of a set of vector functions called inputs, elements or subsets of another set of vector functions called outputs. This mapping is called the input-output mapping of the system.

With respect to the fact that the mapping might be multivalued, let us discuss briefly the following example:

EX. 6: Inventory management: in the operation of a warehouse, the inputs are the quantities of the stored goods that have been ordered, the outputs are the quantities sold. Since there is a random component in the demand, the output will not be represented by one function but by a whole bunch of them, each fitted with a weight representing the likelihood of that particular demand.

What are the goals of system theory? Besides the obvious goal of broadening our knowledge, system theory has some very specific ones. The system theorist wants to determine whether the system can accomplish certain tasks preassigned by the operator of the plant and to devise rules or algorithms in order to execute the jobs in the most efficient way, according to criteria preset by the operator.

How does the system theorist go about his job?

a) First he has to construct a mathematical model of the system. This is not an entirely scientific task since it involves, quite often, economic and even political considerations. Usually the model constructed contains unknown parameters which have to be determined. This involves procedures which go under the name of identification of parameters.

b) Once this is accomplished, the mathematical study can start. Questions are asked, problems are solved. Their nature varies from system to system but there are some basic ones such as: controllability, observability, reduction of the model, stability, perturbation decoupling, system decoupling, optimal control.

c) The last step is the numerical implementation of the results found in b).

A broad classification of the models used.

There are several categories of models:

C) Deterministic versus stochastic.

D) Continuous versus discrete time.

E) Ordinary versus distributed control.

C and D are self-explanatory. As for E, in the ordinary models the inputs and outputs are represented by functions of one parameter, in the distributed models by functions of several.

CHAPTER I

LINEAR SYSTEMS

§0 Introduction

The models most commonly used in the praxis are the so-called linear systems. They are also the best known ones.

DEFINITION 0: A linear input-output system is a sextuple (U, X, Y, B, A, C) where U, X, Y are three finite dimensional vector spaces over $\mathbf{R}$ called, respectively, the input or control, state and output spaces.

B, A, C are three linear mappings $B : U \to X$, $A : X \to X$, $C : X \to Y$. B, A are called, respectively, the control vector and the drift vector. C is called the output mapping. An element of X is called a state of the system.

REMARK 1: The preceeding notation is very cumbersome; usually we shall represent the system by the triple (B, A, C) and forget about the spaces. When C does not play any role, we simply write (B, A).

§1 The "operation" of a linear system

Given a linear system (B, A, C), its input-output mapping (IOM) is defined as follows.

Let $L^1(U)$ denote the set of all mappings $u : [\alpha(u), \beta(u)] \to U$ which are L^1. Here $\alpha(u)$ is a real number and $\beta(u)$ is real and larger than $\alpha(u)$ or else $\beta(u) = +\infty$. Further, let $AC(X)$ (resp. $AC(Y)$) denote the space of all mappings $x : [\alpha(x), \beta(x)] \to X$ (resp. $y : [\alpha(y), \beta(y)] \to Y$), α and β as above, which are absolutely continuous.

DEFINITION 1: 1) A trajectory of a system (B, A, C) is a pair $(x, u) : [a, b] \to X \times U$ belonging to $AC(X) \times L^1(U)$ such that for almost all $t \in [a, b]$, $\frac{dx}{dt}(t) = Ax(t) + Bu(t)$.

2) The IOM Φ of the system (B, A, C) is the mapping $\Phi : X \times L^1(U) \to AC(Y)$ defined as follows: $\Phi(x_0, u) = y$ where $u : [\alpha(u), \beta(u)] \to U$ belongs to $L^1(U)$, $x : [\alpha(u), \beta(u)] \to X$ is the unique solution of

$$\begin{cases} \frac{dx}{dt}(t) = Ax(t) + Bu(t) \\ x(\alpha(u)) = x_0 \end{cases}$$

defined on $[\alpha(u), \beta(u)]$ and $y : [\alpha(u), \beta(u)] \to Y$ is the function $y(t) = Cx(t)$.

REMARK 2: An important property of Φ is time stationarity, that is: for any $\tau > 0$, $\Phi(x_0, u)_\tau = \Phi(x_0, u_\tau)$. (If $f : [a, b] \to ?$, then $f_\tau : [a - \tau, b - \tau] \to ?$ is given by $f_\tau(t) = f(t + \tau)$.)

REMARK 3: In practice, it can happen that constraints are imposed on the input functions. We formalize this in the next definition.

DEFINITION 2: An admissible space of controls is a subspace $\mathcal{U}$ of $L^1(U)$ having the following properties:

(i) for any $u : [a(u), b(u)] \to U$, $v : [a(v), b(v)] \to U$ belonging to $\mathcal{U}$, if $a(v) = b(u)$, then the function $w = u^* v : [a(u), b(v)] \to U$,

$w(t) = u(t)$ if $t \in [a(u), b(u)]$ and $w(t) = v(t)$ if $t \in [a(v), b(v)]$, belongs to u;

(ii) if $u : [a(u), b(u)] \to U$ belongs to $\mathcal{U}$ and if $\tau \in \mathbf{R}$ then $u_\tau : [a(u) - \tau, b(u) - \tau] \to U$, $u_\tau(t) = u(t + \tau)$, belongs to $\mathcal{U}$.

Example of a linear system.

Ex.: The accelerated vehicle. A vehicle V moves along a prescribed oriented path. Let $x(t)$ be the algebraic arc length between the initial position and the one at time t, m be the mass of V. Then the movement of V is regulated by Newton's law:

$$\text{(NL)} \qquad m\frac{d^2x}{dt^2}(t) = u(t) - e\frac{dx}{dt}(t),$$

where $u(t)$ is the acceleration or braking force of V ($u > 0$ acceleration, $u < 0$ braking) and $-e\frac{dx}{dt}$ represents the resistance of the medium to the movement. To simplify, let $m = 1$.

Clearly u is the control. To set up a linear model of (NL), it is sufficient to take as X the phase space $\mathbf{R}^2$, $x_1 = x$, $x_2 = \frac{dx}{dt}$ and as U, the line $\mathbf{R}$. Then $Bu = ue_2$, and A is represented by the matrix $\begin{bmatrix} 0 & 1 \\ 0 & -e \end{bmatrix}$ in the canonical basis: $e_1 = (1, 0)$, $e_2 = (0, 1)$.

As the output space we can take $Y = X$ and $C =$ the identity of X.

Two of the basic properties of a system are the accessibility (or reachability, or controllability, or transitivity) property and the observability property. Controllability answers the question about what the system can do.

§2 Accessibility or controllability

DEFINITION 3: α) Given a system (B, A) and two states x_0 and x_1

in X, we say that x_1 is accessible from x_0 if there exists a trajectory $(x,u):[0,T_u]\to X\times U$, such that $x(0)=x_0$, $x(T_u)=x_1$. We say that u steers x_0 to x_1.

β) The set of all points x_1 accessible from a given state x_0 is called the accessibility set of x_0 and denoted by $A(x_0)$.

γ) Given a time T, the set of all points x_1 which can be steered to, from x_0, in time T (that is, using a control u with $T_u=T$) is called the accessibility set of x_0 at time T and denoted by $A(x_0,T)$.

REMARK 4: More generally, if $J\subset \mathbf{R}_+$ is a set of times, $A(x,J)$ will denote the union $U\{A(x,T)\mid T\in J\}$.

It is obvious that $A(x)=A(x,\mathbf{R}_+)$.

§3 Main theorem on accessibility

THEOREM 0. *Given a linear system (B,A):*

(i) For any $T>0$, $A(0)$ is the same as $A(0,T)$ and equals the A-invariant vector subspace of X generated by the image $\mathcal{B}$ of B, that is, $A(0)=\sum_{n=0}^{d-1}A^n(\mathcal{B})$, $d=\dim X$.

(ii) For any $T>0$ and $x\in X$, $A(x,T)$ is the affine space $e^{TA}x+A(0,T)$.

(iii) For any $x\in X$, $A(x)$ is the sum $\gamma_+(x)+A(0)$, where $\gamma_+(x)$ is the positive semi-trajectory of $x:\gamma_+(x)=\{e^{tA}x\mid t\geq 0\}$.

PROOF: (i) Clearly $A(0,T)$ is a vector subspace of X for any $T>0$, being the image of the operator $W_T:L^1([0,T];U)\to X$, $W_T(u)=\int_0^T e^{(T-t)A}Bu(t)dt$. Let us determine the annihilator V of $A(0,T)$. By a well known theorem, the annihilator V in X' (dual of X) is the kernel of $W'_T : X' \to L^1([0,T];U)'$ ($'$ denotes the dual or the transpose). It is clear that $p \in X'$ belongs to V if and only if: $B'e^{(T-t)A'}p' = 0$ for all $0 \leq t \leq T$. By the theorem of Hamilton-Cayley this last condition is equivalent to the fact that $p \in \bigcap_{n=0}^{d-1} \mathrm{Ker}(B'A'^n)$. Hence $V = \bigcap_{n=0}^{d-1} \mathrm{Ker}(B'A'^n)$. Since the annihilator of $\mathrm{Ker}\, B'A'^n$ in X is $A^n(\mathcal{B})$, it follows that $\mathcal{A}(0,T) = \sum_{n=0}^{d-1} A^n(\mathcal{B})$. By remark 4 this implies that $\mathcal{A}(0) = \mathcal{A}(0,T) = \sum_{n=0}^{d-1} A^n(\mathcal{B})$.

(ii) follows from (i) and the well known formula

$$x(T) = e^{TA}x_0 + W_T(u)$$

for the solution x of $\frac{dx}{dt} = Ax + Bu$, $x(0) = x_0$. (iii) follows from (ii) by Remark 4. □

DEFINITION 4: A linear system (B,A) is called transitive or controllable if for any pair of states (x_0, x_1), x_0 can be steered to x_1.

COROLLARY 1. *A system (B,A) is transitive if and only if:*

(iv) X is the A-invariant subspace generated by the image of B. This condition is equivalent to either one of the following two:

(v) The rank of the mapping $U^d \to X$,

$$(u_0, \ldots, u_{d-1}) \mapsto \sum_{n=0}^{d-1} A^n B u_n,$$

is d.

(vi) Given two scalar products on U and X respectively, the mapping $L(T) : X \to X$, $L(T) = \int_0^T e^{tA}BB^*e^{tA^*}dt$ is positive

definite for any $T > 0$. $B^* : X \to U$, $A^* : X \to X$ denote the adjoints of B and A with respect to the given scalar products.

PROOF: The Corollary is clear: if the system is controllable then in particular $A(0) = X$. If $A(0) = X$, then (iii) shows that the system is controllable. (v) is a restatement of the condition $A(0) = X$. As for (vi) it is clear that $L(T)$ is positive semi-definite. It is also easy to check that a vector $x \in X$ is orthogonal to $A(0, T)$ if and only if $L(T)x = 0$. □

From a theoretical point of view the preceeding results are nice, but they are not very realistic.

§4 A more realistic point of view

In practice, the control functions are subjected to restrictions. In particular their size is limited. Hence usually, given a linear system (U, X, B, A), there exists a compact convex neighborhood U_1 of 0 such that one may use controls with values in U_1 only. As in the general case, but using controls with values in U_1 only, we can define accessibilily sets $A(x, U_1)$, $A(x, T, U_1)$ exactly as before. We have the following result:

THEOREM 1. *a) The system (U_1, X, B, A) is transitive if and only if the equivalent conditions of Corollary 1 are satisfied and the spectrum of A is purely imaginary. We assume this in b) and c) below.*

b) There exists a function $\tau : X \times X \to \mathbf{R}_+$ such that given any $(x_0, x_1) \in X \times X$, x_1 is reachable from x_0 at time T if and only if $T \geq \tau(x_0, x_1)$.

c) Let $H : X \times X' \times U \to \mathbf{R}$. *($X'$ dual of X) be the function* $H(x,p,u) = \langle Ax + Bu, p\rangle$. *For any pair* $(x_0, x_1) \in X \times X$ *there exists a curve* $(\overline{x}, \overline{p}, \overline{u}) : [0, \tau(x_0, x_1)] \to X \times X' \times U_1$ *such that:*

1) $\overline{x}, \overline{p}$ *are absolutely continuous,* $\overline{u}$ *is measurable.*

2) For almost all $t \in [0, \tau(x_0, x_1)]$

$$\frac{d\overline{x}}{dt}(t) = \frac{\partial H}{\partial p}(\overline{x}(t), \overline{p}(t), \overline{u}(t)), \quad -\frac{d\overline{p}}{dt}(t) = \frac{\partial H}{\partial x}(\overline{x}(t), \overline{p}(t), \overline{u}(t))$$

and

$$H(\overline{x}(t), \overline{p}(t), \overline{u}(t)) = \inf\{H(\overline{x}(t), \overline{p}(t), v) \mid v \in U_1\}.$$

3) $\overline{x}(0) = x_0$, $\overline{x}(\tau(x_0, x_1)) = x_1$.

REMARK 5: Since $\frac{\partial H}{\partial p}(x,p,u) = Ax + Bu$, $\frac{d\overline{x}}{dt}(t) = A\overline{x}(t) + B\overline{u}(t)$ and $(\overline{x}, \overline{u})$ is a trajectory of the restricted system (U_1, X, B, A) taking x_0 to x_1.

PROOF: We are not going to prove this result here. It is a fairly easy application of the so-called maximum principle and the facts that, for any $t > 0$, $A(x_0, t, U_1)$ is a compact convex set and the mapping $t \mapsto A(x_0, t, U_1)$ is continuous (in the Hausdorff topology). □

§5 Observability

The state of an input-output system depends on a lot of parameters ($\dim X$ is large!). Most of these parameters are unknown. All the information available about the system is contained in the observed quantities $y = Cx$. Usually the number ($\dim Y$) of these is small. This leads to the following problem: does the output determine the state of the system?

This can also be stated as follows: given a pair of states (x_1, x_2), $x_1 \neq x_2$, does there exist an input $u \in L^1(U)$ such that $\Phi(x_1, u) \neq \Phi(x_2, u)$?

DEFINITION 5: Two states x_1, x_2, $x_1 \neq x_2$, are called indistinguishable if for all $u \in L^1(U)$, $\Phi(x_1, u) = \Phi(x_2, u)$.

Introducing the functional $X \to \mathrm{Map}(L^1(U), AC(Y))$, $x \mapsto \Phi_x$, $\Phi_x(u) = \Phi(x, u)$, x_1 and x_2 are indistinguishable if $\Phi_{x_1} = \Phi_{x_2}$.

THEOREM 2. *A pair $(x_1, x_2) \in X \times X$ is indistinguishable if and only if $x_1 - x_2 \in I$, where $I = \bigcap_{n=0}^{d-1} \operatorname{Ker} CA^n$.*

PROOF: The proof is trivial since it is clear that $I = \{x \in X \mid Ce^{tA}x = 0 \text{ for all } t\}$. □

REMARK 6: It is clear that indistinguishability does not depend on the inputs. This is a very special phenomenon due to the linearity of the system.

DEFINITION 6: I is called the indistinguishability distribution.

DEFINITION 7: A system (U, X, Y, B, A, C) is called observable if $I = \{0\}$.

COROLLARY 2. *The following properties are equivalent:*

(vii) $I = \{0\}$.

(viii) The mapping $X \to Y^d$, $x \mapsto (Cx, CAx, \ldots, CA^{d-1}x)$ is injective.

(ix) Given any two scalar products on X and Y respectively, for any $T > 0$ the mapping $M(T) = \int_0^T e^{tA^*} C^* C e^{tA} dt$ is positive definite.

PROOF: It is trivial that (vii) and (viii) are equivalent. It is also clear that $M(T)$ is always positive semi-definite. It is not hard to check that x belongs to I if and only if $M(T)x = 0$. □

Ivan Kupka

CHAPTER II

THE CATEGORY OF SYSTEMS

§0 Introduction

In a natural way, the set $\sum$ of all linear input-output systems forms a category. Its objects are the systems. Its morphisms are quite natural but harder to describe. To do this conveniently, let us introduce a nice representation.

MATRIX REPRESENTATION OF SYSTEMS: given a linear system (U, X, Y, B, A, C), it can be represented as a linear mapping $X \times U \to X \times Y$ with matrix $\begin{bmatrix} A & B \\ C & 0 \end{bmatrix}$. This matrix determines the system uniquely.

MORPHISMS OF THE CATEGORY OF SYSTEMS: a morphism $g : (U, X, Y, B, A, C) \to (U_1, X_1, Y_1, B_1, A_1, C_1)$ is represented by two matrices

$$g_r = \begin{bmatrix} \alpha & 0 \\ \beta_1 & \beta_2 \end{bmatrix} : X \times U \to X_1 \times U_1$$

and

$$g_\ell = \begin{bmatrix} \alpha & \gamma_1 \\ 0 & \gamma_2 \end{bmatrix} : X \times Y \to X_1 \times Y_1,$$

such that

$$g_\ell \begin{bmatrix} A & B \\ C & 0 \end{bmatrix} = \begin{bmatrix} A_1 & B_1 \\ C_1 & 0 \end{bmatrix} g_r.$$

Sometimes we shall represent g by (β, α, γ), where

$$\beta : X \times U \to U_1, \quad \beta = \beta_1 + \beta_2, \quad \gamma : Y \to X_1 \times Y_1, \quad \gamma = \gamma_1 + \gamma_2.$$

In this manner we get an additive category but not an abelian one [see appendix]. This category is self-dual: given a system (U, X, Y, B, A, C), its dual is the system (Y', X', U', C', A', B'). In matrix representation, the matrix $\begin{bmatrix} A' & C' \\ B' & 0 \end{bmatrix}$ of the second system is the transpose of the matrix $\begin{bmatrix} A & B \\ C & 0 \end{bmatrix}$ of the first system.

The dual of a morphism $g = (g_\ell, g_r)$ is (g_r', g_ℓ').

The category of systems has finite direct products which are equal to finite direct sums. Let $\Sigma_i = (U_i, X_i, Y_i, B_i, A_i, C_i)$, $1 \leq i \leq p$, be p systems. Then the direct product $\Sigma = \prod_{i=1}^{p} \Sigma_i$, is the system (U, X, Y, B, A, C), where $U = \prod_{i=1}^{p} U_i$, $X = \prod_{i=1}^{p} X_i$, $Y = \prod_{i=1}^{p} Y_i$, $B = \prod_{i=1}^{p} B_i$, $A = \prod_{i=1}^{p} A_i$, $C = \prod_{i=1}^{p} C_i$.

§1 The feedback transformation groups

Given U, X and Y, let us denote by $\sum(U, X, Y)$ the full subcategory of $\sum$ having U, X, Y as control, state and output spaces respectively. Denote by $LT(X \times U)$ (resp. $UT(X \times Y)$) the subgroup of $\mathrm{Isom}(X \times U)$ (resp. $\mathrm{Isom}\,(X \times Y)$) of all lower triangular (resp. upper triangular) invertible matrices

$$\begin{bmatrix} \alpha & 0 \\ \beta_1 & \beta_2 \end{bmatrix} : X \times U \hookleftarrow \qquad (\text{resp. } \begin{bmatrix} \alpha & \gamma_1 \\ 0 & \gamma_2 \end{bmatrix} : X \times Y \hookleftarrow).$$

Definition 0: (i) The extended feedback group, denoted by $EFB(U,X,Y)$, is the subgroup of $UT(X \times Y) \times LT(X \times U)$ of all pairs $g = (g_\ell, g_r)$, $g_\ell = \begin{bmatrix} \alpha & \gamma_1 \\ 0 & \gamma_2 \end{bmatrix}$, $g_r = \begin{bmatrix} \alpha & 0 \\ \beta_1 & \beta_2 \end{bmatrix}$; in other words, all pairs such that $pr_x^1 \circ g_\ell \circ i_x^1 = pr_x \circ g_r \circ i_x$, where $i_x : X \to X \times U$, $i_x^1 : X \to X \times Y$ are the canonical injections and $pr_x : X \times U \to X$, $pr_x^1 : X \times Y \to X$ the canonical projections.

(ii) The control feedback group, denoted by $CFB(U,X,Y)$, is the subgroup of $EFB(U,X,Y)$ of all pairs $g = (g_\ell, g_r)$ such that $\gamma_1 = 0$, $\gamma_2 = Id_Y$.

(iii) The observation feedback group, denoted by $OFB(U,X,Y)$, is the subgroup of $EFB(U,X,Y)$ of all pairs g such that $\beta_1 = 0$, $\beta_2 = Id_U$.

(iv) The restricted control (resp. observation) feedback group is the normal subgroup of $CFB(U,X,Y)$ (resp. $OFB(U,X,Y)$) of all g such that $\alpha = Id_X$, $\beta_2 = Id_U$ (resp. $\gamma_2 = Id_Y$).

Remark: The space Y (resp. U) does not play any role in the definition of $CFB(U,X,Y)$ (resp. $OFB(U,X,Y)$). Hence we will often abbreviate it to $CFB(U,X)$ (resp. $OFB(X,Y)$).

Proposition 1. *(v) The subset $\sum_c(U,X,Y)$ of $\sum(U,X,Y)$ of all controllable systems is invariant under $CFB(U,X,Y)$.*

(vi) The subset $\sum_0(U,X,Y)$ of $\sum(U,X,Y)$ of all observable systems in invariant under $OFB(U,X,Y)$.

Definition 1: We say that a property is a control (resp. observation) feedback invariant if the set of all systems having this property is invariant under CFB (resp. OFB). We say that a function $\lambda : \sum \to E$ (E some set) is control (resp. observation) feedback invariant if λ is constant on the orbits of CFB (resp. OFB).

§2 Kronecker indices

The purpose of this section is to define some invariants of unobserved (resp. uncontrolled) systems (B, A) (resp. (A, C)) under the action of $CFB(U, X)$ (resp. $OFB(X, Y)$).

Let (B, A) be a controllable system. Then $X = \sum_{n=0}^{d-1} A^n(\mathcal{B})$, where $\mathcal{B}$ is the image of B and $d = \dim X$. For each integer k define $X(k) = \sum_{n=0}^{k} A^n(\mathcal{B})$.

LEMMA 0. *(vii) The $X(k)$ form a non-decreasing sequence of subspaces of X and $X(k) = X$ if $k \geq d - 1$. The flag $\{0 = X(-1) \subseteq X(0) = \mathcal{B} \subseteq \cdots \subseteq X(d-1) = X\}$ is a control feedback invariant.*

(viii) For every $k \geq 0$, A induces a surjective feedback invariant, linear mapping $A(k) : \frac{X(k)}{X(k-1)} \to \frac{X(k+1)}{X(k)}$.

The proof of this lemma is trivial.

Let $f(k) = \dim \frac{X(k)}{X(k-1)}$. Lemma 0 implies that f is non-increasing and $f(k) = 0$ if $k \geq d$.

DEFINITION 2: The control Kronecker indices of a controllable system (B, A) are the jump points of f, i.e., k is a Kronecker index if $f(k) < f(k-1)$. This is equivalent to saying that Ker $A(k-1) \neq 0$. There are at most $d(= \dim X)$ Kronecker indices. We order them in a decreasing sequence $\kappa(1) > \kappa(2) > \cdots > \kappa(r)$ and call it Kronecker's list.

PROPOSITION 2. *Kronecker's list is a control feedback invariant.*

OBSERVABILITY KRONECKER INDICES: given an observed system (A, C) we can define its observability Kronecker indices as the control Kronecker indices of the controlled pair (C', A'). They can also be defined directly as follows: for any k, let $X(k) = \bigcap_{n=0}^{k} \text{Ker } CA^n$.

LEMMA 0. *(ix) The $X(k)$ form a non-increasing sequence of subspaces of X and $X(k) = 0$ if $k \geq d-1$. The flag $\{X = X(-1) \supset X(0) = Ker\ C \supset \cdots \supset X(d-1) = 0\}$ is an observability feedback invariant.*

(x) For any $k \geq 0$, A induces an injective output feedback invariant, linear mapping $A(k) : \frac{X(k)}{X(k+1)} \to \frac{X(k-1)}{X(k)}$.

DEFINITION 3: The observability Kronecker indices of an observable system (A, C) are the jump points of the function f, $f(k) = \dim \frac{X(k)}{X(k+1)}$, that is: k is a Kronecker index if $f(k-1) > f(k)$.

We shall not proceed any further in this direction, since: 1) the observability Kronecker indices are much less used than the controllability ones; 2) using the duality, one can pass from one kind of index to the other.

§3 Partial canonical forms

Our goal in the next two sections is to find a canonical (or normal) form of a system under the action of the controllability feedback group. Since we will consider objects with only the controllability (as opposed to observability) label attached to them, we shall drop it altogether.

Consider a controllable system (B, A). To each integer k we associate the subspace $\mathcal{B}(k) = \mathcal{B} \cap A^{-k}(X(k-1))$ of $\mathcal{B}$.

LEMMA 1. *The $\mathcal{B}(k)$'s form a non-decreasing sequence of subspaces of $\mathcal{B}$. They are feedback invariants and $\mathcal{B}(k) = \mathcal{B}(k-1)$, except when k is a Kronecker index of (B, A).*

PROOF: Let $q_k : X(k) \to \frac{X(k)}{X(k-1)}$ be the canonical projection. Then $q_k A^k : \mathcal{B} = X(0) \to \frac{X(k)}{X(k-1)}$ is equal to the composition $A(k-1) \circ \cdots \circ A(0) \circ \mathcal{B}(k)$ is the kernel of $q_k A^k$. This shows that $\mathcal{B}(k) \neq \mathcal{B}(k-1)$ if and only if Ker $A(k-1) \neq 0$, since each factor of $A(j-1) \circ \cdots \circ A(0)$ is onto. It is clear that the $\mathcal{B}(k)$ are feedback invariants. □

DEFINITION 4: The integers $\beta(i) = \dim \mathcal{B}(\kappa(i)) - \dim \mathcal{B}(\kappa(i) - 1)$ are called the Brunovsky indices. They are feedback invariants.

Given any linear mapping $K : X \to U$, denote by A_K the mapping $A + BK$. The following proposition gives a partial normal form for the feedback transformed system (B, A_K). Denote by $\kappa(1) \geq \kappa(2) \geq \cdots \geq \kappa(r)$ the Kronecker list of (B, A) and hence also of (B, A_K).

PROPOSITION 3. *For each integer i, $1 \leq i \leq r$, choose a subspace $\mathcal{B}_{i,0}$ of $\mathcal{B}(\kappa(i))$ complementary to $\mathcal{B}(\kappa(i+1))$. Let $\mathcal{B}^K_{i,j}$ denote the subspace $A^i_K(\mathcal{B}_{i,0})$ of X if $0 \leq j \leq \kappa(i) - 1$. Then:*

(a) X is the direct sum $\displaystyle\bigoplus_{i=1}^{r} \bigoplus_{j=0}^{\kappa(i)-1} \mathcal{B}^K_{i,j}$.

(b) For all i, j, $1 \leq i \leq r$, $0 \leq j \leq \kappa(i) - 2$, A_K maps $\mathcal{B}^K_{i,j}$ isomorphically into $\mathcal{B}^K_{i,j+1}$.

PROOF: $A^{\kappa(i)}_K(\mathcal{B}_{i,0}) \subset \displaystyle\sum_{j=0}^{\kappa(i)-1} A^j_K(\mathcal{B}(\kappa(i))$. This and the controllability of (B,A) imply that X is equal to the sum in a). That this sum is direct, is easy to see as follows: were the sum not direct, one could find vectors $b_{ij} \in \mathcal{B}^K_{i,j}$, not all 0, such that $\sum b_{ij} = 0$. Let q be the largest j such that there is a non zero b_{ij}. Then $\sum_{j=q} b_{ij} = -\sum_{j<q} b_{ij}$. Since $b_{ij} \in X(q-1)$ if $j<q$, $\sum_{j=q} b_{ij} \in X(q-1)$. Let $I_q = \{i \mid b_{iq} \neq 0\}$. For any $i \in I_q$ there is a $b(i) \in \mathcal{B}_{i,0}$ such that $b_{iq} = A^q_K b(i)$, therefore $\sum_{i \in I_q} b(i) \in \mathcal{B}(q)$. Let t be the integer such that $\kappa(t) \leq q < \kappa(t-1)$. Since $b_{iq} \in \mathcal{B}^K_{iq}$ if $i \in I_q$, $q \leq \kappa(i)-1$, and all the i's in I_q are smaller than t and $\sum_{i \in I_q} b(i) \in \mathcal{B}(q) \cap \bigoplus_{i=1}^{t-1} \mathcal{B}_{i,0}$. Since $\mathcal{B}(q) = \mathcal{B}(\kappa(t))$, this intersection is 0. Hence $b(i) = 0$ for all $i \in I_q$, which is a contradiction.

To prove b) assume that for some i,j, $j < \kappa(i)-1$, A_K restricted to $\mathcal{B}^K_{ij}$ is not injective. Then there is a nonzero b in $\mathcal{B}_{i,0}$ such that $A^{j+1}b = 0$. This shows that $b \in \mathcal{B}(j+1)$. Since $j+1 < \kappa(i)$, $b \in \mathcal{B}(\kappa(i+1))$. But then $b \in \mathcal{B}_{i,0} \cap \mathcal{B}(\kappa(i+1)) = 0$. □

§4 The normal form

We shall show that one can choose the linear mapping $K : X \to U$ in such a way that $A^{\kappa(i)}_K(\mathcal{B}_{i,0}) = 0$ for all $i = 1,\ldots,r$.

LEMMA 2. *There is a feedback $K : X \to U$ such that $A^{\kappa(i)}_K(\mathcal{B}_{i,0}) = 0$ for all $1 \leq i \leq r$.*

PROOF: We can always assume that B is injective. Since $A^{\kappa(i)}(\mathcal{B}_{i,0}) \subseteq \sum_{j=0}^{\kappa(i)-1} A^j(\mathcal{B})$, it follows from Proposition 3 that there exist uniquely defined linear mappings $\varphi_{in\ell} : \mathcal{B}_{i,0} \to \mathcal{B}_{\ell,0}$ such that:

$$A^{\kappa(i)} \mid \mathcal{B}_{i,0} = \sum_{n=0}^{\kappa(i)-1} \sum_{k=1}^{r} A^{\kappa(i)-n-1} \varphi_{ink}.$$

To determine K, it is sufficient to compute BK. We compute both the $\mathcal{B}_{ij}^K$ and the restrictions $BK \mid \mathcal{B}_{ij}^K$ by induction on j. For $j = 0$, $\mathcal{B}_{i0}^K = \mathcal{B}_{i0}$, and $BK \mid \mathcal{B}_{i,0} = -\sum_{n=1}^{r} \varphi_{i0n}$. Assume the $\mathcal{B}_{ij}^K$ have been computed for $j \leq m$ and the restrictions $BK \mid \mathcal{B}_{ij}^k$ for $j \leq m-1$. Then we set $BK \mid \mathcal{B}_{i,m}^K = -\sum_{k=1}^{r} \varphi_{imk} \circ L_{im}$, where $L_{im} : \mathcal{B}_{im} \to \mathcal{B}_{i0}^K$ is the inverse of the restriction of A_K^m to $\mathcal{B}_{i,0}$. (This is already defined). Once BK is defined on all the $\mathcal{B}_{i,m}^K$, $i \leq i \leq r$, we set $\mathcal{B}_{i,m+1}^K = (A + BK)(\mathcal{B}_{i,m}^K)$ if $m \leq \kappa(i) - 2$; otherwise $\mathcal{B}_{i,m+1}^K$ is taken to be 0.

We have to check that with this K, $A_K^{\kappa(i)}(\mathcal{B}_{i,0}) = 0$ for all $1 \leq i \leq r$. For this, I claim that for any i, $1 \leq i \leq r$, and j, $0 \leq j \leq \kappa(i) - 1$,

$$A_K^j \mid \mathcal{B}_{i,0} = -\sum_{n=0}^{j-1} \sum_{\ell=1}^{r} A^{j-n-1} \circ \varphi_{in\ell} + A^j \mid \mathcal{B}_{i,0}.$$

We prove this by induction on j. Assume we have proved it for $j \leq m$; then:

$$A_K^{m+1} \mid \mathcal{B}_{i,0} = (A + BK) \circ (A_K^m \mid \mathcal{B}_{i,0}).$$

Since $A_K^m(\mathcal{B}_{i,0}) \subset \mathcal{B}_{i,m}^K$, the above definition of BK implies

$$A_K^{m+1} \mid \mathcal{B}_{i,0} = (A - \sum_{\ell=1}^{r} \varphi_{im\ell} \circ L_{im}) \circ A_K^m \mid \mathcal{B}_{i,0}$$

$$= A \circ A_K^m \mid \mathcal{B}_{i,0} - \sum_{\ell=1}^{r} \varphi_{im\ell};$$

applying the induction hypothesis,

$$A_K^{m+1} \mid \mathcal{B}_{i,0} = - \sum_{n=0}^{m-1} \sum_{k=1}^{r} A^{m-n} \circ \varphi_{ink} - \sum_{\ell=1}^{r} \varphi_{im\ell} + A^{m+1} \mid \mathcal{B}_{i,0},$$

proving the claim for $j = m + 1$.

If we set $j = \kappa(i)$,

$$A_K^{\kappa(i)} \mid \mathcal{B}_{i,0} = - \sum_{n=0}^{\kappa(i)-1} \sum_{\ell=1}^{r} A^{\kappa(i)-n-1} \circ \varphi_{in\ell} + A^{\kappa(i)} \mid \mathcal{B}_{i,0} = 0,$$

which proves Lemma 2. □

Before stating our main Theorem we need a definition.

DEFINITION 5: For any integer n, denote by $\Sigma(n)$ the system defined as follows: $U = \mathbf{R}$, $X = \mathbf{R}^n$ with canonical basis $e_1, \ldots, e_n$, $B(1) = e_1$, $Ae_i = e_{i+1}$ if $i \leq n-1$, $Ae_n = 0$.

THEOREM 1. *Let $\sum_c(U, X)$ denote the set of all controllable systems. The orbits of $\sum_c(U, X)$ under the action of the feedback group $CFB(U, X)$ are in $1-1$ correspondence with the set $K\mathcal{B}$ of all sequences $(\kappa(1), \ldots, \kappa(r), \beta(1), \ldots, \beta(r))$ such that:*

(xi) $\kappa(1) > \kappa(2) > \cdots > \kappa(r) > 0$;

(xii) $\beta(i) > 0$ *for all* i, $1 \leq i \leq r$;

(xiiii) $\sum_{i=1}^{r} \kappa(i)\beta(i) = \dim X;$

(xiv) $\sum_{i=1}^{r} \beta(i) = \dim U.$

Any system $(B, A) \in \sum_c(U, X)$ *is feedback equivalent to* $\bigoplus_{i=1}^{r}(\Sigma(\kappa(i)))^{\beta(i)}$ *if* $(\kappa(i), \ldots, \kappa(r), \beta(i), \ldots, \beta(r))$ *is the Kronecker-Brunovsky sequence associated to* (B, A).

Theorem 1 follows directly from Lemma 2 by choosing a basis in each $\mathcal{B}_{i,0}$.

Finally we have the following trivial Proposition.

PROPOSITION 4. *Any system* $\Sigma(n)$ *is indecomposable.*

§5 Relation with the holomorphic vector bundles on $\mathbf{C}P(1)$ (the Riemann sphere)

To any system (B, A) we can associate a differential operator $D : AC(X) \oplus L^1(U) \to L^1(X)$ as follows: $D(x, u) = Ax + Bu - \frac{dx}{dt}$. Let $X_{\mathbf{C}}$, $U_{\mathbf{C}}$ denote the complexifications of X and U.

The Laplace transform $\hat{D}$ is the mapping $\mathbf{C}P(1) \times X_{\mathbf{C}} \times U_{\mathbf{C}} \to \mathbf{C}P(1) \times X_{\mathbf{C}}$, $(\lambda, \hat{x}, \hat{u}) \mapsto (\lambda, A\hat{x} + B\hat{u} - \lambda\hat{x})$ if $\lambda \neq \infty$ and $(\infty, \hat{x}, \hat{u}) \mapsto (\infty, -\hat{x})$.

LEMMA 3. *(xv) If* (B, A) *is controllable, the mapping* $\hat{D}$ *is onto.*

(xvi) The space $V(B, A) = \operatorname{Ker} \hat{D} = \{(\lambda, \hat{x}, \hat{u}) \mid \hat{D}(\lambda, \hat{x}, \hat{u}) = (\lambda, 0)\}$ *is a holomorphic vector bundle of rank* $\dim U$ *on* $\mathbf{C}P(1)$.

(xvii) The correspondence $\mathcal{F} : \sum_c \ni (B, A) \mapsto V(B, A) \in$ *holomorphic vector bundles on* $\mathbf{C}P(1)$*, is functorial. If a system*

(B, A) is the direct sum of two systems (B_1, A_1) and (B_2, A_2) then $V(B, A) = V(B_1, A_1) \oplus V(B_2, A_2)$.

PROOF: If $\hat{D}$ were not onto, there would exist a $\lambda \in \mathbf{C}P(1)$ and a linear form $p' \in X'_{\mathbf{C}}$, $p' \neq 0$, such that $D(\lambda)'p' = 0$. Then $\lambda \neq \infty$, $\lambda p' = A'p'$ and $B'p' = 0$. These relations imply that $p \in \bigcap_{n=0}^{\infty} \operatorname{Ker} B'A'^n$. (A', B') is not observable, hence (A, B) is not controllable.

Since $\hat{D}$ is holomorphic, (xvi) is an immediate consequence of (xv).

Let $\mu : (B, A) \to (B_1, A_1)$ be a morphism in $\sum(U, X)$: $\alpha A_1 = A\alpha + B\beta_1$, $\alpha B_1 = B\beta_2$. The linear mapping $\tau : X_{\mathbf{C}} \times U_{\mathbf{C}} \to X_{\mathbf{C}} \times U_{\mathbf{C}}$ defined by the matrix $\begin{bmatrix} \alpha & 0 \\ \beta_1 & \beta_2 \end{bmatrix}$ induces a bundle mapping $V(B, A) \to V(B_1, A_1)$. □

Now we state the main Theorem of this section:

THEOREM 2. *(xviii) Given any controllable system (B, A) having the sequence $(\kappa(1), \ldots, \kappa(r), \beta(1), \ldots, \beta(r))$ as the Kronecker-Brunovsky list, the vector bundle $V(B, A)$ associated to it is isomorphic to $\bigoplus_{i=1}^{r} \mathcal{O}(-\kappa(i))^{\beta(i)}$. $\mathcal{O}(-1)$ is the tautological bundle on $\mathbf{C}P(1)$, $\mathcal{O}(-n) = \mathcal{O}(-1)^{\oplus n}$ if $n \geq 0$, $\mathcal{O}(n)^m$ is the direct sum $\mathcal{O}(n) \oplus \cdots \oplus \mathcal{O}(n)$, m times.*

PROOF: (B, A) is feedback equivalent to $S = \bigoplus_{i=1}^{r} (\Sigma(\kappa(i)))^{\beta(i)}$ by Theorem 1, therefore, in view of Lemma 3-(xvii), it is sufficient to prove (xviii) for this last system S. Now by Lemma 3-(xvii),

$V(S) = \oplus V(\Sigma(\kappa(i)))^{\beta(i)}$. Hence we have to determine $V(\Sigma(n))$, n integer. Since the control space of $\Sigma(n)$ is one dimensional, $V(\Sigma(n))$ is a line bundle. Calling $e_1, \ldots, e_n$ the canonical basis of $\mathbf{C}^n$, the following two mappings

$$s : \mathbf{C}P(1) - \{\infty\} \to V(\Sigma(n)), \quad \sigma : \mathbf{C}P(1) - 0 \to V(\Sigma(n))$$

are sections of $V(\Sigma(n))$ and generate $V(\Sigma(n))$ on $\mathbf{C}P(1) - \{\infty\}$ and $\mathbf{C}P(1) - \{0\}$ respectively:

$$s(\lambda) = (\lambda^{n-1}e_1 + \cdots + \lambda e_{n-1} + e_n, \lambda^n) \in \mathbf{C}^n \times \mathbf{C}, \quad \lambda \in \mathbf{C}P(1) - \{\infty\},$$

$$\sigma(z) = (ze_1 + \cdots + z^n e_n, 1) \in \mathbf{C}^n \times \mathbf{C},\ z \text{ coordinate on } \mathbf{C}P(1) - \{0\}.$$

On $\mathbf{C}P(1) - \{0, \infty\}$, $z\lambda = 1$. Hence $s(\frac{1}{z}) = \frac{1}{z^n}\sigma(z)$, which shows that s has a pole of order n at ∞: $V(\Sigma(n)) = \mathcal{O}(-n)$. □

§6 Complex systems

One may ask whether the function $\mathcal{F}$ introduced in Lemma 3(xvii) induces an equivalence of categories. We show in this appendix that this is so if we allow complex coefficients. To be more precise, let $\sum_c(U, X)$ be defined as before, except that the control and state space may be complex vector spaces. Then we can define the functor $\mathcal{F} : \sum_c \to \mathcal{H}$, where $\mathcal{H}$ denotes the full subcategory of the category of holomorphic vector bundles whose objects are all negative (i.e., the Chern classes of their factors are < 0).

THEOREM 3. *$\mathcal{F}$ is an equivalence of categories.*

PROOF: It is easy to see that Lemma 3 and Theorem 2 are valid for complex coefficients. To prove Theorem 3 it is sufficient to show that for any integers n, m, $\mathcal{F}$ maps the feedback $\mathrm{Hom}(\Sigma(n),\ \Sigma(m))$, $1-1$ onto $\mathrm{Hom}(V(\Sigma(n)),\ V(\Sigma(m)))$. We leave this proof to the reader. $\square$

CHAPTER III

APPLICATIONS OF CONTROLLABILITY AND OBSERVABILITY

§0 Pole assignment

A necessary and sufficient condition for the 0 state of a linear system $\frac{dx}{dt} = Ax$ to be asymptotically stable, is that all the eigenvalues of A have negative real parts. In this chapter, we shall see that it is always possible to insure this condition in a controllable linear system. In fact we prove much more.

DEFINITION 0: A poset in $\mathbf{C}$ is a mapping $\varphi : \mathbf{C} \to \mathbf{N}$ ($\mathbf{N}$ is the set of all integers, 0 included) having a finite support. The norm, $\|\varphi\|$, of a poset φ, is $\sum\{\varphi(z) \mid z \in \mathbf{C}\}$. A poset is called conjugate-invariant if $\varphi(\overline{z}) = \varphi(z)$ for all $z \in \mathbf{C}$ ($\overline{z}$ = complex conjugate of z).

EX.: Given a real vector space X of finite dimension and a linear endomorphism A of X, the spectrum of A, Spec (A), is the poset $\varphi : \mathbf{C} \to \mathbf{N}$ defined as follows: $\varphi(z)$ is the multiplicity of z as an eigenvalue of A (if z is not an eigenvalue of A this multiplicity is 0). Clearly Spec (A) is a conjugate-invariant poset of norm $\dim X$.

DEFINITION 1: Let (B, A) be a system. We say that it has the pole assignment property if the mapping $(B_1, A_1) \mapsto \text{Spec}(A_1)$, from the orbit of (B, A) under the control feedback group into the set of all conjugate-invariant posets of norm equal to $\dim X$, is onto. In other words, for any conjugate invariant poset p of norm $\dim X$ there exists a system (B_p, A_p), control feedback equivalent to (B, A), such that $\mathrm{Spec}(A_p) = p$.

THEOREM 0. *A system (B, A) has the pole assignment property if and only if it is controllable.*

PROOF: First assume that (B, A) is a $\Sigma(n)$. A poset φ defines a unique unitary polynomial $P_\varphi[T]$ as follows: if $z \in \mathbf{C}$, z is a zero of $P_\varphi[T]$ with multiplicity $\varphi(z)$. The degree of P_φ is the norm of φ. If φ is conjugate invariant, P_φ has real coefficients.

If $\|\varphi\| = n$, then $P_\varphi[T] = T^n - \sum_{j=0}^{n-1} k_{n-j} T^j$.

Let $k : \mathbf{R}^n \to \mathbf{R}$ be the linear form $k(x) = \sum_{j=1}^{n} k_j x_j$ and define the endomorphism $A_k : \mathbf{R}^n \to \mathbf{R}^n$, $A_k = A + e_1 \otimes k$. Then (B, A_k) is feedback equivalent to $\Sigma(n)$ and its Hamilton-Cayley polynomial is $P_\varphi[T]$.

In the general case, since (B, A) is feedback equivalent to $\bigoplus_{i=1}^{r} (\Sigma(\kappa(i)))^{\beta(i)}$, without loss of generality we can assume that $\mathrm{Ker}\, B = 0$. To prove the "if" part of Theorem 0, it is sufficient to do it for such a system. Call $e(i,j)$ the control vector of the j^{th} copy of $\Sigma(\kappa(i))$, $1 \le j \le \beta(i)$. The set of vectors $\{A^k e(i,j) | 1 \le i \le r, 1 \le j \le \beta(i), 0 \le k \le \kappa(i) - 1\}$ is a basis of the state space. Define a linear mapping $K : X \to U$ as follows: identify U with the image of B; it admits the set $\{e(i,0) \mid 1 \le i \le r\}$

as basis. Then:

$$K(A^k(e(i,j))) = \begin{cases} 0 & \text{if } k < \kappa(i) - 1 \\ e(i, j+1) & \text{if } k = \kappa(i) - 1 \text{ and } j < \beta(i) \\ e(i+1, 1) & \text{if } k = \kappa(i) - 1 \text{ and } j = \beta(i) \end{cases}.$$

The couple (B, A_K) is feedback equivalent to (B, A). The system (B_1, A_K) with control space $\mathbf{R}$ and control mapping $B_1(1) = e(1,1)$, is isomorphic to $\Sigma(n)$, $n = \sum_{i=1}^{r} \kappa(i)\beta(i)$, the dimension of the state space of (B, A). We can apply what has already been proved to (B_1, A_K). This completes the proof of the "if" part of Theorem 0.

Assume now that (B, A) is not controllable. Then $S = \sum_{n=0}^{d-1} A^n(\operatorname{Im} B) \neq X$, $(d = \dim X)$. Let V be the annihilator of this space S in the complex dual $X'_{\mathbf{C}}$ of X. Since S is A-stable, so is V under the action of A'. Hence V contains at least one eigenvector p of A'. Let λ be its eigenvalue: $A'p = \lambda p$. Since V annihilates Im B, $B'p = 0$. Let (B_1, A_1) be feedback equivalent to (B, A). Then $A_1 = \overline{\alpha}^1 A\alpha + \overline{\alpha}^1 B\beta_1$, $B_1 = \overline{\alpha}^1 B\beta_2$. If $q = \alpha'(p)$, then $A_1' q = \lambda q$ since $B'p = 0$. This shows that λ is an eigenvalue for any system feedback equivalent to (B, A). Hence (B, A) does not have the pole assignment property. □

§1 Stabilization

Given a linear system (B, A), the actual evolution of the state of the system is not governed by the equation $\frac{dx}{dt}(t) = Ax(t) + Bu(t)$ but more realistically by

$$\frac{dx}{dt}(t) = Ax(t) + Bu(t) + w(t),$$

where w represents an unknown perturbation. An important practical problem is to insure that perturbations do not blow up the system, damage it or simply throw the system off a prescribed path.

Presently we shall show that if (B, A) is controllable, we can prevent the occurrence of such an unwanted phenomenon by using a feedback device, at least if we know that the perturbations are bounded.

THEOREM 1. *Assume (B, A) is controllable. For any number $\alpha > 0$, there exist a system (B, A_α) feedback equivalent to (B, A), ($A_\alpha = A_{K_\alpha} = A + BK_\alpha, K_\alpha : X \to U$) and an Euclidean norm $\| \ \|_\alpha$ on X such that if (x, u), $(\tilde{x}, u)$ are respectively the solutions of:*

$$\begin{cases} \frac{dx}{dt}(t) = A_\alpha x(t) + Bu(t) \\ x(0) = x_0 \end{cases}, \quad \begin{cases} \frac{d\tilde{x}}{dt}(t) = A_\alpha \tilde{x}(t) + Bu(t) + w(t), \\ \tilde{x}(0) = \tilde{x}_0 \end{cases}$$

both defined on $[0, +\infty)$, then, for any $T > 0$,

$$\begin{aligned} \sup_{t \geq 2T} \|x(t) - \tilde{x}(t)\|_\alpha &\leq \frac{1}{\alpha} \sup_{t \geq T} \|w(t)\|_\alpha \\ &\quad + \frac{e^{-\alpha T}}{\alpha} \sup_{0 \leq t \leq T} \|w(t)\|_\alpha + e^{-2\alpha T} \|x_0 - \tilde{x}_0\|_\alpha . \end{aligned}$$

Actually, $\left\| e^{tA_\alpha}(v) \right\|_\alpha \leq e^{-t\alpha} \|v\|_\alpha$ for all $t \geq 0$ and $v \in X$.

PROOF: Choose K_α in such a way that all the eigenvalues of $A_{K_\alpha} = A_\alpha$ have their real parts smaller than $-\alpha$. Define a scalar product $\langle \ , \ \rangle_\alpha$ on X as follows: let $\langle \ , \ \rangle$ be any scalar product and set $\langle x', x'' \rangle_\alpha = \int_0^\infty e^{2\alpha s} \left\langle e^{sA_\alpha}(x'), e^{sA_\alpha}(x'') \right\rangle ds$. Then $\left\| e^{tA_\alpha}(x') \right\|_\alpha \leq e^{-\alpha t} \|x'\|_\alpha$ if $t \geq 0$. Set $z(t) = x(t) - \tilde{x}(t)$; since $\frac{dz}{dt}(t) = A_\alpha z(t) - w(t)$ and $z(t) = e^{tA_\alpha}(x_0 - \tilde{x}_0) - \int_0^t e^{(t-s)A_\alpha} w(s) ds$, it follows that $\sup\limits_{t \geq 2T} \|z(t)\|_\alpha \leq e^{-2\alpha T} \|x_0 - \tilde{x}_0\|_\alpha + S$, where $S = \sup\limits_{t \geq 2T} \int_0^t \left\|e^{(t-s)A_\alpha} w(s)\right\|_\alpha ds \leq \sup\limits_{t \geq 2T} \int_0^T \left\|e^{(t-s)A_\alpha} w(s)\right\|_\alpha ds + \sup\limits_{t \geq 2T} \int_T^t \left\|e^{(t-s)A_\alpha} w(s)\right\|_\alpha ds \leq \frac{e^{-T\alpha}}{\alpha} \sup\limits_{0 \leq t \leq T} \|w(t)\|_\alpha + \frac{1}{\alpha} \sup\limits_{t \geq T} \|w(t)\|_\alpha$.

□

Let us give a few examples illustrating the use of this Theorem.

EX. 1: Balancing a pointer on the tip of your finger. In order to simplify the problem we assume that:

1) your finger moves on a horizontal axis, taken to be the x-axis;

2) the pointer is a homogeneous cylindrical rod, of negligible diameter, having mass m and length L;

3) the pointer moves in a vertical plane P containing the x-axis. We take a vertical line oriented upwards as the y-axis in P.

The control is the movement of your finger: let ξ be its abscissa, on the x-axis. The position of the rod is given if we specify the coordinates x_G, y_G of its center of gravity and the angle φ of the pointer with the x-axis.

There are two constraints on the rod, expressing the fact that the lower end of the rod lies on your finger:

$$x_G = \xi + \frac{L}{2} \cos \varphi,$$
$$y_G = \frac{L}{2} \sin \varphi.$$

The kinetic energy T of the rod is:

$$T = \frac{1}{2} m(\dot{x}_G^2 + \dot{y}_G^2) + \frac{mL^2}{24} \dot{\varphi}^2.$$

There are two forces acting on the rod: its weight and the reaction of your finger on the rod. Since the constraints are holonomic we can eliminate x_G, y_G and the reaction in the Lagrange equation:

$$T = \frac{1}{2}m\dot{\xi}^2 + \frac{mL^2}{6}\dot{\varphi}^2 - \frac{mL\sin\varphi}{2}\dot{\varphi}\dot{\xi}.$$

Since $\frac{d}{dt}(\frac{\partial T}{\partial \dot{\varphi}}) - \frac{\partial T}{\partial \varphi} = -\frac{mgL\cos\varphi}{2}$ is the work of the weight mg when φ varies, we have

$$\text{(E)} \qquad \frac{2L}{3}\ddot{\varphi} = -g\cos\varphi + \sin\varphi\ddot{\xi},$$

where $\ddot{\xi}$, the acceleration of your finger, is your control.

The vertical position of the pointer is an equilibrium position of the rod, albeit an unstable one. Let us try to stabilize it using a feedback. For this we linearize (E) around the equilibrium position $\varphi = \frac{\pi}{2}$. Let $\psi = \frac{\pi}{2} - \varphi$; then

$$\text{(lin E)} \qquad \frac{2L}{3}\ddot{\psi} = g\psi - \ddot{\xi}.$$

The equivalent linear system $\frac{dx}{dt} = Ax + Bu$ is: $x = (\psi, \dot{\psi}) \in X = \mathbf{R}^2$, $U = \mathbf{R}$, $u = \ddot{\xi}$,

$$A = \begin{bmatrix} 0 & 1 \\ \frac{3g}{2L} & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ -\frac{3}{2L} \end{bmatrix}.$$

The pair (A, B) is obviously controllable.

Let $e_1 = \begin{bmatrix} 0 \\ -\frac{3}{2L} \end{bmatrix}$, $e_2 = \begin{bmatrix} -\frac{3}{2L} \\ 0 \end{bmatrix}$. Then $B(1) = e_1$, $Ae_1 = e_2$.

If we define the feedback $K_0 : \mathbf{R}^2 \to \mathbf{R}$ by $K_0e_1 = 0$, $K_0e_2 = \frac{3g}{2L}$, then $(A_0 = A + BK_0, B)$ is in normal form with respect to the basis (e_1, e_2).

Assume we want $\{-1\}$ as spectrum, with multiplicity 2. The corresponding feedback is $K_1 : K_1 e_1 = -2$, $K_1 e_2 = -1$. Hence the total feedback is $\overline{K} : \overline{K} e_1 = -2$, $\overline{K} e_2 = \frac{3g}{2L} - 1$. The corresponding control $\overline{u}$ is $\overline{u}(\varphi) = -\frac{4L}{3}\dot{\varphi} + (\frac{2L}{3} - g)(\frac{\pi}{2} - \varphi)$; notice that it depends on the angular velocity $\dot{\varphi}$.

Does there exist a stabilizing feedback, depending on φ only? If $u^1(\varphi) = k\psi$ were such a feedback then the equation of the evolution of the pointer would be

$$\frac{2L}{3}\ddot{\psi} = (g - k)\psi,$$

but $(0, 0)$ is not asymptotically stable for this equation. Hence any stabilizing feedback has to take into account the speed with which the pointer is falling.

EX. 2: Libration point satellite: on the segment joining the Earth to the Moon, there is a point L called the libration point where the pull of the Earth on the satellite is equal to the sum of the pull of the Moon and the centrifugal force due to the rotation of the Moon around the Earth. This L is an unstable equilibrium point.

The satellite can be guided using a small reaction engine. One wants to design a feedback in order to stabilize the satellite at L. Taking a planar system of cartesian coordinates attached to the Moon, the linearized equations of motion of the satellite around the libration point are:

$$\text{(L)} \qquad \begin{cases} \ddot{x} - 2\omega\dot{y} - 9\omega^2 x = 0 \\ \ddot{y} + 2\omega\dot{x} + 4\omega^2 y = u \end{cases},$$

where $\omega = \frac{2\pi}{29}$ radians per day, $u = \frac{Th}{m\omega^2}$, Th is the thrust of the engine in the y direction and m is the mass of the satellite. The linear system associated to (L) is: $\frac{dz}{dt} = Az + Bu$, where the state space is $\mathbf{R}^4$, the control space $\mathbf{R}$, $z = (z_1, z_2, z_3, z_4)$: $z_1 = x$, $z_3 = y$, $z_2 = \dot{x} - 2\omega y$, $z_4 = \dot{y} + 2\omega x$,

$$B(1) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \qquad A = \begin{bmatrix} 0 & 1 & 2\omega & 0 \\ -9\omega^2 & 0 & 0 & 0 \\ -2\omega & 0 & 0 & 1 \\ 0 & 0 & -4\omega^2 & 0 \end{bmatrix}.$$

The characteristic polynomial of A is $\lambda^4 - \omega^2\lambda^2 + 36\omega^4$.

The pair (A, B) is controllable: $A(B(1)) = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$,

$$A^2(B(1)) = \begin{bmatrix} 2\omega \\ 0 \\ 0 \\ -4\omega^2 \end{bmatrix}, \qquad A^3(B(1)) = \begin{bmatrix} 0 \\ 18\omega^3 \\ -8\omega^2 \\ 0 \end{bmatrix}.$$

To compute a stabilizing feedback we shall use a method due to Bass and Gina. Since this method is applicable in general, we assume that $A \in \mathrm{End}(\mathbf{R}^d)$ and $B \in \mathrm{Lin}(\mathbf{R}, \mathbf{R}^d)$. Let $P(\lambda) = \det(\lambda I - A)$ be the characteristic polynomial of A and $P_K(\lambda) = \det[\lambda I - A - BK]$ that of a feedback $K : \mathbf{R}^4 \to \mathbf{R}$; then $P_K(\lambda) = \det[(\lambda I - A)[I + (\lambda I - A)^{-1}BK]]$. K is a linear form, the range of B being one dimensional. Hence:

$$P_K(\lambda) = P(\lambda)(1 + K)((\lambda I - A)^{-1}B(1)).$$

Let $P(\lambda) = \lambda^d + \sum_{j=0}^{d-1} a_{d-j}\lambda^j$, $a_{d-j} \in \mathbf{R}$. It is easy to show that $(\lambda I - A)^{-1} = \frac{1}{P(\lambda)}Q(\lambda)$, where

$$Q(\lambda) = \lambda^{d-1}I + \sum_{j=0}^{d-2} Q_{d-1-j}\lambda^j, \quad Q_{d-1-j} \in \mathrm{End}(\mathbf{R}^d).$$

The operators Q_{d-1-j} can be computed using the relation $P(\lambda)I = (\lambda I - A)Q(\lambda)$: $Q_0 = I$, $Q_{d-1-j} = A^{d-1-j} + \sum_{r=0}^{d-j-2} a_{d-j-1-r}A^r$.

Hence

$$P_K(\lambda) = P(\lambda) + \sum_{j=0}^{d-1} \left[KA^{d-1-j}B(1) + \sum_{r=0}^{d-j-2} a_{d-j-1-r} KA^r B(1) \right] \lambda^j$$

with the convention: $a_n = 0$ if $n \leq 0$.

If the system (A, B) is controllable, the vectors $\{A^r B(1) \mid 0 \leq r \leq d-1\}$ form a basis of $\mathbf{R}^d$ and K is determined by the numbers $\{KA^r B(1) \mid 0 \leq r \leq d-1\}$. Then if $P_K(\lambda) = \lambda^d + \sum_{j=0}^{d-1} a_{d-j}^K \lambda^j$ is given, K can be computed using the recurrence relations:

$$KB(1) = a_1^K - a_1$$

and, for $j < d-1$,

$$KA^{d-1-j}B(1) = a_{d-j}^K - a_{d-j} - \sum_{r=0}^{d-j-2} a_{d-j-1-r} KA^r B(1).$$

Let us go back to the libration point problem. We shall stabilize the satellite using a feedback K such that Spec $(A + BK) = \{-\omega, -\omega, -(1+\sqrt{-1})\omega, -(1-\sqrt{-1})\omega\}$:

$$P_K(\lambda) = \lambda^4 + 4\omega\lambda^3 + 7\omega^2\lambda^2 + 6\omega^3\lambda + 2\omega^4,$$

$$P(\lambda) = \lambda^4 - \omega^2\lambda^2 + 36\omega^4,$$

$a_1^K = 4\omega$, $a_1 = 0$, $a_2^K = 7\omega^2$, $a_2 = -\omega^2$, $a_3^K = 6\omega^3$, $a_3 = 0$, $a_4^K = 2\omega^4$, $a_4 = 36\omega^4$, $KB(1) = 4\omega$, $KAB(1) = 8\omega^2$, $KA^2B(1) = 6\omega^3 + \omega^2 4\omega = 10\omega^3$, $KA^3B(1) = -34\omega^4 + 8\omega^4 = -26\omega^4$.

§2 Observers

An observer is a device enabling us to reconstruct a trajectory of the system (B, A, C) from the input and output data.

DEFINITION 2: Given a system (B, A, C), an observer is a functional $\Psi : \mathbf{R}_+ \times L^1([0, \alpha]; U) \times AC([0, \alpha]; Y) \to X$ (α depending on Ψ) such that for any trajectory $(x, u) : [a, b] \to X \times U$ of (B, A, C) with output $y = Cx$ and $b - a > \alpha$, we have $x(b) = \Psi(b, u_\alpha, y_\alpha)$, where $(u_\alpha, y_\alpha) : [0, \alpha] \to U \times Y$ is given by $u_\alpha(t) = u(b - \alpha + t)$, $y_\alpha(t) = y(b - \alpha + t)$.

The main theorem about observers is the following:

THEOREM 2. *A system (B, A, C) has observers if and only if it is observable.*

PROOF: Assume that the system has an observer ψ but is not observable. Then one can find two trajectories $(x_j, u) : [a, b] \to X \times U$, $j = 1, 2$, $b - a > \alpha$, having the same output y and such that $x_1(b) \neq x_2(b)$. On the other hand $x_1(b) = \Psi(b, u_\alpha, y_\alpha) = x_2(b)$, a contradiction.

If the system is observable, endow X and Y with scalar products $\langle\ ,\ \rangle_X$ and $\langle\ ,\ \rangle_Y$. For any $\alpha > 0$, the endomorphism $L(a) = \int_0^\alpha e^{-tA^*} C^* C e^{-tA} dt$ is positive definite by Corollary 2-(ix), §5, Chapter I.

Given any trajectory $(x, u) : [a, b] \to X \times U$ with $b - a > \alpha$ and output $y = Cx$, we have

$$y(t) = Ce^{(t-b)A} x(b) + C \int_b^t e^{(t-s)A} Bu(s) ds;$$

therefore

$$\int_{b-\alpha}^{b} e^{(t-b)A^*}C^*y(t)dt = \int_{b-\alpha}^{b} e^{(t-b)A^*}C^*Ce^{(t-b)A}x(b)dt$$
$$+ \int_{b-\alpha}^{b} e^{(t-b)A^*}C^*C\int_{b}^{t} e^{(t-s)A}Bu(s)dsdt$$

and

$$x(b) = L(\alpha)^{-1}\left[\int_{b-\alpha}^{b} e^{(t-b)A^*}C^*y(t)dt\right.$$
$$\left. - \int_{b-\alpha}^{b}\int_{b}^{t} e^{(t-b)A^*}C^*Ce^{(t-s)A}Bu(s)dsdt\right]$$
$$= L(\alpha)^{-1}\left[\int_{b-\alpha}^{b} e^{(t-b)A^*}C^*y(t)dt\right.$$
$$\left. + \int_{b-\alpha}^{b}(W(\alpha) - W(b-s))e^{(b-s)A}Bu(s)ds\right].$$

where $W(t) = \int_0^t e^{-\sigma A^*}C^*Ce^{-\sigma A}d\sigma$. Hence

$$\Psi(T, z, v) = L(\alpha)^{-1}\left[\int_{0}^{\alpha} e^{(t-\alpha)A^*}C^*z(t)dt\right.$$
$$\left. + \int_{0}^{\alpha}(W(\alpha) - W(t))e^{tA}Bv(s)ds\right].$$

□

This exact method of reconstructing the trajectory from the input and output data is not always the most convenient to implement. In the next paragraph we give an approximate (asymptotic) method.

§3 Asymptotic observers

Given a system $\Sigma = (U, X.Y, B, A, C)$, we want to construct a system $\Sigma_1 = (U \times Y, X, Y, B_1, A_1, C_1)$ such that given any trajectory $(\overline{x}, \overline{u}) : [a, +\infty) \to X \times U$ of the first system with output $\overline{y} = C\overline{x}$, any trajectory $(z, \overline{u}, \overline{y}) : [a, +\infty) \to X \times U \times Y$ of the second system has the property: $\lim_{t \to +\infty} \|\overline{x}(t) - z(t)\| = 0$ ($\| \ \|$ = some norm on X). The system Σ_1 is called an asymptotic observer of Σ.

It is easy to see that such a system $\sum_1$ will exist only if $\sum$ is observable. Under this obvious condition, it can always be found:

THEOREM 3. *Given any observable system* (U, X, Y, B, A, C) *and any number* $\alpha > 0$, *there exist a system* $(U \times Y, X, Y, B_\alpha, A_\alpha, C_\alpha)$ *and a norm* $\| \ \|_\alpha$ *on* X *such that for any trajectory* $(\overline{x}, \overline{u}) : [a, +\infty) \to X \times U$ *of the first system with output* $\overline{y} = C\overline{x}$, *any trajectory* $(z, \overline{u}, \overline{y}) : [a, +\infty) \to X \times U \times Y$ *of the second system has the property:*

$$\limsup_{t \to +\infty} e^{\alpha t} \|\overline{x}(t) - z(t)\|_\alpha < +\infty.$$

PROOF: Since (A, C) is observable, (C', A') is controllable and hence has the pole assignment property (see Theorem 0 and Chapter I). Applying Theorem 1 to the pair (C', A'), we can find a feedback $L'_\alpha : X' \to Y'$ dual of $L_\alpha : Y \to X$, and a scalar product $\langle \ , \ \rangle'_\alpha$ on X' such that $A'_\alpha = A' + L'_\alpha C'$ has the property: $\left\| e^{tA'_\alpha}(v') \right\|'_\alpha \le e^{-t\alpha} \|v'\|'_\alpha$ for all $t \ge 0$ and all $v' \in X'$.

$\langle \ , \ \rangle'_\alpha$ induces a dual scalar product $\langle \ , \ \rangle_\alpha$ on X, dual space of X'. If $A_\alpha = A + L_\alpha C$, for all $t \ge 0$ and $v \in X$ we have $\left\| e^{tA_\alpha}(v) \right\|_\alpha \le e^{-t\alpha} \|v\|_\alpha$. This can be seen as follows: for any $v \in X$, $\|v\|_\alpha =$

$\sup\left\{\frac{v'(v)}{\|v'\|'_\alpha} \mid v' \in X' - \{0\}\right\}$, hence

$$\left\|e^{tA_\alpha}(v)\right\|_\alpha = \sup_{v'} \frac{v'(e^{tA_\alpha}(v))}{\|v'\|'_\alpha} = \sup_{v'} \frac{e^{tA'_\alpha}v'(v)}{\|v'\|'_\alpha}$$

$$\leq \|v\|_\alpha \sup_{v'} \frac{\left\|e^{tA'_\alpha}(v')\right\|'_\alpha}{\|v'\|'_\alpha} \leq e^{-t\alpha}\|v\|_\alpha.$$

Now define the observer system as follows: $A_\alpha = A + L_\alpha C$, $B_\alpha : U \oplus Y \to X$, $B_\alpha(u, y) = Bu - L_\alpha y$, $C_\alpha = C$. Then if $(z, \overline{u}, \overline{y}) : [a, +\infty) \to X \times U \times Y$ is a trajectory of this new system,

$$\frac{dz(t)}{dt} = A_\alpha z(t) + B\overline{u}(t) - L_\alpha \overline{y}(t).$$

But

$$\frac{d\overline{x}(t)}{dt} = A\overline{x}(t) + B\overline{u}(t) = A_\alpha \overline{x}(t) - L_\alpha \overline{y}(t) + B\overline{u}(t).$$

hence

$$\frac{d(z(t) - \overline{x}(t))}{dt} = A_\alpha(z(t) - \overline{x}(t))$$

and

$$z(t) - \overline{x}(t) = e^{(t-a)A_\alpha}(z(a) - \overline{x}(a)).$$

Using the above inequality,

$$e^{t\alpha}\|z(t) - \overline{x}(t)\|_\alpha \leq \left\|e^{-aA_\alpha}(z(a) - \overline{x}(a))\right\|_\alpha.$$

□

Ivan Kupka

CHAPTER IV

NONLINEAR SYSTEMS

§0 Introduction

In practice most systems are represented by nonlinear mathematical models. The study of these models is much harder than that of linear systems and is far from complete. This may be the main reason why nonlinear systems have been little used in applications, up to now. Most practitioners prefer to approximate the nonlinear model by a linear one. The fact that there are powerful and fast computers available, makes this procedure realistic and efficient.

Nonetheless it is of interest to study nonlinear systems because of their theoretical importance. Since we want to include the linear systems among the more general "nonlinear" ones, it is better to talk about general systems instead of nonlinear ones.

§1 What is a general system?

As with linear systems, we have input, output and state spaces. The state space is usually a connected smooth (C^∞ or C^ω) real manifold X, the output space another smooth manifold Y, usually a vector space, and the input space is a subspace U of a smooth manifold MU.

The dynamics of the system are modelled by a parametrized vector field $F : X \times U \to TX$ (the tangent space of X), $F(x,u) \in T_xX$ for all $(x,u) \in X \times U$, with parameter space U. F is assumed to be smooth in the following sense: it has a smooth extension $\tilde{F} : X \times MU \to TX$.

Finally, the output mapping is a smooth mapping $h : X \to Y$. Let us sum this up, in a formal definition.

DEFINITION 0: A general system $\sum$ is a quintuple (U, X, Y, F, h) such that:

1) X and Y are real smooth manifolds, and U is a subspace of a smooth manifold MU;

2) $F : X \times U \to TX$ is a vector field on X parametrized by U, having a smooth extension to $X \times MU$;

3) $h : X \to Y$ is a smooth mapping.

REMARK 0: We could have assumed that U is just a topological space and that F is continuous in both x and u and smooth in x. For technical reasons we feel that the more involved definition above is easier to handle.

The next thing is to define trajectories.

DEFINITION 1: A trajectory of Σ is a pair $(x,u) : [a,b] \to X \times U$ (where eventually $b = +\infty$ and then $[a,b] = [a,+\infty)$) such that:

(i) x is absolutely continuous, u is measurable;

(ii) for almost every $t \in [a,b]$, $\frac{dx}{dt}(t) = F(x(t), u(t))$;

(iii) if $(x', u) : [a', b'] \to X \times U$, $[a', b'] \subset [a,b]$, satisfies (i) and (ii), and if $x'(t') = x(t')$ for some $t' \in [a', b']$, then $x'(t) = x(t)$ for all $t \in [a', b']$.

In practice it happens very often that there are restrictions on the controls in the sense that $u : [a,b] \to U$ must belong to a family $\mathcal{U}$ of admissible functions $u : [a_u, b_u] \to U$, where a_u, b_u are real numbers depending on u, or else $b_u = +\infty$ and $[a_u, b_u]$ means $[a_u, +\infty)$. A very commonly used family $\mathcal{U}$ is $PC(U)$, the set of all functions $u : [a_u, b_u] \to U$ that are piecewise constant: there exists a partition $a_u = t_0 < t_1 < \cdots < t_N < b_u = t_{N+1}$ (or $a_u = t_0 < t_1 < t_2 < \cdots < t_n \to +\infty$ if $b_u = +\infty$) such that u is constant on each interval $[t_k, t_{k+1})$, $k = 0, \ldots, N$ (resp. $k \in \mathbf{N}$).

If a family $\mathcal{U}$ of admissible controls is given, we shall always assume that it satisfies the following three conditions:

(R) if $u : [a,b] \to U$ belongs to $\mathcal{U}$ and $[a', b'] \subset [a,b]$, then the restriction $u \mid [a', b']$ belongs to $\mathcal{U}$;

(T) if $u : [a,b] \to U$ belongs to $\mathcal{U}$ and $c \in \mathbf{R}$, then $u_c : [a-c, b-c] \to U$, $u_c(t) = u(t+c)$, is in $\mathcal{U}$.

(C) if $u : [a,b] \to U$ and $v : [b,c] \to U$ belong to $\mathcal{U}$, then their concatenation u^*v belongs to $\mathcal{U}$, where

$$u^*v(t) = \begin{cases} u(t) & a \leq t < b \\ v(t) & b \leq t \leq c \end{cases}.$$

§2 Input-output mapping

The most important object associated to a system is its input-output mapping.

NOTATIONS: (iv) Given any topological space T, let $\mathrm{Mes}(T)$ denote the set of all measurable functions $u : J \to T$, where J is a finite closed interval or else is infinite but closed on the left. J depends on u.

(v) Given any smooth manifold M, denote by $AC(M)$ the set of all absolutely continuous curves $x : J \to M$, where J is as in (iv) and J depends on x.

DEFINITION 2: Let $(U, X, Y, F, h, \mathcal{U})$ be a system with a set of admissible controls $\mathcal{U}$. The input-output mapping of the system is the functional mapping $\Phi : \text{dom}(\Phi) \to AC(Y)$ defined as follows: $\text{dom}\,\Phi \subset X \times \mathcal{U}$ is the set of all pairs (x_0, u), $u : [a, b] \to U$, such that there exists a trajectory $(x, u) : [a, b] \to X \times U$ with $x(a) = x_0$ and $\Phi(x_0, u) = h \circ x \in AC(Y)$.

Given $q \in X$, its response mapping $\Phi_q : \text{dom}(\Phi_q) \to AC(Y)$ is the restriction of Φ to the set $\{u \mid u \in \mathcal{U},\ (q, u) \in \text{dom}(\Phi)\}$. Given $u \in U$, its flow $\Phi^u : \text{dom}(\Phi^u) \to AC(Y)$ is the restriction of Φ to the set $\{q \mid q \in X,\ (q, u) \in \text{dom}(\Phi)\}$.

REMARK 1: (vi) If no family $\mathcal{U}$ is given, we take $\mathcal{U} = \text{Mes}(U)$.

(vii) It can happen that $\text{dom}(\Phi_q)$ or $\text{dom}(\Phi^u)$ are empty for some $q \in X$ or some $u \in \mathcal{U}$.

§3 Examples of general systems

EX. 1: Obviously any linear system is a particular case of a general one.

EX. 2: Attitude of a satellite. A satellite travels in space and the path of its center of gravity is prescribed. We are interested in the motion of the satellite around its center of gravity. Assuming the satellite S to be a rigid body, its position in phase space is given by an orthonormal frame (e_1, e_2, e_3) rigidly linked to S and by the angular velocity ω of (e_1, e_2, e_3) with respect to itself. To simplify the equations, we choose the principal axes of inertia of S as the axes of the frame (e_1, e_2, e_3). The motion of S around its center of gravity is controlled by m small two-sided reaction engines.

The state space X of S is the product: (orthonormal frames of $\mathbf{R}^3$) $\times \mathbf{R}^3$. Since the orientation of a frame does not change during a continuous motion we can identify the state space with the product $S0(3) \times so(3)$ of the special orthogonal group by its Lie algebra. If $\alpha_1, \ldots, \alpha_m$ are the maximal thrusts of the small rockets, we take as U the m-dimensional cube $[-\alpha_1, \alpha_1] \times \cdots \times [-\alpha_m, \alpha_m]$ and as MU the space $\mathbf{R}^m$.

The motion of S is determined by the equations:

$$\frac{de_k(t)}{dt} = \omega(t) \times e_k(t), \quad k = 1, 2, 3,$$

where $\times$ denotes the cross product in $\mathbf{R}^3$ and the evolution of ω is determined by Euler's equations: $\omega = \omega_1 e_1 + \omega_2 e_2 + \omega_3 e_3$, with:

$$\frac{d\omega_1(t)}{dt} = a_{23}\omega_2(t)\omega_3(t) + \sum_{j=1}^{m} u_j(t) b_1(j),$$

$$\frac{d\omega_2(t)}{dt} = a_{31}\omega_3(t)\omega_1(t) + \sum_{j=1}^{m} u_j(t) b_2(j),$$

$$\frac{d\omega_3(t)}{dt} = a_{12}\omega_1(t)\omega_2(t) + \sum_{j=1}^{m} u_j(t) b_3(j);$$

here a_{ij} are constants depending only on the geometry of S, $b(j) = b_1(j)e_1 + b_2(j)e_2 + b_3(j)e_3$ are vectors depending on the location of the reaction engine and u_j is the thrust of the j^{th} engine; $u_j \geq 0$ if the thrust is in the direction of $b(j)$, and $u_j < 0$ otherwise.

As output we can take, for example, one of the Euler angles defining the position of the frame (e_1, e_2, e_3) with respect to a fixed one.

§4 Accessibility

Given a system $\Sigma = (U, X, F)$, the definitions of the accessibility set and the accessibility set at time T are similar to their namesakes in the linear case.

DEFINITION 3: 1) For any $x \in X$, the accessibility set of x with respect to Σ, denoted by $A(x, \Sigma)$ or simply by $A(x)$, is the set $\{z(T_u) \mid (z, u) : [0, T_u] \to X \times U$ is a trajectory of Σ and $z(0) = x\}$.

2) For any $x \in X$ and $T > 0$, the accessibility set of x with respect to Σ, at time T, denoted by $A(x, T, \Sigma)$ or simply by $A(x, T)$, is the set $\{z(T) \mid (z, u) : [0, T] \to X \times U$ is a trajectory of Σ and $z(0) = x\}$.

In the general case the problem of accessibility is much harder than in the linear case. Moreover one is lead to consider two notions of accessibility, a global and a local one. Let us discuss the global one, first. Very few general results about transitivity or about accessibility are known. Let us state a basic result in accessibility (see [S]).

DEFINITION 4: A system Σ is called weakly reversible if for any x and y in X, $y \in A(x, \Sigma)$ is equivalent to $x \in A(y, \Sigma)$.

THEOREM 0. (Sussmann [**S1**]) *Assume Σ is weakly reversible.*

1) The accessibility set $A(x,\Sigma)$ of $x \in X$ is an immersed submanifold of X (see the appendix for a definition).

2) If, moreover, Σ is analytic or Σ is locally finitely generated (see appendix), then for any $x \in X$, $T_x(A(x,\Sigma)) = Lie(\Sigma)(x)$, where $Lie(\Sigma)$ is the Lie algebra of vector fields generated by $\{F_u \mid u \in U\}$.

In the general case, it is not true that $T_x A(x,\Sigma) = \mathrm{Lie}(\Sigma)(x)$; we only have an inclusion $T_x A(x,\Sigma) \supset \mathrm{Lie}(\Sigma)(x)$. (See the counterexample below.)

The proof will be given in the appendix.

If Σ is a general system, then one can define a weakly reversible system $\tilde{\Sigma} = (\tilde{U}, X, \tilde{F})$ as follows: $\tilde{U} = U_+ \cup U_-$ is the disjoint union of two copies U_+, U_- of U and $\tilde{F}_u = \begin{cases} F_u & u \in U_+ \\ -F_u & u \in U_- \end{cases}$. It is clear that $\tilde{\Sigma}$ is weakly reversible and that for any $x \in X$, $A(x,\Sigma) \subset A(x,\tilde{\Sigma})$. But we can say more. To start with, notice that by the above Theorem, $A(x,\tilde{\Sigma})$ is an immersed submanifold of X.

PROPOSITION 0. *If Σ is analytic or if Σ is locally finitely generated, then the interior of $A(x,\Sigma)$, with respect to the intrinsic manifold topology of $A(x,\tilde{\Sigma})$, is dense in $A(x,\Sigma)$, with respect to the same topology.*

In particular, if Σ is any system such that Lie $(\Sigma)(x) = T_x X$ for all $x \in X$, then the interior of $A(x,\Sigma)$ is dense in $A(x,\Sigma)$ for all $x \in X$.

For the proof see the appendix.

COUNTEREXAMPLE: Let Σ be the system $(\mathbf{R}^2, \mathbf{R}^2, F)$, where $F_u = u_1 F_1 + u_2 F_2$ is given by $F_1 = \frac{\partial}{\partial x_1}$ and $F_2 = \alpha(x_1)\frac{\partial}{\partial x_2}$, with α C^∞ and equal to 1 for $x_1 \geq 1$ and 0 for $x_1 \leq 0$. It is easy to see that

$A(x,\Sigma) = \mathbf{R}^2$ for all $x \in \mathbf{R}^2$. On the other hand, at any point x with $x_1 < 0$, $\dim \operatorname{Lie}(\Sigma)(x) = 1$. Hence $\operatorname{Lie}(\Sigma)(x) \neq T_x A(x,\Sigma)$.

§5 Miscellany about accessibility

Let us now state a few facts.

A) Given a system Σ, if the accessibility set $A(x)$ is a neighborhood of x for all $x \in X$ then the system is transitive, i.e. $A(x) = X$ for all $x \in X$.

B) The structure of $A(x)$ is unstable under small perturbations of Σ.

C) If Σ is real analytic, $A(x)$ may well not be subanalytic.

D) If, for all $x \in X$, the positive convex cone in $T_x X$ generated by $\{F_u(x) \mid u \in U\}$ has a non empty interior then, for any $x \in X$, the boundary of $A(x)$ is a locally Lipschitz submanifold of X: locally its boundary is the graph of a Lipschitz function.

These are rather negative results. One would like to know sufficient conditions for transitivity. Several such conditions are known:

E) the Kalman condition for linear systems (see Chapter 1 §3).

F) a general condition for all systems on homogeneous spaces of semi-simple Lie groups induced by the group action (see [**JK**]).

Let us state and prove another such condition (see [**B**], [**L**]).

PROPOSITION 1. *Let $\Sigma = (U, X, F)$ be a system such that:*

(viii) $F_u = F_0 + \sum_{j=1}^{m} u_j F_j$ and U is a subset of $\mathbf{R}^m$ containing $\{\pm\varepsilon_j \mid j = 1, \ldots, m\}$, where $\varepsilon_1, \ldots, \varepsilon_m$ is a basis of $\mathbf{R}^m$.

(ix) The set of all states $x \in X$, recurrent for F_0, is dense in X.

(x) For any $x \in X$, Lie $(F_0, \ldots, F_m)(x) = T_x X$, where Lie $(F_0, \ldots, F_m)$ is the Lie algebra of vector fields generated by $F_0, \ldots, F_m$.

Then Σ is controllable: in fact for any $x, y \in X$ there exists a piecewise constant control $u_{x,y} : [0, T_{x,y}] \to \{\pm\varepsilon_j \mid j = 1, \ldots, m\}$ steering x to y.

PROOF: In the first place, taking a linear transformation on the controls and on the F_k, $1 \leq k \leq m$, we may assume that $\varepsilon_1, \ldots, \varepsilon_m$ is the canonical basis $e_1, \ldots, e_m$ of $\mathbf{R}^m$: $\varepsilon_i = \sum_{j=1}^{m} a_{ji} e_j$, $1 \leq i \leq m$. Take as new controls $\overline{u}_1, \ldots, \overline{u}_m$: $u_i = \sum_{j=1}^{m} a_{ij} \overline{u}_j$; as new control fields $\overline{F}_1, \ldots, \overline{F}_m$: $\overline{F}_i = \sum_{j=1}^{m} a_{ji} F_j$. Then $\sum_{j=1}^{m} \overline{u}_j \overline{F}_j = \sum_{k=1}^{m} u_k F_k$, and ε_j is sent to e_j, $1 \leq j \leq m$.

Denote by U_0 the subset $\{\pm e_i \mid 1 \leq i \leq m\}$ of U. I claim that it is sufficient to show that the closure $\overline{A(x, U_0)}$ of $A(x, U_0)$ is always X. This being so, condition (x) and Theorem 0 imply that the interior int $A(x, U_0)$ of $A(x, U_0)$ is dense in X for any $x \in X$. Note that the system $\Sigma_0^- = (U_0, X, -F)$ satisfies the same assumptions as $\Sigma_0 = (U_0, X, F)$. Hence for any $x \in X$, the interior int $A^-(x, U_0)$ of the accessibility set $A^-(x, U_0)$ of x with respect to Σ^-, is dense in X too. For any $x, y \in X$, take a $z \in \text{int}(A(x, U_0)) \cap \text{int}(A^-(y, U_0))$. There exists a piecewise constant control taking x into z and a piecewise constant control taking y into z along the system Σ_0^-. But it is easy to check that if $u : [0, T] \to U_0$ steers y to z along the system Σ^-, then $u' : [0, T] \to U_0$, $u'(t) = u(T - t)$ steers z to y

along Σ.

To show that $\overline{A(x,U_0)} = X$ for any $x \in X$, note that condition (x) and Theorem 0 imply that given any $y \in X$, there exists a function $f : \{1,\ldots,N\} \to \{F_0,\ldots,F_m\}$, N depending on f, and numbers $t_1,\ldots,t_N \in \mathbf{R}$, such that $y = e^{t_N f(N)} \circ \cdots \circ e^{t_1 f(1)}(x)$. Using induction on N, we may assume that we know that $y_1 = e^{t_{N-1} f(N-1)} \circ \cdots \circ e^{t_1 f(1)}(x)$ belongs to $\overline{A(x,U_0)}$. We shall discuss the possible cases separately:

Case 1: If $f(N) = F_0$ and $t_N \geq 0$, then $y \in A(y_1,U_0)$. Hence $y \in \overline{A(x,U_0)}$.

Case 2: if $f(N) = F_0$ but $t_N < 0$, we can find a sequence $\{z(n) \mid n \geq 1\}$ of recurrent points for F_0 contained in int $A(x,U_0)$ such that $z(n)$ tends to y_1. Each $z(n)$ being recurrent, so is $e^{t_N F_0}(z(n))$, hence we can find a $T_n > |t_N|$ such that the distance of $e^{t_N F_0}(z(n))$ to $e^{(T_n+t_N)F_0}(z(n))$ is less than $\frac{1}{n}$. Then $\lim e^{(T_n+t_N)F_0}(z(n)) = \lim e^{tF_0}(z(n)) = e^{tF_0}(y_1) = y$, proving that $y \in \overline{\text{int} A(x,U_0)} = \overline{A(x,U_0)}$.

Case 3: $f(N) = F_k$, $k > 0$; let σ be the sign of t_N. Then $t_N F_k = |t_N| \left[(F_0 + \sigma F_k) + (-F_0)\right]$. By the Trotter-Kato formula, $e^{t_N F_k}$ is the uniform limit, on compact subsets of X, of the sequence of diffeomorphisms $\varphi_n = \left(e^{\frac{|t|}{n}(F_0+\sigma F_k)} e^{\frac{-|t|}{n}F_0}\right)^n$ ($n \geq 1$ and $(\ )^n$ in the sense of composition). Hence $y = \lim_{n\to+\infty} \varphi_n(y_1)$ and by Case 2, $\varphi_n(y_1) \in \overline{A(x,U_0)}$. □

§6 Application to the attitude control of a satellite

Let us recall the equations given at the end of §3. X is the product $SO(3)\times\mathbf{R}^3$, U is the m dimensional cube $[-\alpha_1,\alpha_1]\times\cdots\times[-\alpha_m,\alpha_m]$, $x=(e_1,e_2,e_3,\omega)$, $\omega=\omega_1e_1+\omega_2e_2+\omega_3e_3$ and

$$\frac{de_k}{dt}=\omega\times e_k,\quad k=1,2,3,$$

$$\frac{d\omega}{dt}=Q(\omega)+\sum_{j=1}^{m}u_jb_j,$$

where $b_1,\dots,b_m$ are constant vector fields on $\mathbf{R}^3$ and

$$Q(\omega)=(a_{23}\omega_2\omega_3,a_{31}\omega_3\omega_1,a_{12}\omega_1\omega_2).$$

We have a control-affine model $F_0+\sum_{j=1}^{m}u_jF_j$. Here F_0 is the quadratic field $(\omega\times e_1,\omega\times e_2,\omega\times e_3,Q(\omega))$ and $F_1,\dots,F_m$ are the fields $b_1,\dots,b_m$.

It is a well known fact that the flow $\frac{d\omega}{dt}=Q(\omega)$ is the geodesic flow on S^3. Hence it is recurrent. The flow $\frac{de_k}{dt}=\omega\times e_k$, $k=1,2,3$, is a flow on $SO(3)$ generated by a time dependent Lie algebra element ω. Hence the field F_0 is recurrent.

Using Proposition 1, we see that if Lie $(Q,b_1,\dots,b_m)$ is of maximum rank at each point, the system is controllable (For more details see [**B**]). In fact, it is easy to see that it suffices that Lie $(Q,b_1,\dots,b_m)(0)=\mathbf{R}^3$; this can be realized even if there is only one control $(m=1)$. On the other hand, this last observation has no practical use since, to control the system, we need the recurrence of F_0. Now the return times can be very large, and we might have to wait too long to get the satellite in the wanted position. This justifies the discussion of the next paragraph.

§7 Local accessibility

Let Σ be the system (U,X,F) and let $(x,u):[0,T]\to X\times U$

be a trajectory of Σ. A property of the pair $(\Sigma, (x, u))$, very useful in practice, is the following: whenever at some time $t \in [0, T]$, the trajectory x is perturbed and the point $x(t)$ is sent to some position ξ not too far from x, there exists a small time τ and an admissible control $\tilde{u} : [t, t+\tau] \to U$ close to u, such that the trajectory $(\tilde{x}, \tilde{u}) : [t, t+\tau] \to X \times U$ starting at ξ at time τ, ends at $x(t+\tau)$.

We are going to formalize this property. For technical reasons, we define the corresponding property for the system $\Sigma^- = (U, X, -F)$.

Denote by d a metric on U, compatible with the topology of U. For any $\varepsilon > 0$, $\tau > 0$, and any control $\overline{u} : [0, \overline{T}] \to U$, $\overline{T} > \tau$, let $\mathcal{U}(\overline{u}, \varepsilon, \tau)$ denote the set of all controls $v : [0, \tau] \to U$ such that $\sup\{d(\overline{u}(t), v(t)) \mid 0 \leq t \leq \tau\} < \varepsilon$.

DEFINITION 5: Given a system $\Sigma = (U, X, F)$ and a trajectory $(\overline{x}, \overline{u}) : [0, \overline{T}] \to X \times U$ of Σ, we say that it has the local controllability property at $\overline{x}(0)$ if, given $\varepsilon > 0$, there exists a $\delta > 0$, such that for any $\tau > 0$, $0 < \tau < \delta$, the accessibility set $A(\overline{x}(0), \tau, \mathcal{U}(\overline{u}, \varepsilon, \tau))$ of $\overline{x}(0)$ at time τ, using controls from $\mathcal{U}(\overline{u}, \varepsilon, \tau)$ only, is a neighborhood of $\overline{x}(\tau)$.

A particular case of Definition 5 occurs when $\overline{x}$ is reduced to a point, that is, we have an equilibrium point for the equation $\frac{d\overline{x}}{dt}(t) = F(\overline{x}(t), \overline{u}(t))$.

In the general situation, very little is known about local controllability but in the case of control-affine systems, we have presently two results.

THEOREM 1. *Let* $\Sigma = (U, X, F)$ *be a control-affine system* $F = F_0 + \sum_{j=1}^{m} u_j F_j$. *Assume that* U *is a neighborhood of 0 in* $\mathbf{R}^m$. *Then* *a drift trajectory* $\overline{x} : [0, \overline{T}] \to X$, $\frac{d\overline{x}}{dt}(t) = F_0(\overline{x}(t))$, *has the local controllability property at* $\overline{x}(0)$ *if the following so-called ad-condition is verified: the vectors* $\{ad^k F_0(F_j)(\overline{x}(0)) \mid 1 \leq j \leq m, k \geq 1\}$ *generate* $T_{\overline{x}(0)}X$, *where* $ad^k F_0(F_j) = adF_0(ad^{k-1}F_0(F_j)), adF_0(G) = [G, F_0]$.

For the proof see the appendix.

The next result generalizes part of Theorem 1 (see [**S3**]).

THEOREM 2. *Let* $\Sigma = (U, X, F)$ *be a control-affine system with one control vector:* $F = F_0 + u_1 F_1$. *Assume that* U *is a neighborhood of 0 in* $\mathbf{R}$ *and let* x_0 *be an equilibrium point of* F_0. *Then* x_0 *is locally controllable if the following conditions are satisfied: (xi) Lie* $(F_0, F_1)(x_0) = T_{x_0}X$; *(xii) Let* S_k *be the subspaces of Lie* (F_0, F_1) *defined as follows:* $S_0 = \mathbf{R}F_0$, $S_1 =$ *linear span of* S_0 *and* $\{ad^k F_0(F_1) \mid k \geq 1\}$ *and* $S_k =$ *linear span of* S_{k-1} *and* $\{ad^k F_0 adF_1(G) \mid G \in S_{k-1}\}$. *Then for every odd* k, $S_k(x_0) = S_{k+1}(x_0)$.

Nice as they are, very seldom one can use them in practice. Let us illustrate our point returning to the attitude control of a satellity.

§8 Local controllability of the satellite

Let us take a fixed frame (e_1^0, e_2^0, e_3^0). Then $x_0 = (e_1^0, e_2^0, e_3^0, 0)$ is an equilibrium of F_0, that is, of the satellite without control action. At x_0 the "*ad*-space" is 0, hence Theorem 1 is not applicable. As for Theorem 2, assuming we have only one control vector b_1, $S_0(x_0) = 0$, $S_1(x_0) = 0$, $S_2(x_0)$ is either 0 or $\mathbf{R}F_{011}(x_0)$, where

F_{011} is the constant vector field $(0, [b_1, [b_1, Q]])$ (0 is in the $SO(3)$ part). If F_{011} is not zero, the conditions of Theorem 2 are violated.

It is not hard to see that if we have only one control vector, there is no local controllability at x_0.

APPENDIX

PROOF OF THE ORBIT THEOREM

Let $\Sigma = (U, X, F)$ be a system. Denote by $\mathcal{G}_0$ the set of all triples (X_1, φ, X_2), X_1, X_2 open subsets of X, φ diffeomorphism from X_1 onto X_2, having the following property: there exists a smooth mapping $\Phi : X_1 \times [0, \alpha] \to X$ such that:

(i) $\Phi(x, 0) = 0$ for all $x \in X_1$; (ii) $\Phi(x, \alpha) = \varphi(x)$ for all $x \in X_1$; (iii) there is a $u \in U$ such that $\frac{\partial \Phi}{\partial t}(x, t) = F(\Phi(x, t), u)$ for all $x \in X_1$ and $t \in [0, \alpha]$.

DEFINITION 0: The semi-groupoid of Σ denoted by $S\mathcal{G}(\Sigma)$ is the smallest set of triples (X_1, φ, X_2) as above, containing $\mathcal{G}_0$ and closed under the following operations:

(iv) if $X_1 \subset X$ is open, (X_1, Id_{X_1}, X_1) belongs to $S\mathcal{G}(\Sigma)$;

(v) if $(X_1, \varphi, X_2) \in S\mathcal{G}(\Sigma)$ and $X_3 \subset X_1$ is an open subset, then $(X_3, \varphi \mid X_3, \varphi(X_3)) \in S\mathcal{G}(\Sigma)$;

(vi) if $\{(X_1(j), \varphi_j, X_2(j)) \mid j \in J\}$ is a family of elements of $S\mathcal{G}(\Sigma)$ with the property that $\varphi_j \mid X_1(j) \cap X_1(k) = \varphi_k \mid X_1(j) \cap X_1(k)$ for any pair $(j, k) \in J \times J$ such that $X_1(j) \cap X_1(k) \neq \phi$, then the triple (X_1, φ, X_2), union of $\{(X_1(j), \varphi_j, X_2(j)) \mid j \in J\}$, is in $S\mathcal{G}(\Sigma) : X_1 = \bigcup_{j \in J} X_1(j)$, $X_2 = \bigcup_{k \in J} X_2(k)$, $\varphi \mid X_1(j) = \varphi_j$.

(vii) If (X_1, φ, X_2) and (X_2, ψ, X_3) belong to $S\mathcal{G}(\Sigma)$, then the composition $(X_1, \psi \circ \varphi, X_3)$ belongs to it too.

DEFINITION 1: A system Σ is called reversible if for any triple (X_1, φ, X_2) in $S\mathcal{G}(\Sigma)$ and any $x \in X_1$ there exists another triple (X_3, ψ, X_4) in $S\mathcal{G}(\Sigma)$ such that $\varphi(x) \in X_3$ and $\psi(\varphi(x)) = x$. (The 2^{nd} triple usualy depends on x).

DEFINITION 2: Given $x \in X$, the orbit $0(x)$ of x is the set of all $\varphi(x)$ with $(X_1, \varphi, X_2) \in S\mathcal{G}(\Sigma)$ such that $x \in X_1$.

NOTATION 0: (viii) We shall denote by Lie(F) the Lie algebra of vector fields generated by the family $\{F_u \mid u \in U\}$.

(ix) We shall denote by $\mathcal{L}$ the Lie algebra sheaf generated by the family $\{\varphi_*(F_u) \mid u \in U, (X_1, \varphi, X_2) \in S\mathcal{G}(\Sigma)\}$; $\varphi_*(F)$ is the vector field on X_2 defined as follows: $\varphi_*(F)(y) = d\varphi(\varphi^{-1}(y))F(\varphi^{-1}(y))$. It is clear that the elements of Lie(F) are sections of $\mathcal{L}$.

DEFINITION 3: A module $\mathcal{M}$ of vector fields on X is called locally finitely generated if for any $x \in X$ there exists an open neighborhood V of x and a finite set $F_1, \ldots, F_q$ of sections of $\mathcal{M}$ over V such that if F is any section of $\mathcal{M}$ over an open subset W of V then there exist functions $a_1, \ldots, a_q \in C^{sm}(V; \mathbf{R})$ satisfying $F = \sum_{j=1}^{q} a_j F_j$.

DEFINITION 4: A subset Y of X is called an immersed submanifold of X, if Y has a smooth manifold structure and the injection $Y \hookrightarrow X$ of Y into X is a smooth immersion.

THEOREM 0. *Let Σ be a system.*

(x) For any $x \in X$ the orbit $0(x)$ of x under $S\mathcal{G}(\Sigma)$ is equal to the accessibility set of x under Σ.

(xi) If Σ is reversible then $0(x)$ is an immersed submanifold of X for each $x \in X$.

(xii) For any $x \in X$, the tangent space of $0(x)$ at x, $T_x 0(x)$, is equal to the space $\mathcal{L}(x)$ obtained by evaluating $\mathcal{L}$ at x.

(xiii) If Lie(F) is locally finitely generated, then $T_x 0(x) =$ Lie$(F)(x)$.

(xiv) The conclusion of $(xiii)$ is always true if Σ is an analytic system.

PROOF: Is is clear that $0(x)$ is contained in $A(x, \Sigma)$, the accessibility set of x under Σ. If $(x, u) : [0, T] \to X \times U$ is any trajectory of Σ such that u is piecewise constant, then $x(t) \in 0(x(0))$ for all $t \in [0, T]$. Hence $0(x) = A(x, PC(U), \Sigma)$ the accessibility set of x using only piecewise constant controls.

The basic fact for the proof of Theorem 0 is the following: for any $x \in X$ and any $(X_1, \varphi, X_2) \in S\mathcal{G}(\Sigma)$ such that $x \in X_1$, $d\varphi(x)\mathcal{L}(x) = \mathcal{L}(\varphi(x))$.

This can be seen as follows: that $d\varphi(x)\mathcal{L}(u) \subseteq \mathcal{L}(\varphi(x))$ is a trivial consequence of the definition of $\mathcal{L}$. Since Σ is reversible, there exists a triple $(X_3, \psi, X_4) \in S\mathcal{G}(\Sigma)$ such that $X_3 \ni \varphi(x)$ and $\psi(\varphi(x)) = x$. Then $d\psi(\varphi(x))\mathcal{L}(\varphi(x)) \subseteq \mathcal{L}(x)$. Since both $d\varphi(x)$ and $d\psi(\varphi(x))$ are injective, and $\mathcal{L}(x)$, $\mathcal{L}(\varphi(x))$ are finite dimensional, we get $d\varphi(x)\mathcal{L}(x) = \mathcal{L}(\varphi(x))$.

A Σ-chart will be a smooth diffeomorphism $\varphi : M \times N \to W$, of the product of a manifold M by an open subset N of an Euclidean space $\mathbf{R}^q$ onto an open subset W of X satisfying the following conditions:

(xv) $q = \inf\{\dim \mathcal{L}(x) \mid x \in W\}$; q will be denoted by $\nu(\varphi)$;

(xvi) for any $m \in M$, $\varphi(m, N)$ is contained in an orbit of $S\mathcal{G}(\Sigma)$;

(xvii) for any $(m, t) \in M \times N$, $d\varphi(m, t)[T_t N]$ is contained in

$\mathcal{L}(\varphi(m,t))$.

It follows from (xv) and (xvii) that if $\dim \mathcal{L}(\psi(m,t)) = \nu(\varphi)$, $d\varphi(m,t)[T_tN] = \mathcal{L}(\varphi(m,t))$. In what follows, we shall denote by $pr_M : M \times N \to M$ the canonical projection and, for any m, by $\varphi_m : N \to X$ the mapping $\varphi_m(t) = \varphi(m,t)$.

Take any $x_0 \in X$ and any set $F_1, \ldots, F_q$ of vector fields belonging to $\mathcal{L}(V)$, V open neighborhood of x_0, such that $\{F_1(x_0), \ldots, F_q(x_0)\}$ is a basis of $\mathcal{L}(x_0)$. There exist open neighborhoods $W_1, W_2, \ldots, W_q, W$ of x_0, $W_1 \subseteq W_2 \subseteq W_3 \cdots \subseteq W_q \subset W$, $\varepsilon > 0$ and q smooth mappings $\varphi_j : W \times [-\varepsilon, \varepsilon] \to X$ such that:

(xviii) $\varphi_j(x, 0) = x$ for all $x \in W$;

(xix) $\frac{\partial \varphi_j}{\partial t}(x,t) = F_j(\varphi(x,t))$ for all $(x,t) \in W \times [-\varepsilon, \varepsilon]$;

(xx) $\varphi_{jt}(W_k) \subseteq W_{k+1}$ for all $t \in [-\varepsilon, \varepsilon]$ and all $1 \leq k \leq q$, where $\varphi_{jt} : W \to X$ is the mapping $\varphi_{jt}(x) = \varphi_j(x,t)$.

Then we can define a smooth mapping $\tilde{\varphi} : W_1 \times [-\varepsilon, \varepsilon]^q \to X$ as follows: $\tilde{\varphi}(x, t_1, \ldots, t_q) = \varphi_{1,t_1} \circ \varphi_{2,t_2} \circ \cdots \circ \varphi_{q,t_q}(x)$. Let $\tilde{Z}$ be a $(\dim X - q)$-dimensional submanifold of X, $\tilde{Z} \ni x_0$, transversal to $\mathcal{L}(x_0)$. Let φ be the restriction of $\tilde{\varphi}$ to $\tilde{Z} \times [-\varepsilon, \varepsilon]^q$. Since $d\tilde{\varphi}(x_0, 0)$ maps $T_{x_0}\tilde{Z} \times \mathbf{R}^q$ isomorphically onto $T_{x_0}X$, there exist a connected neighborhood $N(x_0)$ of 0 in $[-\varepsilon, \varepsilon]^q$, a neighborhood $Z(x_0)$ of x_0 in $\tilde{Z}$ such that $\varphi : Z(x_0) \times N(x_0) \to X$ is a diffeomorphism of $Z(x_0) \times N(x_0)$ onto the open neighborhood $\varphi(Z(x_0) \times N(x_0))$ of x_0. It is clear that φ is a Σ-chart with $\nu(\varphi) = q = \dim \mathcal{L}(x_0)$.

Now we define a manifold structure $M\Sigma$ on X as follows: its charts are the pairs (Y, χ_Y) of a submanifold Y of X and a diffeomorphism of Y onto an open subset in some Euclidean space $\mathbf{R}^q$ with the following property: there exist a Σ-chart $\varphi : M \times N \to X$, a point m in M, an open subset ω in N such that $\dim \mathcal{L}(\varphi(m,t)) =$

$\nu(\varphi) = q$ for all $t \in \omega$ and (Y, χ_Y) is the pair $(\varphi_m(\omega), \varphi_m^{-1})$, where $\varphi_m : N \to X$ is the mapping $\varphi_m(t) = \varphi(m, t)$.

Let us see that these charts are compatible: let (Y, χ_Y), $(Y', \chi_{Y'})$ be two charts such that $Y \cap Y' \neq \phi$. Then $(Y, \chi_Y) = (\varphi_m(\omega), \varphi_m^{-1}), (Y', \chi_{Y'}) = (\varphi'_{m'}(\omega'), \varphi'^{-1}_{m'})$ for some Σ-charts $\varphi : M \times N \to X$, $\varphi' : M' \times N' \to X$, with $\nu(\varphi) = \dim Y$, $\nu(\varphi') = \dim Y'$. Let $y \in Y \cap Y'$. $T_y Y = \operatorname{Ker} d(pr_M \circ \varphi^{-1})(y) = \mathcal{L}(y) = \operatorname{Ker} d(pr_{M'} \circ \varphi'^{-1})(y) = T_y Y'$. This shows that $pr_M \circ \varphi^{-1} \circ \varphi'_{m'}$ (resp. $pr_{M'} \circ \varphi'^{-1} \circ \varphi_m$) is constant on any connected component of $\varphi'^{-1}_{m'}(\text{Image } \varphi)$ (resp. φ_m^{-1} (Image φ')). At any $y \in Y \cap Y'$, Y and Y' define the same manifold germ. Hence the pairs (Y, χ_Y) define a manifold structure $M\Sigma$ on X and it follows immediately from condition (xvi) of the Σ-charts that Y is contained in an orbit of $S\mathcal{G}(\Sigma)$. Hence the connected components of $M\Sigma$ are contained in the orbits of $S\mathcal{G}(\Sigma)$. In fact they coincide: let $x_0 \in X$ and let $(x, u) : [0, T] \to X \times U$ be a trajectory of Σ such that $x(0) = x_0$. I claim that $x : [0, T] \to X$ is contained in a unique component of $M\Sigma$ and that it is absolutely continuous for that structure. Since $X \ni x \mapsto \dim \mathcal{L}(x)$ is lower semi-continuous, so is $[0, T] \ni t \mapsto \delta(t) = \dim \mathcal{L}(x(t))$. Let $\mathcal{O} \subseteq [0, T]$ be the set of all $t_0 \in [0, T]$ such that $\delta(t)$ is constant in a neighborhood of t_0; $\mathcal{O}$ is an open subset of $[0, T]$. It is everywhere dense: let $P = [0, T] - \mathcal{O}$ and assume that the interior P^0 of P is not empty. Let $r = \sup\{\delta(t) \mid t \in P^0\}$; if $\tau \in P^0$ and $\delta(\tau) = r$, there exists an open subset J of P^0 containing τ, such that $\delta(t) = r$ in J. But then $J \subseteq \mathcal{O}$, a contradiction.

Assume P is empty. Then $\delta(t)$ is constant on $[0, T]$. It does not restrict the generality of our reasonning to assume that the curve x is contained in a coordinate chart $\overline{\varphi}'$: Image $\varphi \to M \times N$ such that $\delta(t) = \nu(\varphi)$ for all t in $[0,T]$. For almost every $t \in [0,T]$, x is differentiable and $\frac{dx}{dt}(t) = F(x(t), u(t)) \in \mathcal{L}(x(t))$. Since $\dim \mathcal{L}(x(t)) = \nu(\varphi)$, $\mathcal{L}(x(t)) = \operatorname{Ker} d(pr_m \circ \varphi^{-1})(x(t))$. So for almost all $t \in [0,T]$, $\frac{d}{dt}(pr_m \circ \varphi \circ x)(t) = 0$. The conclusion follows.

If P is not empty, we can apply to the space P what we have done above. Let $\omega \subseteq P$ be the set of all $t_0 \in P$ such that δ restricted to a *nbd* of t_0 in P is constant. ω is an open subset of the space P and, as for $\mathcal{O}$, we can show that it is everywhere dense in P. Hence there exists an open interval J in $[0,T]$ such that δ restricted to $J \cap P$ is constant and equal to σ, say.

$J \cap \mathcal{O}$ is a disjoint union of a sequence of open intervals. On each of these intervals δ is constant. Since $J \cap P$ is not empty, there is at least one of these intervals, (a,b), on which δ is equal to ℓ and $\ell \neq \sigma$. Either a or b or both belong to P. We shall assume $b \in P$. If not, one could take the same argument as below replacing b by a.

There exists a Σ-chart $\psi : M_1 \times N_1 \to X$ such that $x(b)$ belongs to the image of ψ and $\nu(\psi) = \dim \mathcal{L}(x(b)) = \sigma$; $\psi^{-1}(x) = (m, \xi) \in M_1 \times N_1$ and $\psi_*^{-1}(F)(m, \xi, u) = (E(m, \xi, u), G(m, \xi, u))$, $E : M_1 \times N_1 \times U \to TM_1$, $G : M_1 \times N_1 \times U \to TN_1$.

There is an $\eta > 0$ such that $x([b-\eta, b+\eta])$ is contained in the image of ψ and $\psi^{-1}(x(t)) = (m(t), \xi(t))$ if $t \in [b-\eta, b+\eta]$. We have the relations: $\frac{dm}{dt}(t) = E(m(t), \xi(t), u(t))$, $\frac{d\xi}{dt}(t) = G(m(t), \xi(t), u(t))$ for almost all $t \in [b-\eta, b+\eta]$. Since $\dim \mathcal{L}(y) = \sigma$ for all y such that $pr_{M_1}(\psi^{-1}(y)) = pr_{M_1}\psi^{-1}(x(b)) = m(b)$, it follows that $F(y,u) \in \operatorname{Ker} d(pr_{M_1} \circ \psi^{-1})(y)$ for all such y's. Hence for any $(\xi, u) \in N_1 \times U$, $E(m(b), \xi, u) = 0$. By the uniqueness of solutions, $m(t) = m(b)$ for all $t \in [b-\eta, b+\eta]$. Hence

$pr_{M_1}\psi^{-1}(x(t)) = pr_M\psi^{-1}(x(b))$. Then $\dim \mathcal{L}(x(t)) = \sigma < \rho$ for $t \in [b-\eta, b+\eta]$, which contradicts the fact that if $t \in [b-\eta, b)$, $\dim \mathcal{L}(x(t)) = \delta(t) = \rho$. Our assumption that P is not empty leads to a contradiction.

Now let us prove assertion (xiii). Assume that Lie(F) is locally finitely generated. We are going to show that if (X_1, φ, X_2) is a triple from $S\mathcal{G}(\Sigma)$ and G is a section of Lie(F) over X_1, then $\varphi_*(F)$ belongs to Lie(F). It is sufficient to prove this when the triple belongs to $\mathcal{G}_0$: there exists a smooth mapping $\Phi : X_1 \times [0,\alpha] \to X$ satisfying the conditions (i) and (ii) on the beginning of this Appendix. It is easy to see that all we need to show is: for any $x_0 \in X_1$ there exist a open neighborhood V of x_0 and a number ε, $0 < \varepsilon \leq \alpha$, such that for any G in Lie(F) defined on V, $\Phi_{t*}(G)$ belongs to Lie(F) if $t \in [0,\varepsilon]$.

Choose a neighborhood $V_1 \subseteq X_1$ of x_0 such that, on V_1, Lie(F) has a finite set of generators $F_1, F_2, \ldots, F_q$. We can find an open neighborhood $V_2 \subseteq V_1$ of x_0 and a number ε, $0 < \varepsilon \leq \alpha$, such that $\Phi_t^{-1}(V_2) \subseteq V_1$ for all $t \in [0,\varepsilon]$. On V_2, $\frac{\partial}{\partial t}\Phi_{t*}(F_i) = \Phi_{t*}[F_i, F_u] = \Phi_{t*}[\sum_{j=1}^{q} c_{ji}F_j]$, where the c_{ji} are smooth functions defined on V_2, since $[F_i, F_u] \in$ Lie(F). Hence, $\frac{\partial \Phi_{t*}(F_i)}{\partial t} = \sum_{j=1}^{q} \bar{c}_{ji}\Phi_{t*}(F_j)$, where $\bar{c}_{ji}(x,t) = c_{ji} \circ \Phi_t^{-1}(x,t)$. Let $\Omega : V_2 \times [0,\varepsilon] \to \{q \times q \text{ real matrices}\}$ be the solution of the Cauchy problem: $\frac{\partial \Omega_{ij}}{\partial t} = \sum_{k=1}^{q} \Omega_{ik}\bar{c}_{ki}$, $\Omega(x,0) =$ Identity matrix. On V_2, we have $\Phi_{t*}(F_i) = \sum_{j=1}^{q} F_j\Omega_{ji,t}$, for $t \in [0,\varepsilon]$, proving assertion (xiii).

It is well known that if Σ is analytic, the module generated by $\{F_u \mid u \in U\}$ is locally finitely generated. On the other hand it is easy to prove (xiv) directly. □

COUNTEREXAMPLE: If we do not have uniqueness of solutions along the trajectory $(x, u) : [0, T] \to X \times U$, it may happen that the trajectory x is not contained in the orbit of $x(0)$ under $S\mathcal{G}(\Sigma)$: let $X = \mathbf{R}^3$, $U = \mathbf{R}^2$, $F(x, u) = u_1 F_1(x) + u_2 F_2(x)$, $F_2(x) = \frac{\partial}{\partial x_3}$, $F_1(x) = -(x_1 + x_2)\frac{\partial}{\partial x_1} + (x_1 - x_2)\frac{\partial}{\partial x_2}$. Then if $x_3 \neq 0$, $0(x)$ is two-dimensional; if $x_3 = 0$, $0(x)$ is just the x_3-axis. I claim that one can go from $(1,0,0)$ say, to $(0,0,1)$: let $\check{u} : [0,2] \to U$ be the function $\check{u}(t) = (\frac{1}{1-t}, 0)$ if $0 \le t < 1$, $\check{u}(t) = (0,1)$ if $1 \le t \le 2$. The trajectory (x, u) starting at $(1,0,0)$ is given by

$$\check{x}(t) = [(1-t)\cos(\log\frac{1}{1-t}), (1-t)\sin(\log\frac{1}{1-t}), 0], \text{ for } 0 \le t \le 1$$

$$\check{x}(t) = (0, 0, t-1), \text{ for } 1 \le t \le 2.$$

Clearly x is absolutely continuous and $\frac{d\check{x}}{dt}(t) = F(\check{x}(t), \check{u}(t))$ for $t \neq 1$.

PROOF OF THEOREM 1: For simplicity let $\varphi : V \times [-\varepsilon, \varepsilon] \to X$ be a smooth mapping such that $\frac{\partial \varphi}{\partial t} = F_0(\varphi)$, $\varphi(x, 0) = x$. Let $\varphi_t : V \to X$ be the submapping $\varphi_t(x) = \varphi(x, t)$. To prove Theorem 1 we apply the transformation $x(t) = \varphi(z(t), t)$ to the equation $\frac{dx}{dt} = F_0(x) + \sum_{j=1}^{m} u_j F_j(x)$, which becomes $\frac{dz}{dt}(t) = \sum_{j=1}^{m} u_j(t) G_j(t, z(t))$, where G_j is the time dependent field $G_j(t, z) = \varphi_{t^*}(F_j)(z)$. Denote by z_u the solution of this last equation satisfying the initial condition $z_u(0) = \overline{x}(0)$.

All we need to prove is that the mapping $\mathcal{U}(0,\varepsilon,\tau) \ni u \mapsto z_u(\tau)$ covers a neighborhood of $\overline{x}(0)$. Choose a coordinate neighborhood W of $\overline{x}(0)$ contained in V. In W, $z_u(\tau) - \overline{x}(0)$ is given by:

$$\int_0^\tau \sum_{j=1}^m u_j(t)G_j(t,\overline{x}(0))dt + \int_0^\tau \int_0^t \sum_{j,\ell=1}^m W_{j\ell}(t,s)u_j(t)u_\ell(s)dsdt.$$

where $W_{j\ell}(t,s) = dG_j(t,z_u(s))[G_\ell(s,z_u(s))]$. To prove the above assertion, it is sufficient to prove that the linear mapping $L_\tau : L^1([0,\tau];\mathbf{R}^m) \to \mathbf{R}^d$, defined by

$$L_\tau(u) = \int_0^\tau \sum_{j=1}^m u_j(t)G_j(t,\overline{x}(0))dt,$$

is onto.

If it were not, one could find a $p_\tau \in (\mathbf{R}^d)'$, (dual of $\mathbf{R}^d$), $p_\tau \neq 0$, such that for all u, $\int_0^\tau \sum_{j=1}^m u_j(t)\langle p_\tau, G_j(t,\overline{x}(0))\rangle\, dt = 0$. This implies that $\langle p_\tau, G_j(t,\overline{x}(0))\rangle = 0$ for all $t \in [0,\tau]$ and all j, $1 \leq j \leq m$. Taking a derivative and evaluating at $t = 0$, we get $\langle p_\tau, ad^k F_0(F_j)(\overline{x}(0))\rangle = 0$ for all $k \in \mathbf{N}$ and all j, $1 \leq j \leq m$, contradicting the hypothesis of Theorem 1. □

CHAPTER V

OPTIMAL CONTROL THEORY

§0 Introduction to the problems of optimal control theory

DEFINITION 0: Given a system $\Sigma = (U, X, F)$, a cost function for Σ is a smooth function $c : X \times U \to \mathbf{R}$.

DEFINITION 1: Given a trajectory $(x, u) : [a, b] \to X \times U$, its cost $\mathcal{C}(x, u)$ is defined as follows: let $c_+ = \sup(c, 0)$, $c_- = \sup(-c, 0)$; both integrals $\int_a^b c_+(x(t), u(t))dt$ and $\int_a^b c_-(x(t), u(t))dt$ are defined in the extended sense and take their values in $[0, +\infty]$, $+\infty$ included. Then: $\mathcal{C}(x, u) = \int_a^b c_+(x(t), u(t))dt - \int_a^b c_-(x(t), u(t))dt$, with the conventions:

$$\begin{cases} +\infty - \alpha = +\infty, & \\ \alpha - (+\infty) = -\infty, & \alpha \in \mathbf{R}_+ \\ +\infty - (+\infty) = +\infty. & \end{cases}$$

An optimal control problem is the following setup: 1) a system $\Sigma = (U, X, F)$ with a cost function c, 2) two subsets A and B of X, 3) a class $\mathcal{U} \subset \mathrm{Mes}(U)$ of controls; and, in the case of a fixed time problem, 4) a positive number T.

NOTATIONS: $Tr(A, B, \mathcal{U}) = \{(x, u) : [0, T_u] \to X \times U$ trajectory of Σ, $u \in \mathcal{U}$, $x(0) \in A$, $x(T_u) \in B\}$. $Tr(A, B, \mathcal{U}, T) = \{(x, u) : [0, T_u] \to X \times U$, $(x, u) \in Tr(A, B, \mathcal{U})$ and $T_u = T\}$.

OPTIMAL CONTROL PROBLEM: Find a trajectory $(\overline{x}, \overline{u}) : [0, \overline{T}] \to X \times U$ in $Tr(A, B, \mathcal{U})$ such that

$$\mathcal{C}(\overline{x}, \overline{u}) \leq \mathcal{C}(x, u) \quad \text{for all} \quad (x, u) \in Tr(A, B, \mathcal{U}).$$

OPTIMAL CONTROL PROBLEM WITH FIXED-TIME T: Same statement replacing $Tr(A, B, \mathcal{U})$ by $Tr(A, B, \mathcal{U}, T)$ and $\overline{T}$ by T.

DEFINITION 2: The curve $(\overline{x}, \overline{u})$, if it exists, is called an optimal trajectory.

§1 Some examples

EX. 1: *The classical problem of calculus of variation.* One is given an open subset $\mathcal{O}$ of some space $\mathbf{R}^d$, a smooth function $L : \mathcal{O} \times \mathbf{R}^d \to \mathbf{R}$, two points α and β in $\mathcal{O}$. One wants to find, among the absolutely continuous curves $\overline{x} : [0, \overline{T}] \to \mathcal{O}$ satisfying: i) $\overline{x}(0) = \alpha$, $\overline{x}(\overline{T}) = \beta$; ii) the function $t \mapsto L(\overline{x}(t), \frac{d\overline{x}}{dt}(t))$ is integrable on $[0, \overline{T}]$, one such that $\int_0^{\overline{T}} L(\overline{x}(t), \frac{d\overline{x}}{dt}(t))dt$ is minimal.

The optimal control setup for this problem is as follows: $U = \mathbf{R}^d$, $X = \mathcal{O}$, $F(x, u) = u$, $A = \{\alpha\}$, $B = \{\beta\}$, $c = L$.

EX. 2: *Accelerated vehicle problem.* A particule moves along a prescribed smooth curve Γ, but we can control its acceleration. If we choose an origin 0_Γ and an orientation on Γ and denote by $x(t)$ the curvilinear abscissa of the particle at time t, the equation of motion is: $m\frac{d^2x}{dt^2} = u$, where m is the mass of the particle, u its acceleration. The origin 0_Γ and another point E on Γ are given. One wants to determine u in such a way that the time needed for the particle to go from 0_Γ to E is minimal, assuming the particle starts at 0_Γ from the rest position and arrives at E with zero speed.

The setup for this problem is as follows: $U = [-a, b]$, where b is the maximum acceleration, a the maximum braking force, $X = \mathbf{R}^2$, $\frac{dx_1}{dt} = x_2$, $x_2 = \frac{u}{m}$, $F(x, u) = (x_2, \frac{u}{m})$, c is the constant function 1, A is the point $(0, 0)$, B the point $(L, 0)$, with L the abscissa of E.

In case the particle is really a vehicle, the stability of the vehicle during sharp turns can cause some anxiety. To allay these fears, we may impose a bound K on the centrifugal force: if we denote by $R(x)$ the radius of curvature of Γ at the point of abscissa x, the setup for this new problem is as above except that X is now the open subset $\{(x_1, x_2) \mid x_2^2 < KR(x_1)\}$ of $\mathbf{R}^2$.

Let us note that these two problems cannot be treated using classical calculus of variation.

EX. 3: *Flight of a rocket plane with minimal fuel consumption.* We shall assume that the flight takes place in a vertical plane P with horizontal and vertical coordinates x and h respectively. It is sufficient to study the trajectory of the center of gravity of the plane. Let γ denote the angle of attack of the wings, ε the angle between the thrust and the velocity vector, g the gravity constant, V_E the effective velocity of the outgoing combustion products, ρ the rate of combustion, m the mass of the plane (with the fuel), D, L the drag and the lift respectively, V the velocity of the center of gravity, x, h its coordinates. The duration of the flight is given and is equal to T. We want to minimize the fuel consumption.

The equations of the motion are $\frac{dx}{dt}(t) = V(t)\cos\gamma(t)$, $\frac{dh}{dt}(t) =$

$V(t)\sin\gamma(t)$, $\frac{dm}{dt}(t) = -\rho(t)$, with

$$\frac{dV}{dt}(t) = -g\sin\gamma(t) + \frac{V_E\rho(t)}{m(t)}\cos\varepsilon(t) - \frac{D}{m(t)},$$
$$V(t)\frac{d\gamma}{dt}(t) = -g\cos\gamma(t) + \frac{V_E\rho(t)}{m(t)}\sin\varepsilon(t) + \frac{L}{m(t)}.$$

Note that the mass varies with time since we are burning fuel.

One wants to minimize the amount Q of fuel burned during the flight: $Q = m(0) - m(T)$. D and L are usually functions of h and V. The control parameters are ρ and ε; ρ varies between 0 and $\rho_s > 0$, ε between $-\varepsilon_s$ and $\varepsilon_s > 0$.

To set up a control theoretic model of the plane, we take as U the square $[0, \beta_s] \times [-\varepsilon_s, \varepsilon_s]$. The state of the plane depends on the five parameters, x, h, m, V, γ. As state space X we take the subset $\mathbf{R}^2 \times [0, m_s] \times \mathbf{R}_+ \times S^1$ of the manifold $\mathbf{R}^4 \times S^1$, m_s being the mass of the plane with a full tank. The field F is given by:

$$F_1(x,u) = V\cos\gamma, \quad F_2(x,u) = V\sin\gamma, \quad F_3(x,u) = -u_1,$$
$$F_4(x,u) = -g\sin\gamma + \frac{u_1 V_E}{m}\cos u_2 - \frac{D(h,V)}{m},$$
$$F_5(x,u) = -\frac{g\cos\gamma}{V} + \frac{u_1 V_E}{Vm}\sin u_2 + \frac{L(h,V)}{Vm}.$$

The amount $Q = m(0) - m(T)$ is equal to $\int_0^T \rho(t)dt = \int_0^T u_1(t)dt$. Hence $c(x,u) = u_1$ is the cost function.

Optimal control problems do not always have a solution. Let us give a couple of examples of this occurrence.

Ex. 4: Let $U = X = \mathbf{R}$, $F(x,u) = u$, $c(x,u) = \frac{x^2+u^2}{2}$, $A = \{0\}$, $B = \{1\}$. We have to minimize $\int_0^{T_x} \left[x(t)^2 + \left(\frac{dx}{dt}(t)\right)^2\right] dt$ among all absolutely continuous curves $x : [0, T_x] \to \mathbf{R}$ such that $x(0) = 0$,

$x(T_x) = 1$. This is a classical calculus of variations problem. An optimal curve $\overline{x} : [0, \overline{T}] \to \mathbf{R}$, if it exists, satisfies the Euler equation $\frac{d^2\overline{x}}{dt^2}(t) = \overline{x}(t)$, $0 \le t \le \overline{T}$. Since $\overline{x}(0) = 0$, $\overline{x}(t) = asht$, where $a = \frac{1}{sh\overline{T}}$. The cost $\varphi(\overline{x})$ of $\overline{x}$ is $\int_0^{\overline{T}} a^2 \left(\frac{ch^2t+sh^2t}{2}\right) dt$: $\varphi(\overline{x}) = \frac{a^2}{4} sh(2\overline{T}) = \frac{a^2}{2} sh\overline{T} ch\overline{T} = \frac{1}{2}\coth(\overline{T})$. Hence the minimum is $\frac{1}{2}$. It cannot be attained since $\coth \overline{T} > 1$, for all $\overline{T} \ge 0$.

Ex. 5: $U = [-1, +1] \subset \mathbf{R}$, $X = \mathbf{R}^2$, $F(x, u) = (u, u^2)$, $c(x, u) = x_1^2$, $A = \{(0,0)\}$, $B = \{(0,1)\}$. Then $\mathcal{C}(x, u) \ge 0$ for all $(x, u) \in Tr(A, B)$. For any integer $n \ge 1$ define $u_n : [0,1] \to U$ as

$$u_n(t) = \begin{cases} +1 & \frac{2k}{2n} \le t \le \frac{2k+1}{2n} \\ -1 & \frac{2k+1}{2n} \le t < \frac{2k+2}{2n} \end{cases}.$$

Then, if $x_n : [0,1] \to X$ is the function:

$$x_{n1}(t) = \begin{cases} t - \frac{2k}{2n} & \frac{2k}{2n} \le t \le \frac{2k+1}{2n} \\ \frac{2k+2}{2n} - 1 & \frac{2k+1}{2n} \le t \le \frac{2k+2}{2n} \end{cases}$$

$x_{n2}(t) = t$, then $(x_n, u_n) : [0,1] \to X \times U$ is a trajectory in $Tr(A, B)$ and $\mathcal{C}(x_n, u_n) \le \frac{1}{2n^2}$. Hence: $\inf\{\mathcal{C}(x, u) \mid (x, u) \in Tr(A, B)\} = 0$. It is not attained: were $(\overline{x}, \overline{u}) : [0, \overline{T}] \to X \times U$ an optimal trajectory, then $0 = \mathcal{C}(\overline{x}, \overline{u}) = \int_0^{\overline{T}} x_1^2(t)dt$. Hence $\overline{x}_1(t) = 0$. But then $\overline{u} = 0$ and $\overline{x}_2(t) = 0$, which is a contradiction since $\overline{x}_2(\overline{T}) = 1$.

§2 The most famous problem in optimal control

It is the LQC problem. In this problem, the system is a linear one and the cost c is quadratic:

$$c(x, u) = \frac{1}{2}\langle Pu, u\rangle + \langle Qx, u\rangle + \frac{1}{2}\langle Rx, x\rangle,$$

where

$$P : U \to U', \quad Q : X \to U', \quad R : X \to X'$$

are linear (U', X' are the duals of U and X respectively) and P and R are symmetric:

$$\langle Pu_1, u_2 \rangle = \langle Pu_2, u_1 \rangle \quad \langle Rx_1, x_2 \rangle = \langle Rx_2, x_1 \rangle ;$$

α and β are two points in X and it is a fixed time problem, the duration of the process being $T > 0$.

REMARK 0: The reason for studying the fixed time problem only, is that the non-fixed time problem has no solution. This is shown by Example 4: Ex. 4 is a LQC problem with $\alpha = 0$, $\beta = 1$, $P = 1$, $R = 1$, $Q = 0$.

Our basic assumption will be:

(H) $\qquad c$ is positive definite.

(H) implies that the cost $\mathcal{C}(x, u)$ of a trajectory $(x, u) : [a, b] \to X \times U$ is finite if and only if $u \in L^2([a, b]; U)$. Hence in what follows we shall assume that the controls are in L^2.

NOTATIONS: $AC^2([0, T]; X) = \{x : [0, T] \to X,\ x$ absolutely continuons, $\frac{dx}{dt} \in L^2([0, T]; X\})$; $E = AC^2([0, T]; X) \times L^2([0, T]; X)$.

§3 Necessary conditions in the LQC problem

Assume that $(\overline{x}, \overline{u}) : [0, \overline{T}] \to X \times U$ belongs to $E \cap Tr(\alpha, \beta, T)$ and is optimal for the cost $\mathcal{C}$ in $Tr(\alpha, \beta, T)$. To find conditions satisfied by $(\overline{x}, \overline{u})$, we use the variation method.

Let $(z, v) \in E \cap Tr(0,0,T)$. Define $(x_\lambda, u_\lambda) : [0,T] \to X \times U$ as follows: $x_\lambda = \overline{x} + \lambda z$, $u_\lambda = \overline{u} + \lambda v$, $\lambda \in \mathbf{R}$. Then $(x_\lambda, u_\lambda) \in Tr(\alpha, \beta, T)$ and for all $\lambda \in \mathbf{R}$, $\mathcal{C}(x_\lambda, u_\lambda) \geq \mathcal{C}(\overline{x}, \overline{u})$. This implies that for all $(z, v) \in E \cap Tr(0,0,T)$,

$$(\mathcal{L}) \qquad \int_0^T \left[\frac{\partial c}{\partial x}(\overline{x}, \overline{u}) z + \frac{\partial c}{\partial u}(\overline{x}, \overline{u}) v \right] dt = 0.$$

F is a linear system: $F(x, u) = Ax + Bu$, $A \in \mathrm{End}(X)$, $B \in \mathcal{L}$ in (U, X). Let $D : E \cap Tr(0, X, T) \to L^2([0,T]; X) \times X$ be given by

$$D(y, w) = \left[-\frac{dy}{dt} + Ay + Bw, y(T) \right]$$

and denote by $L : E \to \mathbf{R}$ the linear form

$$L(y, w) = \int_0^T \left[\frac{\partial c}{\partial x}(\overline{x}, \overline{u}) y + \frac{\partial c}{\partial u}(\overline{x}, \overline{u}) w \right] dt.$$

Then it is clear that $E \cap Tr(0,0,T)$ is the kernel of D.

Condition $(\mathcal{L})$ can be rephrased as:

$$(\mathcal{L}') \qquad \mathrm{Ker}\, L \supset \mathrm{Ker}\, D$$

and we have:

LEMMA 0. *A necessary condition for $(\overline{x}, \overline{u})$ to be optimal is that L belongs to the image of D', D' the transpose of D.*

To get a handier condition on $(\overline{x}, \overline{u})$ out of this Lemma, we have to get a concrete realization of D'. As dual of $L^2([0,T]; X) \times X$, it is natural to take $L^2([0,T]; X') \times X'$, X' dual of X. If $AC^2([0,T]; 0, X)$

is the subspace of $AC^2([0,T];X)$ of all curves starting at 0, a realisation of the dual of $AC^2([0,T];0,X)$ is $L^2([0,T];X')$ with the pairing: $x \in AC^2([0,T];0,X)$, $q \in L^2([0,T];X')$

$$\langle x, q\rangle = \int_0^T \left\langle \frac{dx}{dt}(t), q(t)\right\rangle dt,$$

where $\langle\ ,\ \rangle$ is the natural pairing $X \times X' \to \mathbf{R}$. Then a realization of the dual E_0' of $E \cap Tr(0,X,T)$ is $L^2([0,T];X') \times L^2([0,T];U')$, ($U'$ dual of U) and $D' : L^2([0,T];X') \times X' \to L^2([0,T];X') \times L^2([0,T];U')$ is given by

$$D'(p,\xi)(t) = \left[\int_t^T A'p(s)ds + \xi - p(t), B'p(t)\right],$$

$A' \in \mathrm{End}(X')$, $B' : \mathrm{Lin}(X',U')$ being the transposes of A and B.

Let us find the representation of $L \in E_0'$ in the representation $L^2([0,T];X') \times L^2([0,T];U')$ of E_0': if $(y,w) \in E_0 \cap Tr(0,X,T)$,

$$L(y,w) = \int_0^T \left[c_1(t)\frac{dy}{dt} + \frac{\partial c}{\partial u}(\overline{x},\overline{u})w\right] dt,$$

$$c_1(t) = \int_t^T \frac{\partial c}{\partial x}(\overline{x}(s),\overline{u}(s))ds.$$

The Lemma 0 says that there exists a pair $(\overline{p},\xi) \in L^2([0,T];X')\times X'$ such that $D'(\overline{p},\overline{\xi}) = L$, that is,

$$\text{(EXT1)} \qquad \begin{cases} \overline{\xi} + \int_t^T A'\overline{p}(s)ds - \overline{p}(t) = \int_t^T \frac{\partial c}{\partial x}(\overline{x}(s),\overline{u}(s))ds \\ B'\overline{p}(t) = \frac{\partial c}{\partial u}(\overline{x}(t),\overline{u}(t)). \end{cases}$$

The first condition implies that $\overline{p}$ is absolutely continuous. Hence if $(\overline{x},\overline{u})$ is optimal, one can find an absolutely continuous $\overline{p} : [0,T] \to$

X' such that

$$\text{(EXT2)} \qquad \begin{cases} \frac{d\overline{p}}{dt} + A'\overline{p} = \frac{\partial c}{\partial x}(\overline{x}, \overline{u}) = Q'\overline{u} + R\overline{x} \\ B'\overline{p} = \frac{\partial c}{\partial u}(\overline{x}, \overline{u}) = P\overline{u} + Q\overline{x}. \end{cases}$$

A more ingenious and deeper way of expressing these relations is to introduce the function $H : X \times X' \times U \to \mathbf{R}$, $H(x, p, u) = \langle p, Ax + Bu \rangle - c(x, u)$. H is a 2^{nd} degree polynomial in x, p and u. Since c is positive definite, P is positive definite and hence for fixed $(x, p) \in X \times X'$, $H(x, p, u) \to -\infty$ as $u \to \infty$. This shows that $\sup_u H(x, p, u)$ exists and is attained at $\check{u}$ such that:

$$\frac{\partial H}{\partial u}(x, p, \check{u}) = 0;$$

in other words: $B'p - \frac{\partial c}{\partial u}(x, \check{u}) = 0$. Hence the condition (EXT2) above is equivalent to:

THEOREM 0. *An optimal trajectory* $(\overline{x}, \overline{u}) : [0, T] \to X \times U$ *with* $\mathcal{C}(\overline{x}, \overline{u}) < +\infty$ *is the projection on* $X \times U$ *of a curve* $(\overline{x}, \overline{p}, \overline{u}) : [0, T] \to X \times X' \times U$ *satisfying the following conditions:*

$$\text{(EXT3)} \qquad \begin{cases} \overline{x}, \overline{p} \text{ absolutely continuous, } \overline{u} \text{ measurable,} \\ \frac{d\overline{x}}{dt}(t) = \frac{\partial H}{\partial p}(\overline{x}(t), \overline{p}(t), \overline{u}(t)), \\ -\frac{d\overline{p}}{dt}(t) = \frac{\partial H}{\partial x}(\overline{x}(t), \overline{p}(t), \overline{u}(t)), \\ H(\overline{x}(t), \overline{p}(t), \overline{u}(t)) = \sup\{H(\overline{x}(t), \overline{p}(t), u) \mid u \in U\}, \end{cases}$$

for almost all $t \in [0, T]$.

This is a particular case of the celebrated maximum principle which applies in the general situation.

DEFINITION 3: A curve $(x, p, u) : [a, b] \to X \times X' \times U$ satisfying EXT3 is called an extremal.

Actually, in this case, we can go much further, due to the fact that the supremum in the 3^d condition of EXT3, is attained at a unique point $\check{u}(x,p) = P^{-1}B'p - P^{-1}Qx$.

Let $\mathcal{H} : X \times X' \to \mathbf{R}$ denote the function $\mathcal{H}(x,p) = H(x,p,\check{u}(x,p))$. $\mathcal{H}$ is a quadratic form in (x,p):

$$\mathcal{H}(x,p) = \langle A_1 x, p\rangle + \frac{1}{2}\langle BP^{-1}B'p, p\rangle - \frac{1}{2}\langle R_1, x, x\rangle,$$
$$A_1 = A - BP^{-1}Q, \quad R_1 = R - Q'P^{-1}Q.$$

Then we have the following Corollary.

COROLLARY. *If a trajectory $(\overline{x},\overline{u}) : [0,T] \to X \times U$ is optimal, it is the projection of a trajectory $(\overline{x},\overline{p}) : [0,T] \to X \times X'$ of the linear Hamiltonian system:*

$$\text{(EXT4)} \qquad \begin{cases} \frac{d\overline{x}}{dt}(t) = \frac{\partial \mathcal{H}}{\partial p}(\overline{x}(t),\overline{p}(t)) = A_1\overline{x}(t) + BP^{-1}B'\overline{p}(t), \\ -\frac{d\overline{p}}{dt}(t) = \frac{\partial \mathcal{H}}{\partial x}(\overline{x}(t),\overline{p}(t)) = A_1'\overline{p}(t) - R_1\overline{x}(t), \\ \overline{u}(t) = P^{-1}B'\overline{p}(t) - P^{-1}Q\overline{x}(t). \end{cases}$$

In other words, EXT3 is equivalent to EXT4.

§4 Optimal control synthesis in the *LQC* case

Our next step will be to analyze more deeply the system EXT1 and the relation between its trajectories and the optimal trajectories of the given *LQC* system. This is done in the next Lemma.

LEMMA 1. *Assume $(\overline{x},\overline{p},\overline{u}) : [0,T] \to X \times X' \times U$ is a solution of EXT4 suth that $\overline{x}(0) = \alpha$, $\overline{x}(T) = \beta$. Then $(\overline{x},\overline{u})$ is an optimal trajectory in $Tr(A,B,T)$ and it is the unique one.*

PROOF: Let $(x, u) : [0, T] \to X \times U$ be a trajectory in $Tr(A, B, T)$ such that $\mathcal{C}(x, u) < +\infty$. Then $\mathcal{C}(x, u) = \mathcal{C}(\overline{x}, \overline{u}) + \mathcal{C}(x - \overline{x}, u - \overline{u}) + \mathcal{C}_1$, where

$$\mathcal{C}_1 = \int_0^T \left[\frac{\partial c}{\partial x}(\overline{x}, \overline{u})(x - \overline{x}) + \frac{\partial c}{\partial u}(\overline{x}, \overline{u})(u - \overline{u}) \right] dt.$$

Using the conditions EXT2,

$$\frac{\partial c}{\partial x}(\overline{x}, \overline{u}) = \frac{d\overline{p}}{dt} + A'\overline{p},$$
$$\frac{\partial c}{\partial u}(\overline{x}, \overline{u}) = B'\overline{p},$$

we get $\mathcal{C}_1 = \int_0^T \left[\left\langle x - \overline{x}, \frac{d\overline{p}}{dt} + A'\overline{p} \right\rangle + \langle u - \overline{u}, B'\overline{p} \rangle \right] dt$ and integrating by parts the first term:

$$\mathcal{C}_1 = \langle x - \overline{x}, \overline{p} \rangle \Big|_0^T + \int_0^T \left\langle -\frac{d(x - \overline{x})}{dt} + A(x - \overline{x}) + B(u - \overline{u}), p \right\rangle dt.$$

But $(x - \overline{x})(0) = \alpha - \alpha = 0$, $(x - \overline{x})(T) = \beta - \beta = 0$ and $\frac{d(x-\overline{x})}{dt} = A(x - \overline{x}) + B(u - \overline{u})$ for almost all t, hence $\mathcal{C}_1 = 0$. Since $\mathcal{C}(x - \overline{x}, u - \overline{u}) \geq 0$, $\mathcal{C}(x, u) \geq \mathcal{C}(\overline{x}, \overline{u})$. In fact $\mathcal{C}(x, u) > \mathcal{C}(\overline{x}, \overline{u})$ unless $\mathcal{C}(x - \overline{x}, u - \overline{u}) = 0$. But since c is positive definite, this last condition entails $x = \overline{x}$, $u = \overline{u}$.

All that remains to be done is to see if, whatever α, β are in X, there exists a solution $(\overline{x}, \overline{p}, \overline{u})$ of EXT4 such that $\overline{x}(0) = \alpha$ $\overline{x}(T) = \beta$.

In order to do this efficiently, let us recall some symplectic geometry: $X \times X'$ has a natural symplectic structure $\Omega : X \times X' \times X \times X' \to \mathbf{R}$ given by:

$$\Omega(x_1, p_1, x_2, p_2) = \langle x_1, p_2 \rangle - \langle x_2, p_1 \rangle .$$

Denote by $\mathbf{R} \ni t \mapsto \varphi_t \in Sp(X \times X')$ (the symplectic linear group of $(X \times X', \Omega)$) the flow of the quadratic Hamiltonian $\vec{\mathcal{H}}$. Then the set of points $\overline{x}(t)$ of all solutions $(\overline{x}, \overline{p}, \overline{u}) : [0,T] \to X \times X' \times U$ of EXT4 satisfying $\overline{x}(0) = \alpha$ is just the projection onto X of the affine subspace $\varphi_T(\{\alpha\} \times X')$ of $X \times X'$. Hence there will be a trajectory joining α to β for any points $\alpha, \beta \in X$ if and only if the projection of $\varphi_T(\{\alpha\} \times X')$ on X is onto. Since the dimensions of $\varphi_T(\{\alpha\} \times X')$ and of X are the same, this is equivalent to saying that $\varphi_T(\{\alpha\} \times X')$ is transversal to X' in $X \times X'$. □

LEMMA 2. *If (B, A) is controllable then $\varphi_t(\{\alpha\} \times X')$ is transversal to X and X' in $X \times X'$, for all $\alpha \in X$ and $t \geq 0$.*

PROOF: Since $\varphi_t(\{\alpha\} \times X')$ is parallel to $\varphi_t(\{0\} \times X') = \varphi_t(X')$, it is sufficient to take $\alpha = 0$. Let $(x,p) \in X \times X'$ and set $\varphi_t(x,p) = (x(t), p(t))$. Then:

$$\frac{d}{dt}\langle x(t), p(t)\rangle = \left\langle \frac{\partial \mathcal{H}}{\partial p}(x(t), p(t)), p(t)\right\rangle - \left\langle x(t), \frac{\partial \mathcal{H}}{\partial x}(x(t), p(t))\right\rangle$$
$$= \frac{1}{2}\left\langle Bp^{-1}B'p(t), p(t)\right\rangle + \frac{1}{2}\langle R_1 x(t), x(t)\rangle ,$$

$R_1 = R - Q'P^{-1}Q$. For any $x \in X$, $c(x, -P^{-1}Qx) = \frac{1}{2}\langle R_1 x, x\rangle$. Hence R_1 is positive definite. Since P^{-1} is positive definite, $\left\langle BP^{-1}B'p(t), p(t)\right\rangle \geq 0$. From this we conclude that for $t > 0$, $\langle x(t), p(t)\rangle \geq \langle x, p\rangle$ and we have equality only if $x(s) = 0$ and $B'p(s) = 0$ for $0 \leq s \leq t$. In this last case: $\frac{dp}{dt}(s) = -A'p(s)$, and $B'p(s) = 0$, $0 \leq s \leq t$. This implies that $p \in \bigcap_{n=0}^{\infty} \operatorname{Ker} B'A'^n$. Since (B, A) is controllable, $p = 0$. Thus $\langle x(t), p(t)\rangle > \langle x, p\rangle$, unless $x = 0$ and $p = 0$, which shows that for any point $(x(t), p(t)) \in \varphi_t(X')$, 0 excepted, $\langle x(t), p(t)\rangle > 0$ and proves the Lemma. □

REMARK 1: The Lemma is still true if we assume that (B, A) is controllable, (A_1, R_1) is observable, P is positive definite, c is positive semi-definite.

Let us sum up our results.

THEOREM 1. *Assume that the cost function c is positive definite and that (B, A) is controllable. Then for any $T > 0$ and any pair α, β in X, there exists a unique trajectory going from α to β in time T and minimizing the cost in $Tr(\alpha, \beta, T)$.*

§5 The Ricatti equation

Using the notations of §4, for any $t \geq 0$, $\varphi_t(X')$ is transversal to X in $X \times X'$. Since $\varphi_t(X')$, X, X' are all Lagrangian subspaces of $(X \times X', \Omega)$, there exists a unique symmetric linear mapping $K(t) : X' \to X$ such that $\varphi_t(X')$ is the graph of $K(t)$ in $X \times X'$: $\varphi_t(X') = \{(K(t)p, p) \mid p \in X'\}$. We want to discuss the function $\mathbf{R}_+ \ni t \mapsto K(t) \in \mathrm{Sym}(X', X)$, the set of all symmetric linear mappings $X' \to X$.

The symplectic flow $\{\varphi_t \mid t \in \mathbf{R}\}$ has a generating function $S : \mathbf{R} \times X \times X' \to \mathbf{R}$, i.e., a function with the following property: if $(x, p) \in X \times X'$ and $\varphi_t(x, p) = (x(t), p(t))$, then

$$\text{(G)} \qquad x(t) = \frac{\partial S}{\partial p}(t, x, p(t)), \quad p = \frac{\partial S}{\partial x}(t, x, p(t)).$$

Since φ_t is a linear isomorphism, the relations (G) show that S is a quadratic form in x and p,

$$S(t, x, p) = \frac{1}{2}\langle S_{11}(t)x, x\rangle + \frac{1}{2}\langle S_{22}(t)p, p\rangle + \langle S_{12}(t)x, p\rangle,$$

where $S_{11}(t): X \to X'$, $S_{22}(t): X' \to X$, $S_{12}(t): X \to X$ and S_{11}, S_{22} are symmetric. Also the first equation of (G) shows that the linear space $\varphi_t(X')$ is the space $\{(\overline{x}, \overline{p}) \mid \overline{x} = \frac{\partial S}{\partial p}(t, 0, \overline{p}) = S_{22}(t)\overline{p}\}$; this implies that $S_{22} = K$.

The function S is the solution of the following Cauchy problem:

$$\begin{cases} \frac{\partial S}{\partial t} + \mathcal{H}(\frac{\partial S}{\partial p}, p) = 0 \\ S(0, x, p) = \langle x, p \rangle \end{cases}$$

(x plays the role of a parameter). Hence the matrix functions S_{11}, S_{12}, $S_{22} = K$ satisfy the following equations:

$$\begin{cases} \frac{\partial S_{11}}{\partial t} - S'_{12} R_1 S_{12} = 0 & S_{11}(0) = 0 \\ \frac{\partial S_{12}}{\partial t} - A_1 S_{12} - S_{22} R_1 S_{12} = 0 & S_{12}(0) = Id_X \\ \frac{\partial K}{\partial t} + A_1 K + K A'_1 + B P^{-1} B' - K R_1 K = 0 & K(0) = 0. \end{cases}$$

PROPOSITION 0. *Assume (B, A) is controllable and c positive definite. Given a point $x_T \in X$ and a number T, the unique optimal trajectory for fixed time T joining 0 to x_T is the projection on X of the unique trajectory of the Hamiltonian $\vec{\mathcal{H}}$ passing through the point $(x_T, K(T)^{-1} x_T)$ at time T. The matrix function $\mathbf{R}_+ \ni t \mapsto K(t) \in Sym(X', X)$ is the unique solution of the Cauchy problem*

$$(\mathcal{K}) \quad \begin{cases} \frac{dK}{dt}(t) + A_1 K(t) + K(t) A'_1 - K(t) R_1 K(t) + B P^{-1} B' = 0 \\ K(0) = 0. \end{cases}$$

REMARK 2: The fact that $K(t)$ is invertible for $t > 0$, is a direct consequence of the fact, from Lemma 2, that $\varphi_t(X')$ is transversal to X' in $X \times X'$ for $t > 0$.

NOTATION: The equation from Proposition 0, satisfied by K, is called the *Ricatti equation* associated to the LQC problem.

Let us now study the evolution of $K(t)$ as t goes to $+\infty$. We need the following Lemmas.

LEMMA 3. *If (B, A) is controllable and c positive definite then:*

1) the Hamiltonian field $\vec{\mathcal{H}}$ of $\mathcal{H}$ has no purely imaginary eigenvalues;

2) $X \times X'$ splits into a direct sum of two $\vec{\mathcal{H}}$-invariant Lagrangian subspaces L_+ and L_- such that the spectrum of the restriction of $\vec{\mathcal{H}}$ to L_+ (resp. L_-) has eigenvalues with positive (resp. negative) real part only.

REMARK 3: The Lemma is still true if we assume that (B, A) is controllable, (A_1, B_1) observable, c positive semi-definite, and $U \ni U \mapsto c(0, u)$ positive definite.

PROOF: Let $(x, p) \neq 0$ be an eigenvector of the field of $\mathcal{H}$ in the complexified $X_{\mathbf{C}} \times X'_{\mathbf{C}}$ with eigenvalue $\lambda \in \sqrt{-1}\mathbf{R}$. Then:

$$\begin{cases} A_1 x + BP^{-1}B'p = \lambda x \\ R_1 x - A'_1 p = \lambda p. \end{cases}$$

From this, multiplying scalarly by p^* the first equation and by x^* the second one, x^*, p^* being the complex conjugates of x and p, we get

$$\langle A_1 x, p^* \rangle + \langle BP^{-1}B'p, p^* \rangle = \lambda \langle x, p^* \rangle ,$$

$$\langle R_1 x, x^* \rangle - \langle A'_1 p, x^* \rangle = \lambda \langle x^*, p \rangle ;$$

this last equation is equivalent to:

$$\langle R_1 x, x^* \rangle - \langle A_1 x, p^* \rangle = -\lambda \langle x, p^* \rangle ,$$

and we obtain

$$\langle R_1 x, x^* \rangle + \langle BP^{-1}B'p, p^* \rangle = 0.$$

By the positive definiteness of c (see the proof of Lemma 2) this implies: $x = 0$, $B'p = 0$; therefore $A_1'p = -\lambda p$ and $p \in \bigcap_{n=0}^{\infty} \operatorname{Ker} B'A'^n$. Since (B, A) is controllable, this intersection is 0. Hence $p = 0$, which is a contradiction. □

The second part of the lemma is a direct consequence of the first and some well known facts about the symplectic derivations.

LEMMA 4 (SEE [**Sh**]). *Assume (B, A) is controllable and c positive definite. If L is any $\vec{\mathcal{H}}$-invariant Lagrangian subspace of $X \times X'$, then $L \cap X' = 0$ and $L \cap X = 0$.*

PROOF: Let $z = (0, p) \in L \cap X'$. Since $\vec{\mathcal{H}}(z) \in L$, $\Omega(z, \vec{\mathcal{H}}(z)) = 0$. Now $\Omega(z, \vec{\mathcal{H}}(z)) = \langle BP^{-1}B'p, p \rangle$, therefore $B'p = 0$ and $\vec{\mathcal{H}}(z) = (0, -A_1'p) \in L \cap X'$. This shows that $L \cap X'$ is A_1'-invariant and contained in the kernel of B'. But $L \cap X' \subset \bigcap_{n=0}^{\infty} \operatorname{Ker} B'A_1'^n = 0$, since (B, A) and also (B, A_1) is controllable.

If $z = (x, 0) \in L \cap X$, then again $\Omega(z, \vec{\mathcal{H}}(z)) = 0$. But $\Omega(z, \vec{\mathcal{H}}(z)) = \langle x, R_1 x \rangle$, which shows that $x = 0$. □

COROLLARY. *There exist symmetric isomorphisms $K_+, K_- : X' \to X$ such that L_+ and L_- are the graphs of K_+ and K_-.*

PROOF: L_+, L_- are $\vec{\mathcal{H}}$-invariant Lagrangian spaces by Lemma 3.

Here is the main result of this section.

PROPOSITION 1. *Assume (B, A) is controllable and c is positive definite. Let $\varepsilon = +$ or $-$.*

1) The solution $K(t)$ of the system $(\mathcal{K})$ of Proposition 0 converges exponentially to K_ε as t goes to $\varepsilon\infty$.

2) K_ε satisfies the algebraic Ricatti equation:

$$A_1 K_\varepsilon + K_\varepsilon A_1' - K R_1 K_\varepsilon + B P^{-1} B' = 0,$$

where $A_1 = A - BP^{-1}Q, \quad R_1 = R - Q'P^{-1}Q$.

3) The linear mapping $F : X \to U$, $F = P^{-1}[B'K^{-1} - Q]$ is a stabilizing feedback for the system (B, A).

PROOF: 1) By Lemma 4, $L_+ \cap X' = L_- \cap X' = 0$. Since L_+, L_- are $\vec{\mathcal{H}}$-invariant, $L_+ \cap \varphi_t(X') = L_- \cap \varphi_t(X') = 0$ for all $t \geq 0$ (φ_t is the flow of $\vec{\mathcal{H}}$). Since $\vec{\mathcal{H}}$ is hyperbolic (Lemma 3-1)) there exist a norm $\| \ \|$ on $X \times X'$ and a positive number $\alpha > 0$ such that:

$$\|\varphi_t(z)\| \geq e^{t\alpha} \|z\| \quad \text{if} \quad z \in L_+, \quad t \geq 0,$$

$$\|\varphi_t(z)\| \leq e^{-t\alpha} \|z\| \quad \text{if} \quad z \in L_-, \quad t \geq 0,$$

and vice-versa if $t < 0$.

If $z \in X' - \{0\}$, $z = z_+ + z_-$, $z_+ \in L_+$, $z_- \in L_-$, $z_+ \neq 0$, $z_- \neq 0$; then

$$\varphi_t(z) = \varphi_t(z_+) + \varphi_t(z_-) = \begin{cases} \varphi_t(z_+) + 0(e^{-t\alpha}), & t > 0 \\ \varphi_t(z_-) + 0(e^{t\alpha}), & t < 0. \end{cases}$$

This shows that in the Grassmannian manifold of all Lagrangians in $X \times X'$, $\varphi_t(X')$ converges exponentially to L_ε as $t \to \varepsilon\infty$. Since L_ε, $\varphi_t(X')$ are the graphs of K_ε and $K(t)$ respectively, we get 1).

2) Since K_ε is the limit, as $t \to \varepsilon\infty$, of the trajectory $\mathbf{R} \ni t \mapsto K(t)$ of $(\mathcal{K})$, K_ε is a stationary point of $(\mathcal{K})$: this gives 2).

3) Let $\Gamma_- = \begin{bmatrix} I_X & 0 \\ K_+^{-1} & I_{X'} \end{bmatrix} : X \times X' \to X \times X'$. Then $L_- = \Gamma_-(X)$ and it is easily checked that:

$$\Gamma_-^{-1} \vec{\mathcal{H}} \Gamma_- = \begin{bmatrix} A_1 + BP^{-1}B'K_-^{-1} & BP^{-1}B' \\ 0 & -A_1' - K_-^{-1}BP^{-1}B' \end{bmatrix}.$$

The spectrum of the restriction of $\Gamma_-^{-1} \vec{\mathcal{H}} \Gamma_-$ to X is that of the restriction of $\vec{\mathcal{H}}$ to L_-. Hence all eigenvalues of $A_1 + BP^{-1}B'K_-^{-1}$ have negative real parts. This proves 3). □

CHAPTER VI

OPTIMAL CONTROL – MAXIMUM PRINCIPLE

§0 Introduction

Let $\Sigma = (U, X, F)$ be a system with a cost function $c : X \times U \to \mathbf{R}$ and boundary conditions (A, B). We assume that A and B are smooth submanifolds of X and $\mathbf{R} \times X$ respectively. We are going to discuss conditions satisfied by the optimal trajectories of the system $\tilde{\Sigma} = (U, X, F, c, A, B)$, that is, trajectories $(\overline{x}, \overline{u}) : [0, \overline{T}] \to X \times U$ of Σ such that

(i) $\overline{x}(0) \in A$, $(\overline{T}, x(\overline{T})) \in B$;

(ii) $\int_0^{\overline{T}} c(\overline{x}(t), \overline{u}(t))dt = \inf\{\int_0^T c(x(t), u(t))dt \mid (x, u) : [0, T] \to X \times U$, trajectory of Σ, $(0, x(0)) \in A$, $(T, x(T)) \in B\}$.

Obviously there are other optimal control problems but most of them can be reduced to the one above. Let us give an example. Quite frequently, the cost of a trajectory $(x, u) : [0, T_u] \to X \times U$ has the more general form

$$\int_0^{T_u} c(x(t), u(t))dt + \gamma(T_u, x(T_u)),$$

where γ is a given smooth function on B.

At least in the C^∞ case, this can be reduced to our original problem, provided that $A \cap B = \phi$. In this case, one can extend γ

to $\tilde{\gamma} : X \times \mathbf{R} \to \mathbf{R}$ in a C^∞ fashion, with the following conditions being satisfied: $\tilde{\gamma} \mid B = \gamma$, $\tilde{\gamma} \mid A = 0$. Then let $\tilde{\tilde{c}} : X \times U \times \mathbf{R} \to \mathbf{R}$ be the function $\tilde{\tilde{c}}(x,u,t) = c(x,u) + \frac{\partial \tilde{\gamma}}{\partial x}(x,t)F(x,u) + \frac{\partial \tilde{\gamma}}{\partial t}(x,t)$. The new cost depends on the time, but this can be taken care of by extending the state space to include time as a state variable: the new system $\tilde{\Sigma} = (U, \tilde{X}, \tilde{F})$ has $\tilde{X} = \mathbf{R} \times X$ as new state space. The new dynamic is $\tilde{F}(\tilde{x}, u) = (1, F(x,u))$ (x = projection of $\tilde{x}$ onto X), the cost $\tilde{c} : \tilde{X} \times U \to \mathbf{R}$ is $\tilde{c}(x_0, x, u) = \tilde{\tilde{c}}(x, u, x_0)$, $\tilde{A} = \{0\} \times A$, $\tilde{B} = B$. Hence this case is easy to handle.

On the other hand, if the system is subjected to one-sided constraints (that is: X is now a manifold with boundary), then the problem is quite different and cannot be handled by the methods we are going to discuss (see **[C]** for a treatment of that case).

§1 Statement of the maximum principle

Let us return do the system $\tilde{\Sigma}$ discussed at the beginning of §0. Let $H_\lambda : T^*X \times U \to \mathbf{R}$ denote the smooth function $H_\lambda(p,u) = \langle p, F(x,u)\rangle - \lambda c(x,u)$, where x is the projection of p on X and λ is a parameter taking the values 0 or 1 only.

THEOREM 0. *Let* $(\overline{x}, \overline{u})\colon [0, \overline{T}] \to X \times U$ *be an optimal trajectory of* $\tilde{\Sigma}$.

(iii) There exist a $\lambda \in \{0,1\}$ *and an absolutely continuous lifting of* $\overline{x}$, $\overline{p} : [0, \overline{T}] \to T^*X$, *satisfying the following conditions: for almost all* $t \in [0, \overline{T}]$, $\overline{p}(t) \neq 0_{\overline{x}(t)}$ *and*

$$(EX\lambda) \qquad \begin{cases} \frac{d\overline{p}}{dt}(t) = \vec{H}_\lambda(\overline{p}(t), \overline{u}(t)) \\[2ex] H_\lambda(\overline{p}(t), \overline{u}(t)) = \sup\{H_\lambda(\overline{p}(t), v) \mid v \in U\}; \end{cases}$$

$\vec{H}_\lambda$ *denotes the Hamiltonian vector field associated to the function* $p \mapsto H_\lambda(p,u)$ *(u is considered as a parameter).*

(iv) $\overline{p}$ *satisfies the following boundary conditions:* $\overline{p}(0)$ *annihilates* $T_{\overline{x}(0)}A$ *and there exists a scalar* μ *such that the vector* $(\mu, \overline{p}(T)) \in T^*_q(\mathbf{R} \times X) = T^*_{\overline{T}}\mathbf{R} \times T_{\overline{x}(\overline{T})}X$ *annihilates* T_qB*, where* $q = (\overline{T}, \overline{x}(\overline{T}))$.

DEFINITION: Any curve $(p,u) : [0,T] \ni t \mapsto (p(t), u(t)) \in T^*X \times U$ such that $p(t)$ is absolutely continuons, u is measurable and such that $(p(t), u(t))$ satisfies the condition $EX\lambda$ for almost all $t \in [0,T]$, is called an extremal. The extremal is ordinary if $\lambda = 1$, exceptional if $\lambda = 0$.

COROLLARY. H_λ *is constant along any extremal of* $EX\lambda$.

The proof of these results is given in Chapter VII.

SPECIAL FORMULATION: In case X is an open subset of an $\mathbf{R}^d$ with canonical coordinates $(x_1, \ldots, x_d)$, $T^*X = X \times (\mathbf{R}^d)^*$, dual of $\mathbf{R}^d$, and p can be represented by a pair (x,p), $x = (x_1, \ldots, x_d) \in X$, $p = (p_1, \ldots, p_d) \in (\mathbf{R}^d)^*$. The conditions $EX\lambda$ can be written as follows: for almost all $t \in [0,T]$,

$$(EX\lambda 1) \quad \begin{cases} \frac{dx_k}{dt}(t) = \frac{\partial H_\lambda}{\partial p_k}(x(t), p(t), u(t)) & \\ -\frac{dp_k}{dt}(t) = \frac{\partial H_\lambda}{\partial x_k}(x(t), p(t), u(t)) & k = 1, 2, \ldots, d \\ H_\lambda(x(t), p(t), u(t)) = \sup\{H_\lambda(x(t), p(t), v) \mid v \in U\}. & \end{cases}$$

Moreover $p(t)$ is not identically zero.

In one particular instance can the system $EX\lambda$ be written as a genuine Hamiltonian system:

THEOREM 1. *Assume that there exists a smooth function* $\check{u}$ *:* $T^*X \to U$ *with the property that* $H_\lambda(p, \check{u}(p)) > H_\lambda(p, v)$ *for all*

$v \in U$, $v \neq \check{u}(p)$ and all $p \in T^*X$. Let $\mathcal{H}_\lambda : T^*X \to \mathbf{R}$ be the function $\mathcal{H}_\lambda(p) = H_\lambda(p, \check{u}(p))$. Then $EX\lambda$ is equivalent to the system

$$\begin{cases} \frac{dp}{dt}(t) = \vec{\mathcal{H}}_\lambda(p(t)) \\ u(t) = \check{u}(p(t)); \end{cases}$$

$\vec{\mathcal{H}}_\lambda$ is the Hamiltonian field associated to $\mathcal{H}_\lambda$.

The result stated in Theorem 0 is called the maximum principle. It is the control-theoretic equivalent of the combination: Euler-Lagrange equations and the Weierstrass condition, in the classical calculus of variations (see the applications below). As these two gadgets, it enables us to narrow down the search for optimal trajectories to the projections of the extremals. Quite often there is only one or only a finite number of extremals satisfying the boundary conditions. The problem is therefore reduced to checking that this unique one or one of those is optimal. This, by the way, is not always an easy task.

One last question is: when do the exceptional extremals come into the picture? There is no corresponding object in the classical calculus of variations. It is clear that the exceptional extremals do not depend upon the cost function.

§2 Applications

EX. 1: ***Classical calculus of variations.*** Let X be an open subset of $\mathbf{R}^d$, A, B two points in X, $c : X \times \mathbf{R}^d \to \mathbf{R}$ a smooth function. Wanted is an absolutely continuous curve $\overline{x} : [0, \overline{T}] \to X$ such that $\overline{x}(0) = A$, $\overline{x}(\overline{T}) = B$ and $\int_0^{\overline{T}} c(\overline{x}(t), \frac{d\overline{x}}{dt}(t))dt$ is minimal among the absolutely continuous curves joining A to B.

One can formulate this problem as an optimal control problem taking U as $\mathbf{R}^d$ and F as $u : F(x,u) = u$. Using the form $EX\lambda 1$ of the maximum principle we get the extremal equations:
$(EX\lambda 2)$

$$\begin{cases} \frac{dx}{dt}(t) = u(t) \\ -\frac{dp}{dt}(t) = -\lambda \frac{\partial c}{\partial x}(x(t), u(t)) \\ \langle p(t), u(t)\rangle - \lambda c(x(t), u(t)) = \inf_{v \in U}\{\langle p(t), v\rangle - \lambda c(x(t), v)\}. \end{cases}$$

The last condition implies Lagrange's 1^{st} order condition:

$$p(t) = \lambda \frac{\partial c(x(t), u(t))}{\partial u}.$$

In particular, since $p(t)$ cannot be identically 0, $\lambda = 1$. Combining with the 1^{st} and 2^{nd} equations in $EX\lambda_2$ we get:

$$\frac{d}{dt}\left(\frac{\partial c(x,\dot{x})}{\partial \dot{x}}\right) = \frac{\partial c(x,\dot{x})}{\partial x}, \quad \dot{x} = \frac{dx}{dt}.$$

These are the Euler-Lagrange equations. The 3^d equation in $EX\lambda 2$ can be written as follows:

$$\frac{\partial c(x(t),\dot{x}(t))}{\partial \dot{x}}(\dot{x}(t) - v) + c(x(t), v) - c(x(t), \dot{x}(t)) \geq 0.$$

The 1^{st} member of this relation is the Weierstrass $\mathcal{E}$-function; we get the Weierstrass condition.

EX. 2: *The accelerated car.* A car moves along a prescribed curve Γ, controlled by its acceleration. Orient Γ, choosing as origin 0_Γ, the initial position of the car, call $x(t)$ the abscissa of the car's position at time t on Γ, M the mass of the car, and u its acceleration or braking force. The dynamic equation of the problem is $M\frac{d^2x}{dt^2}(t) =$

$u(t)$, $u(t) > 0$ if the car accelerates, $u(t) < 0$ if it brakes. Let a (resp. b) be the maximum braking (accelerating) force. One wants to drive the car starting at 0 from rest to a point L on the curve at a distance L from 0 on Γ, in minimum time, and stop at L. (see Ex. 2 in Chapter V).

The mathematical model of this process is a system $\Sigma = (U, X, F)$, where $U = [-a, b]$, X is $\mathbf{R}^2$ (the phase space), $F(x, u) = (x_2, \frac{u}{M})$, $\frac{dx_1}{dt} = x_2$, $\frac{dx_2}{dt} = \frac{u}{M}$, A is the point (0,0), B the point $(L, 0)$ (the speed at A and B should be 0) and the cost function c is just the constant 1. Let us write the system $EX\lambda 2$:

$$H_\lambda(x, p, u) = p_1 x_2 + \frac{p_2 u}{M} - \lambda,$$

$$\begin{cases} \frac{d\overline{x}_1}{dt}(t) = \overline{x}_2(t) \\ \frac{d\overline{x}_2}{dt}(t) = \frac{\overline{u}(t)}{M} \end{cases}, \qquad \begin{cases} -\frac{d\overline{p}_1}{dt}(t) = 0 \\ -\frac{d\overline{p}_2}{dt}(t) = \overline{p}_1 \end{cases},$$

$$\overline{p}_1(t)\overline{x}_2(t) + \frac{\overline{p}_2(t)\overline{u}(t)}{M} - \lambda = \inf\{\overline{p}_1(t)\overline{x}_2(t) + \frac{\overline{p}_2(t)v}{M} - \lambda \mid v \in U\},$$

with $\overline{x}_1(0) = \overline{x}_2(0) = 0, \quad x_1(\overline{T}) = L, \quad x_2(\overline{T}) = 0.$

Clearly λ does not matter:

$$u(t) = \begin{cases} b & \text{if } p_2(t) > 0 \\ -a & \text{if } p_2(t) < 0. \end{cases}$$

In case $p_2(t) = 0$, $u(t)$ is not determined.

We see that the projections onto X of the extremals belong to four families of curves:

$$(I)\begin{cases} x_1(t) = -\frac{a}{2M}t^2 + x_2(0)t + x_1(0) \\ x_2(t) = -\frac{a}{M}t + x_2(0) \end{cases}$$

$$(II)\begin{cases} x_1(t) = \frac{b}{2M}t^2 + x_2(0)t + x_2(0) \\ x_2(t) = \frac{b}{M}t + x_2(0) \end{cases}$$

$$(III)\begin{cases} x_1(t) = \begin{cases} -\frac{a}{2M}t^2 + x_2(0)t + x_1(0), & 0 \le t \le t_1 \\ \frac{b}{2M}(t-t_1)^2 + x_2(t_1)(t-t_1) + x_1(t_1), & t_1 < t \end{cases} \\ x_2(t) = \begin{cases} -\frac{a}{M}t + x_2(0), & 0 \le t \le t_1 \\ \frac{b}{M}(t-t_1) + x_2(t_1), & t_1 < t \end{cases} \end{cases}$$

$$(IV)\begin{cases} x_1(t) = \begin{cases} \frac{b}{2M}t^2 + x_2(0)t + x_1(0), & 0 \le t \le t_1 \\ -\frac{a}{2M}(t-t_1)^2 + x_2(t_1)(t-t_1) + x_2(t_1), & t_1 < t \end{cases} \\ x_2(t) = \begin{cases} \frac{b}{M}t + x_2(0), & 0 \le t \le t_1 \\ -\frac{a}{M}(t-t_1) + x_2(t_1), & t_1 < t \end{cases} \end{cases}$$

It is clear that a trajectory of the families I and II can not pass through L with zero speed if it starts at 0 with zero speed. If the car performs a trajectory of type III, it will have zero speed only once again; its position x, at that moment, will be $-\frac{t_1^2}{2M}\left(\frac{a^2+ba}{a}\right)$. Since $L > 0$, type III is ruled out. On the other hand, if the car follows a trajectory of type IV, it will have zero speed only once again but at that moment, $T = \frac{a+b}{a}t_1$, its position will be $\frac{ab+b^2}{2Ma}t_1^2$. Hence $t_1^2 = \frac{2MaL}{ab+b^2}$ and $T = \sqrt{\frac{2ML(a+b)}{ab}}$. This is the optimal trajectory; t_1 is the time at which one should start braking.

EX. 3: *Tank with a hole*. Let us denote by h the depth of the liquid, Q the flow rate of the feeding pipe, A_s the area of the section of the tank, V the velocity of escape of the fluid through the hole, A_h the area of the hole. The dynamics of the system are:

$$A_s\frac{dh}{dt} = Q - A_hV.$$

Now V can be taken equal to $0.6\sqrt{2gh}$, where g is the gravity constant. The flow rate Q is the control parameter.

Assuming that the supply of water is limited, by W say, and that the maximum flow rate is Q_m, maximize the time during which the level of the water in the tank is at least a (a given positive constant).

EX. 4: *Motion of a rocket in space*. We identify the rocket with a point x. It is submitted to two forces, one the gravity of the earth, which we denote by $g(x)$, and the other the thrust F of the rocket. If M denotes the total mass of the rocket, c the rate of combustion of the fuel, the equations of the motion of the rocket are:

$$\begin{cases} m\frac{d^2x}{dt^2} = mg + F \\ c\frac{dm}{dt} + \|F\| = 0 \qquad \|F\| = \text{ norm of the force } F. \end{cases}$$

One wants to maximize the fuel consumption in a flight from a point A to a point B, assuming the duration T of the flight is fixed. The first order equations, equivalent to the dynamical equations, are

$$\begin{cases} \frac{dx}{dt} = y \\ \frac{dy}{dt} = g + \frac{1}{m}F \\ \frac{dm}{dt} = -\frac{1}{c}\|F\|. \end{cases}$$

We will assume that g is a central force deriving from a potential $V(\|x\|)$. To get rid of the denominators we choose a new coordinate m_1 instead of $m : m = e^{-m_1/c}$. Then:

$$\begin{cases} \frac{dx}{dt} = y \\ \frac{dy}{dt} = g + e^{-m_1/c} \\ \frac{dm_1}{dt} = -\|F\| e^{-m_1/c}. \end{cases}$$

Let p, q, r be the coordinates dual to x, y, m respectively. The cost is $\|F\|$ and the Hamiltonian of the problem is:

$$H = \langle p, y \rangle + \left\langle q, g + Fe^{-m_1/c} \right\rangle - r \|F\| e^{-m_1/c}.$$

The control is obviously F. We assume that F can be any vector such that $\|F\| \le f$ (f some constant.)

Then, fixing x, y, m, p, q, r, the maximum of H is attained at $F = 0$ if $\|q\| < r$ and at $F = f\frac{q}{\|q\|}$ if $\|q\| > r$.

If $\|q\| = r$, $\|F\|$ is not determined by the maximum principle. Let us check if there are extremals on which $\|q\| = r$. The extremal system is in our case:

$$\frac{dx}{dt} = y, \quad \frac{dy}{dt} = g + e^{-m_1/c}F, \quad \frac{dm_1}{dt} = -\|F\| e^{-m_1/c}$$

$$-\frac{dp}{dt} = q\frac{\partial g}{\partial x}, \quad -\frac{dq}{dt} = p, \quad -\frac{dr}{dt} = \frac{1}{c}[r\|F\| - \langle q, F \rangle]e^{-m_1 c}.$$

Since $F = \|F\| \frac{q}{\|q\|}$, $\|q\|$, r are constants and it remains to determine $\|F\|$.

We have to recall that on any extremal on which $\|q\| = r$, all the derivatives $\frac{d^n\|q\|^2}{dt^n}$ are zero. The formulas for $\frac{d\|q\|^2}{dt}$ and $\frac{d^2\|q\|^2}{dt^2}$ do not involve F, only $\frac{d^3\|q\|^2}{dt^3}$ involves F. Setting it equal to zero gives us the value of $\|F\|$ we are looking for.

EX. 5: *Minimal fuel expenditure in a rocket flight.* We assume the flight takes place in a fixed vertical plane with a horizontal and a vertical axis. The horizontal coordinate will be called x, the vertical one h. Let us denote by $m(t)$ the total mass of the rocket at time t, by v its velocity, by $D(h, v)$, $L(h, v)$ respectively the drag and the lift at the altitude h and velocity v. Let γ be the angle of attack, ε the angle between the thrust and the velocity vectors of the outgoing combustion products and g the gravity constant.

The duration T of the flight is fixed. One wants to minimize the fuel expenditure $m(0) - m(T)$. The control parameters are ε and the rate of combustion of the fuel, ρ. As A we take the origin (0,0) and as B, some other point with coordinates (x_B, h_B).

Let us write the dynamical equations:

$$\frac{dx}{dt}(t) = v(t)\cos\gamma(t),\ \frac{dh}{dt}(t) = v(t)\sin\gamma(t),\ \frac{dm}{dt}(t) = \rho(t),$$

$$\frac{dv}{dt}(t) = -g\sin\gamma(t) + \frac{V_E\rho(t)}{m(t)}\cos\varepsilon(t) - \frac{D(h(t),v(t))}{m(t)},$$

$$\frac{d\gamma}{dt}(t) = -\frac{g}{v(t)}\cos\gamma(t) + \frac{V_E\rho(t)}{m(t)v(t)}\sin\varepsilon(t) + \frac{L(h(t),v(t))}{m(t)v(t)}.$$

The Hamiltonian function H_λ is the sum $H_\lambda^0 + H_\lambda^1$; H_λ^0 does not depend on the controls ρ, ε and $H_\lambda^1 = [-M + \lambda + \frac{V_E}{m}(V\cos\varepsilon + \frac{\Gamma}{v}\sin\varepsilon)]\rho$, where X, H, M, V, Γ denote the variables adjoint to x, h, m, v, γ respectively.

CHAPTER VII

PROOF OF THE MAXIMUM PRINCIPLE

§0 Approximating cones

Let $\Sigma = (U, X, F)$ be a system, A a submanifold of X. We are going to find a good approximation to the accessibility set of A at any of its points. For this we need some regularity assumptions on the trajectories we are going to study.

DEFINITION 0: A trajectory $(x, u) : [0, T] \to X \times U$ of Σ is called tame if there exist an open neighborhood V of $x([0, T])$, a positive number $T_1 > T$, an extension $\tilde{u}$ of u to $[0, T_1]$, and a continuous mapping $\varphi : [0, T_1] \times [0, T_1] \to \mathrm{Diff}(V, X)$, the set of all smooth embeddings of V into X with the usual topology, such that:

(i) $\varphi(s, s)$ is the injection of V into X, for all $s \in [0, T_1]$;

(ii) for any $x \in V$, any $s \in [0, T_1]$, the curve $[0, T_1] \ni t \mapsto \varphi(t, s)x \in X$ is absolutely continuous and for almost all $t \in [0, T_1]$

$$\frac{\partial \varphi(t, s)x}{\partial t} = F(\varphi(t, s)x, \tilde{u}(t));$$

(iii) for any $t_1 > t_2 > t_3$ in $[0, T_1]$, $\varphi(t_1, t_3) = \varphi(t_1, t_2) \circ \varphi(t_2, t_3)$ on a smaller neighborhood $V(t_3)$ of $x(t_3)$, $V(t_3) \subset V$, depending on t_3.

NOTATION: Given a trajectory $(x, u) : [0, T] \to X \times U$, a point $t \in [0, T]$ will be called a Lebesgue point of (x, u) if $\frac{dx}{dt}$ is approximately continuous at t and $\frac{dx}{dt}(t) = F(x(t), u(t))$. (See the appendix.) It is easy to see that if x is approximately continuous at t, it is differentiable there.

DEFINITION 1: Let $(\overline{x}, \overline{u}) : [0,T] \to X \times U$, $x(0) \in A$, be a tame trajectory of Σ. The approximating cone of the accessibility set of A at $\tau \in [0,T]$ along the trajectory $(\overline{x}, \overline{u})$ of Σ, denoted by $K(\overline{x}, \overline{u}, \tau, A, \Sigma)$ or, when no confusion is possible, simply by $K(\tau, A)$, is the convex cone in $T_{\overline{x}(\tau)}X$ generated by the vectors:

(iv) $\frac{\partial\overline{\varphi}(\tau,0)w}{\partial x}$, $w \in T_{x(0)}A$;

(v) $\frac{\partial\overline{\varphi}(\tau,0)}{\partial x}[F(\overline{x}(t), v) - F(\overline{x}(t), \overline{u}(t))]$, $v \in U$, $t \leq \tau$, t a Lebesgue point of $(\overline{x}, \overline{u})$;

(vi) $\varepsilon \frac{\partial\overline{\varphi}}{\partial x} F(\overline{x}(t), \overline{u}(t))$, $\varepsilon = +$ or $-$, $t \leq \tau$, t a Lebesgue point of $(\overline{x}, \overline{u})$,

where $\overline{\varphi}$ is the mapping associated to $(\overline{x}, \overline{u})$ by Definition 0.

LEMMA 0. *Let* $0 \leq \tau \leq \tau_1 \leq T$. *Then* $\frac{\partial\overline{\varphi}(\tau_1,\tau)}{\partial x} K(\tau, A) \subset K(\tau_1, A)$.

PROOF: It is sufficient to check the inclusion for the generators of $K(\tau, A)$. Using property (iii) of $\overline{\varphi}$, this is easy. □

The next lemma shows that $K(\overline{x}, \overline{u}, \tau, A, \Sigma)$ is actually a good approximation of the accessibility set of A at $\overline{x}(\tau)$; in its statement, $(\overline{x}, \overline{u}) : [0,T] \to X \times U$ denotes a tame trajectory of Σ.

LEMMA 1. *Let* $h : \omega \to Y$ *be a submersion of an open neighborhood* ω *of* $\overline{x}(\tau)$ *into some manifold* Y. *Given any* C^1 *curve* $\alpha : [0, \varepsilon] \to Y$ *such that* $\alpha(0) = h(\overline{x}(\tau))$ *and* $\frac{d\alpha}{d\lambda}(0)$ *belongs to the interior of the cone* $dh(\overline{x}(\tau))K(\tau, A)$, *there exists an* ε_1, $0 \leq \varepsilon_1 \leq \varepsilon$, *such that* $\alpha([0, \varepsilon_1])$ *is contained in the image, under* h, *of the accessibility set of* A.

PROOF: There exist generators G_j, $1 \leq j \leq m$, of types (iv), (v), (vi) of Definition 1 and positive scalars a_j, $1 \leq j \leq m$, such that:

(vii) The dimension of the linear space generated by the G_i, $1 \leq i \leq m$, is at least equal to that of Y;

(viii) $\frac{d\alpha}{d\lambda}(0) = \sum_{j=1}^{m} a_j dh(\overline{x}(\tau))G_j$.

We can index the G_j as follows: for $1 \leq j \leq r \leq m$, G_j is of type (iv), for $j > r$, G_j is of type (v) or (vi): there exist $t_{r+1} \leq t_{r+2} \leq \cdots \leq t_m$ such that

$$G_j = \frac{\partial \overline{\varphi}(\tau, t_j)}{\partial x}[F(\overline{x}(t_j), u_j) - F(\overline{x}(t_j), \overline{u}(t_j))]$$

or

$$G_j = \varepsilon_j \frac{\partial \overline{\varphi}(\tau, t_j)}{\partial x} F(\overline{x}(t_j), \overline{u}(t_j)).$$

By modifying slightly the a_j and the t_j, we can assume that $t_{r+1} < t_{r+2} < \cdots < t_m$ (see the appendix). Let C be the cube $[0,c]^m$ in $\mathbf{R}^m$. We take $c > 0$ sufficiently small in order that all the subsequent constructions are possible and we define a mapping $C \ni z \mapsto (x(\cdot, z), u(\cdot, z)) \in Tr(A, X)$ as follows: let I be the set of all i such that G_i is of type (vi). Choose any smooth mapping $\xi : [0,c]^r \to A$ such that $\xi(0) = \overline{x}(0)$ and $\frac{\partial \xi}{\partial z_i}(0) = a_i w_i$, $1 \leq i \leq r$, where $G_i = \frac{\partial \overline{\varphi}(\tau, 0)}{\partial x} w_i$. By induction we define functions $S_k : C \to \mathbf{R}$, $r+1 \leq k \leq m$, as follows: $S_m = 0$, $S_{k-1} = S_k$ if $k \notin I$, $S_{k-1} = S_k + \varepsilon_k a_k z_k$ if $k \in I$. Then $(x(\cdot, z), u(\cdot, z))$ is defined on $[0, T(z)]$, where $T(z) = \tau$ if $t_m < \tau$ or $t_m = \tau$ and $m \in I$ and $T(z) = \tau + a_m z_m$ if $t_m = \tau$ and $m \notin I$, as follows:

$$x(0, z) = \xi(z_1, \ldots, z_r),$$

$$\begin{cases} x(t,z) = \overline{\varphi}(t + S_j(z), t_j + S_j(z))x(t_j, z) \\ u(t,z) = \overline{u}(t + S_j(z)) \end{cases}$$

if $j \in I$ and $t_j \leq t \leq t_{j+1}$,

$$\begin{cases} x(t,z) = \exp[(t - t_j)F_{u_j}]x(t_j, z) \\ u(t,z) = u_j \end{cases}$$

if $j \notin I$ and $t_j \leq t \leq t_j + a_j z_j$ and finally

$$\begin{cases} x(t,z) = \overline{\varphi}(t + S_j(z), t_j + a_j z_j + S_j(z))x(t_j + a_j z_j, z) \\ u(t,z) = \overline{u}(t + S_j(z)) \end{cases}$$

if $j \notin I$ and $t_j + a_j z_j \leq t \leq t_{j+1}$.

It is easy to check the following Lemma by induction on j.

LEMMA 2. *The function $C \ni z \mapsto x(\tau, z) \in X$ is differentiable at $z = 0$ and its differential is*

$$\Phi = \sum_{j=1}^{m} a_j G_j \otimes dz_j.$$

To finish the proof of Lemma 1, we need another Lemma, whose proof will be given in the appendix:

LEMMA 3. *Let G be a relatively compact open connected subset of $\mathbf{R}^n$, whose closure (resp. boundary) we denote by $\overline{G}$ (resp. ∂G). Let $f, g : \overline{G} \to \mathbf{R}^n$ be two continuous mappings and let $L(f,g)$ denote the compact subset $\{(1-t)f(x) + tg(x) \mid x \in \partial G, 0 \leq t \leq 1\}$ of $\mathbf{R}^n$. Assume that g is a homeomorphism. Then any point in $g(\overline{G}) - L(f,g)$ belongs to the image of f.*

We apply Lemma 3 to our situation as follows: by the assumption of Lemma 1 there exists a simplex σ of dimension $e = \dim Y$, with one vertex at $0_{h(\overline{x}(\tau))}$, containing $\frac{d\alpha}{d\lambda}(0)$ in its interior and contained in the cone $dh(\overline{x}(\tau))K(\tau, A)$. Since by condition (vii) Φ is surjective, there exists a simplex Δ with one vertex at 0, contained in $\mathbf{R}^n_+ = \{z \mid z_i \geq 0 \text{ for all } i\}$, which is mapped isomorphically onto σ by Φ. As $\overline{G}$ we shall take the simplex $\lambda\Delta = \{\lambda z \mid z \in \Delta\}$, where the value of λ, $\lambda \in (0,1)$, will be choosen later. For λ small enough, $\lambda\Delta$ is contained in C. Choose a C^1 coordinate chart $(\mathcal{O}, \chi)$ of Y at $y(\tau) = h(\overline{x}(\tau))$, $\chi : \mathcal{O} \to T_{y(\tau)}Y$, such that:

(ix) $\chi(y(\tau)) = 0_{y(\tau)}$;

(x) $d\chi(y(\tau)) : T_{y(\tau)}Y \to T_{y(\tau)}Y$ is the identity mapping;

(xi) $\chi \circ \alpha(\lambda) = \frac{d\alpha}{d\lambda}(0)\lambda$.

We choose λ small enough so that the image of the mapping $\lambda\Delta \ni z \mapsto h(x(\tau, z))$ is contained in $\mathcal{O}$. As f we take the composition $z \mapsto \chi(h(x(\tau, z)))$, as g the differential of f at 0,

$$g(z) = \sum_{j=1}^{m} a_j z_j [dh(\overline{x}(\tau))G_j].$$

For any $v \in T_{y(\tau)}Y$, any number $a > 0$, let $\overline{B}(v, a)$ denote the closed ball $\{w \mid w \in T_{y(\tau)}Y, \|w - v\| \leq a\}$, where $\| \ \|$ is any norm on $T_{y(\tau)}Y$. The set $L(f, g)$ is certainly included in $L' = \bigcup\{B(g(z), \|\varphi(z)\|) \mid z \in \partial(\lambda\Delta)\}$. Since $\|\varphi(z)\| / \|z\|$ tends to 0 as $\|z\|$ goes to 0, it is clear that if λ is small enough there will be an $\varepsilon_1 > 0$ such that $g(\eta, \ldots, \eta) \in g(\lambda\Delta) - L'$ for all $\eta \in (0, \varepsilon_1]$. Then condition (xi) above and Lemma 3 imply that $\alpha([0, \varepsilon_1])$ is contained in the image of f. □

§1 The maximum principle in a special case

Consider the following problem. Let $\Sigma = (U, X, F)$ be a system, A, B two submanifolds of X, $k : B \to \mathbf{R}$ a smooth function of B without singular points. Let $(\overline{x}, \overline{u}) : [0, \overline{T}] \to X \times U$ be a tame trajectory of Σ satisfying $\overline{x}(0) \in A$, $\overline{x}(\overline{T}) \in B$ and $(\overline{x}, \overline{u})$ is optimal with respect to the cost $k(\overline{x}(\overline{T}))$ among all trajectories $(x, u) : [0, T_u] \to X \times U$ of Σ such that $x(0) \in A$, $x(T_u) \in B$. Then we have the following necessary condition on $(\overline{x}, \overline{u})$. Let $H : T^*X \times U \to \mathbf{R}$ be the function $H(p, u) = \langle p, F(x, u)\rangle$, x projection of p on X.

THEOREM 0. *There exists an absolutely continuous lifting $\overline{p} : [0, \overline{T}] \to T^*X$ of $\overline{x}$ satisfying the following conditions:*

(xii) for almost all $t \in [0, \overline{T}]$, $\frac{d\overline{p}}{dt}(t) = \vec{H}(\overline{p}(t), \overline{u}(t))$, where $\vec{H}(\cdot, \overline{u}(t))$ is the Hamiltonian field of the function $H(\cdot, \overline{u}(t))$;

(xiii) for almost all $t \in [0, \overline{T}]$, $H(\overline{p}(t), \overline{u}(t)) = \sup_{u \in U} H(p(t), u)$;

(xiv) $\overline{p}(0)$ belongs to the annihilator of $T_{\overline{x}(0)}A$;

(xv) the restriction $\overline{p}(\overline{T}) \mid T_{\overline{x}(\overline{T})}B = \mu dk(\overline{x}(\overline{T}))$, where μ is a non-negative scalar.

PROOF: Let m be the codimension of B at $\overline{x}(\overline{T})$ plus 1. Then in an open neighborhood ω of $\overline{x}(\overline{T})$ in X, there exists a submersion $\tilde{h} = (h_1, \ldots, h_{m-1}) : \omega \to \mathbf{R}^{m-1}$ such that $B \cap \omega = \tilde{h}^{-1}(0)$. Replacing ω by a smaller neighborhood of $\overline{x}(\overline{T})$ if need be, we can extend k to a smooth function $\tilde{k} : \omega \to \mathbf{R}$ such that the mapping $h = \tilde{h} \times \tilde{k} : \omega \to \mathbf{R}^m = \mathbf{R}^{m-1} \times \mathbf{R}$ is a submersion.

We shall apply Lemma 1, taking $\mathbf{R}^m$ as Y, $\overline{T}$ as τ and as α, the curve $\alpha : [0, +\infty) \to Y$, $\alpha(\lambda) = (0, \ldots, 0, -\lambda + k(\overline{x}(\overline{T})))$. The image of α is the translation $(0, \ldots, 0, k(\overline{x}(\overline{T}))) + \ell$ of the half line $\ell = \{(0, \ldots, 0, t) \mid t \leq 0\}$. Lemma 1 says that if ℓ were contained the interior of the cone $dh(\overline{x}(\overline{T}))K(\overline{T}, A)$, then for any $\eta > 0$, small enough, there would exist a trajectory $(x_\eta, u_\eta) : [0, T_\eta] \to X \times U$ of Σ such that $x_\eta(0) \in A$ and $h(x_\eta(T_\eta)) = \alpha(\eta)$; but then $\tilde{h}(x_\eta(T_\eta)) = 0$, $x_\eta(T_\eta)$ belongs to B and $k(x_\eta(T_\eta)) < k(\overline{x}(\overline{T}))$: $(\overline{x}, \overline{u})$ would not be optimal. Hence ℓ is not contained in the interior of $dh(\overline{x}(\overline{T}))K(\overline{T}, A)$. The Hahn-Banach separation theorem tells us that there exists a $q \in (\mathbf{R}^m)^*$ (dual of $\mathbf{R}^m$), such that $\langle q, w\rangle \geq 0$ for all $w \in dh(\overline{x}(\overline{T}))K(\overline{T}, A)$ and $\langle q, w\rangle \leq 0$ for all $w \in \ell$.

Let $p_{\overline{T}}$ be the element $q \circ dh(\overline{x}(\overline{T}))$ in $T^*_{\overline{x}(\overline{T})}X$. The above shows that $\langle p_{\overline{T}}, G\rangle \geq 0$ for any element G of $K(\overline{T}, A)$. Let us denote by $\overline{p} : [0, \overline{T}] \to T^*X$ the function $t \mapsto p_{\overline{T}} \circ \frac{\partial \overline{\varphi}(\overline{T}, t)}{\partial x}$. Taking G of type (iv), (v), (vi) respectively, we get:

(xvi) $\langle \overline{p}(0), w\rangle \geq 0$ for any $w \in T_{\overline{x}(0)}A$;

(xvii) $\langle \overline{p}(t), F(\overline{x}(t), v)\rangle \geq \langle p(t), F(\overline{x}(t), \overline{u}(t))\rangle$ for any $v \in U$ and any Lebesgue point t of $(\overline{x}, \overline{u})$;

(xviii) $\varepsilon \langle \overline{p}(t), F(\overline{x}(t), \overline{u}(t))\rangle \geq 0$ for any $\varepsilon = +$ or $-$ and any Lebesgue point t of $(\overline{x}, \overline{u})$.

(xvi) shows that $\overline{p}(0)$ annihilates $T_{\overline{x}(0)}A$. (xviii) implies that $\langle \overline{p}(t), F(\overline{x}(t), \overline{u}(t))\rangle = 0$ for any Lebesgue point t. (xvii) and (xviii) imply that $\langle \overline{p}(t), F(\overline{x}(t), v)\rangle \geq 0$ for any $v \in U$ and any Lebesgue point t. Hence we have proved conditions (xiii), (xiv) of the Theorem.

The definition of $\overline{p}(t)$ itself shows that $\overline{p}(t)$ satisfies (xii). Since $\langle q, w\rangle \leq 0$ for all $w \in \ell$, $\langle \overline{p}(\overline{T}), w\rangle = \langle p_{\overline{T}}, w\rangle \leq 0$ for all $w \in T_{\overline{x}(\overline{T})}B$ such that $dk(\overline{x}(\overline{T}))w \leq 0$. This shows that the restriction of $\overline{p}(\overline{T})$ to $T_{\overline{x}(\overline{T})}B$ is a non-negative multiple of $dk(\overline{x}(\overline{T}))$ and finishes the proof of Theorem 0. □

§2 The general maximal principle

Let us come back to the general optimal control problem stated in the introduction to Chapter VI. We can reduce this problem to the one stated in §1 above as follows. First we extend the state space X to $\tilde{X} = \mathbf{R} \times X \times \mathbf{R}$ and the system F to $\tilde{F} = (a, F, c)$. More explicitely, if $d = \dim X$, $\overline{x} \in \tilde{X}$ is a triple (x_0, x, x_{d+1}), $x_0, x_{d+1} \in \mathbf{R}$,

$$\frac{dx_0}{dt} = 1, \quad \frac{dx}{dt} = F(x,u), \quad \frac{dx_{d+1}}{dt} = c(x,u).$$

A is replaced by $\tilde{A} = \{0\} \times A \times \{0\}$, B by $\tilde{B} = \mathbf{R} \times B \times \mathbf{R}$ and the cost by a final cost $k : \tilde{B} \to \mathbf{R}$ defined by $k(x_0, x, x_{d+1}) = x_{d+1} + \gamma(x_0, x)$. (Recall that the initial cost was $\int_0^{T_u} c(x(t), u(t))dt + \gamma(T_u, x(T_u)))$. It is clear that k has no singular points on $\tilde{B}$: $\frac{\partial k}{\partial x_{d+1}} = 1$.

We can apply Theorem 0 of §1 above and interpret the results in our present setting. $\tilde{p}(t) = (\overline{p}_0(t), \overline{p}'(t), \overline{p}_{d+1}(t))$, $H(p_0, p, p_{d+1}, u) = p_0 + \langle p, F(x,u)\rangle + p_{d+1}c(x,u)$, where x is the projection of p on X.

Then (xii) tells us that $\frac{d\overline{p}_0}{dt} = \frac{d\overline{p}_{d+1}}{dt} = 0$. $\overline{p}_0, \overline{p}_{d+1}$ are constants. Let $\overline{p} : [0, \overline{T}] \to T^*X$ be defined as follows: if $\overline{p}_{d+1} \neq 0$, $\overline{p}(t) = \frac{1}{p_{d+1}}\overline{p}'(t)$; if $p_{d+1} = 0$, $\overline{p}(t) = \overline{p}'(t)$. Set $\lambda = 0$ if $\overline{p}_{d+1} = 0$, $\lambda = -\frac{\overline{p}_{d+1}}{|\overline{p}_{d+1}|}$ if $p_{d+1} \neq 0$. Then the last equation of (xii) is: $\frac{d\overline{p}}{dt}(t) = \vec{H}(\overline{p}(t), \overline{u}(t))$ for almost all $t \in [0, \overline{T}]$, where $H_\lambda(p,u) = \langle p, F(x,u)\rangle - \lambda c(x,u)$. Since p_0 is constant, (xiii) is equivalent to:

$$H_\lambda(\overline{p}(t), \overline{u}(t)) = \sup\{H_\lambda(\overline{p}(t), u) \mid u \in U\} = 0.$$

(xiv) says that $\overline{p}(0)$ is in the annihilator of $T_{\overline{x}(0)}A$. Finally (xv) states the the restriction of $p_0 dx_0 + \langle \overline{p}(\overline{T}), dx \rangle + \overline{p}_{d+1} dx_{d+1}$ to $T_{\overline{x}(\overline{T})}\tilde{B}$ is equal to $\mu dx_{d+1} + \mu \frac{\partial \gamma}{\partial x}(\overline{T}, \overline{x}(\overline{T}))dx + \mu \frac{\partial \gamma}{\partial T}(\overline{T}, \overline{x}(\overline{T}))dx_0$. Therefore $\overline{p}_{d+1} = \mu$, and is positive. This shows that the scalar λ is 0 or -1.

We have proved all the assertions of Theorem 0 of Chapter VI.

APPENDIX TO CHAPTER VII

PROOF OF LEMMA 3: Define for every $t \in [0,1]$ a continuous mapping $f_t: \overline{G} \to \mathbf{R}^n$ as follows: $f_t(x) = (1-t)g(x) + tf(x)$. Then $f_1 = f$, $f_0 = g$ and $f_t(\partial G) \subset L(f,g)$ for any $t \in [0,1]$. Let $y \in g(\overline{G}) - L(f,g)$. Then $y \in g(G)$ and $\mathbf{R}^n - \{y\} \supset L(f,g)$. Since $L(f,g)$ is compact, it is easy to find a compact neighborhood Ω of ∂G such that $\partial\Omega \cap G$ is a manifold ($\partial\Omega$ is the boundary of Ω) and $f_t(\Omega \cap \overline{G}) \subset \mathbf{R}^n - \{y\}$ for all $t \in [0,1]$.

The f_t induce homomorphisms of the n^{th} singular homology groups of pairs: $f_{t_*} : H_x(\overline{G}, \Omega \cap \overline{G}) \to H_x(\mathbf{R}^n, \mathbf{R}^n - \{y\})$.

Since $t \mapsto f_t$ is a homotopy of mappings of pairs $(\overline{G}, \Omega \cap \overline{G}) \to (\mathbf{R}^n, \mathbf{R}^n - \{y\})$, it follows that $f_{0*} = f_{1*}$. Hence $f_* = g_*$. Now g_* can be factored as follows:

$$
\begin{array}{ccc}
H_n(\overline{G}, \Omega \cap \overline{G}) & \xrightarrow{g_*} & H_n(\mathbf{R}^n, \mathbf{R}^n - \{y\}) \\
\alpha_1 \nearrow & & \\
H_n(\overline{G - \Omega}, \partial(G - \Omega)) & & \uparrow \gamma \\
\alpha_2 \searrow & & \\
H_n(\overline{g(G-\Omega)}), \partial\overline{g(G-\Omega)}) & \underset{\beta}{\longrightarrow} & H_n(\overline{g(G-\Omega)}, \overline{g(G-\Omega)} - \{y\})
\end{array}
$$

α_1 is an excision of $\Omega \cap G$, hence an isomorphism. α_2 is induced by g and is an isomorphism. γ is an excision of $\mathbf{R}^n - g(G - \Omega)$ and an isomorphism. β is onto since $\overline{g(G-\Omega)}$ is a topological manifold with boundary and y is an interior point. This shows that g_* is onto. If y did not belong to the image of f, f_* would be zero and g_* too, since $g_* = f_*$. But g_* is onto and then $H_n(\mathbf{R}^n, \mathbf{R}^n - \{y\})$ would be zero; since it is isomorphic to $\mathbf{Z}$, we have reached a contradiction.

□

APPROXIMATE CONTINUITY: A measurable function $\varphi : [0,T] \to Z$ of the interval $[0,T]$ into some topological space Z is said to be approximately continuous at $\tau \in [0,T]$ if for any neighborhood V of $\varphi(\tau)$ in Z, the set of all $t \in [0,T]$ with $\varphi(t) \in V$ has density 1 at τ. The set of all these points has full measure. If Z is a vector space, and φ is integrable, then any primitive of φ is differentiable at τ.

Given any point of approximate continuity τ of φ, any open interval J containing τ and any neighborhood V of $\varphi(\tau)$, there exists a point $t \in J$ of approximate continuity of φ such that $t \neq \tau$, and $\varphi(t) \in V$.

A function can be differentiable at a point which is not a point of approximate continuity for the derivative. Define $\varphi_1 : [-1, +1] \to \mathbf{R}$ as follows:

$$\varphi_1(t) = \begin{cases} 1, & \frac{1}{2k} > t \geq \frac{1}{2k+1} \text{ or } \frac{-1}{2k+1} > t \geq -\frac{1}{2k}, k \geq 1 \\ -1, & \frac{1}{2k+1} > t \geq \frac{1}{2k+2} \text{ or } \frac{-1}{2k+2} > t \geq -\frac{1}{2k+1}, k \geq 0 \\ 0, & t = 0. \end{cases}$$

φ_1 is integrable. Let $\varphi(t) = \int_0^t \varphi_1(s)ds$. Then 0 is not an approximate continuity point of φ_1 but φ is differentiable at 0 and its derivate is 0 there.

REFERENCES

[B] B. BONNARD, Contrôlabilité des systèmes non lineaires, C.R.A.S. (Series I) 292 (1981), 535–537.

[Br] R.W. BROCKETT, "Linear systems", Addison-Wesley (1960).

[C] F.H. CLARKE, "Optimization and nonsmooth analysis", John Wiley & Sons (1983).

[JK] V. JURDJEVIC and I. KUPKA, Control systems on semi-simple Lie groups and their homogeneous spaces, Ann. Inst. Four. 31, 4 (1981), 151–179.

[K] T. KAILATH, "Linear systems", Prentice-Hall (1980).

[LM] E.B. LEE and L. MARKUS, "Foundations of optimal control theory", J. Wiley & Sons (1967).

[L] C. LOBRY, Contrôlabilité des systèmes non lineaires, SIAM Jour. Control 8 (1970), 573–605.

[L] C. LOBRY, Controllability of nonlinear systems on compact manifolds, SIAM Jour. Control 12 (1974), 1–4.

[PBGM] L.S. PONTRYAGIN et al, "The mathematical theory of optimal processes", John Wiley & Sons (1962).

[Sh] M. SHAYMAN, Geometry of the algebraic Riccati equation, SIAM Jour. Control and Optimization 21 (1983) part I: 375–394; part II: 395–409.

[S1] H.J. SUSSMANN, A generalization of the closed subgroup theorem to quotients of arbitrary manifolds, J. Diff. Geom. 10 (1975), 151–166.

[S2] H.J. SUSSMANN, Lie brackets, real analyticity and geometric control in "Differential geometric control theory", ed. R.W. Brockett, R.S. Millman and H.J. Sussmann, Birkhaüser PM 27 (1983), 1–116.

[S3] H.J. SUSSMANN, "Orbits of families of vector fields and integrability of distributions". Trans. Amer. Math. Soc. 180 (1973) pp. 171- 188.

[T] A. TANNENBAUM, "Invariance and system theory: algebraic and geometric aspects", Springer Lect. Notes in Math. 845 (1981).

Impressão Gráfica Wagner