ON BANG-BANG CONTROL POLICIES

Roberto GONZALEZ (*) and Edmundo ROFMAN (**)
Instituto de Matemática "Beppo Levi"
Universidad Nacional de Rosario
ARGENTINA

(Work included in the cooperation program with I.R.I.A. — Rocquencourt — FRANCE)

ABSTRACT: In this paper it is proposed a method for the determination of the optimal distribution of N switching points for a bang-bang control applied to a differential system.

After pointing the necessary conditions to be verified by such switching points it is showed the existence of an optimal policy for a fixed number N of them.

Once characterized these points thru the application of the Pontryagin principle the problem, considered till now in the space of step functions, is put into the L_1 space, in order to show the existence of a minimizing succession of the amplified problem and analized its correspondence to an optimal policy.

After reducing the problem into one of optimization on a convex K of $\mathbb{R}^n$ there are added considerations which let us, with the proposed method, obtain the optimal also with a number of switching points n less than the predetermined N .

Now it is proved that the function to optimize is of C^2 class in K and the applied methods are these of the projected gradient and the conjugated gradient conveniently penalized.

Finally, the obtained algorithms are applied in one example: the shut down policy of a nuclear reactor where the optimun is obtained with a finite number of switchings; this number remains constant although increasing values of N are proposed.

§1. STATEMENT OF THE PROBLEM. NECESSARY CONDITIONS OF OPTIMALITY. EXISTENCE OF MINIMUN IN THE CASE OF FIXED NUMBER OF SWITCHING POINTS.

Given the dinamical system governed by the differential equation:

$$\dot{x} = F(t)x + G(t)u \qquad x \in R^{\upsilon} \ , \ u \in R^{1} \tag{1}$$

with initial condition $x(0) = x_0$

and the cost functional

$$J(u(.)) = \int_0^T l(x(s), u(s), s)ds + g(x(T)) \tag{2}$$

we try to find the control function $u(.)$ that minimices J .

(*) Reasercher of the "Consejo de Investigaciones de la Universidad Nacional de Rosario" for the project: "Optimization and Control. Theory and applications".

(**) Director of the above referred project.

The control $u(.)$ belongs to the family $\mathcal{U}_{ad}$ that satisfy the following restrictions.

a) $u(t) = v_1$ or $u(t) = v_2$ $\quad \forall\, t \in [0,T]$

b) $u(.)$ is a step-function with n switchings

c) $u(0) = v_1$.

We denote with $\theta_1, \theta_2, \ldots, \theta_n$ the switching points and this set with the vector $\theta = (\theta_1, \ldots, \theta_n)'$

(3) θ satisfies the restrictions: $0 < \theta_1 < \theta_2 < \ldots < \theta_n < T$.

Then, if we fix θ , we know the value of $u(t)$ $\forall\, t \in [0,T]$ and we can think of $J(u(.))$ as a function $J(\theta)$ of $\theta \in \mathring{\Omega}$, where $\mathring{\Omega} = \{\theta \in R^n \,/\, 0 < \theta_1 < \ldots < \theta_n < T\}$. If the minimun of the problem exists, the necessary conditions will be (because θ belongs to an open set):

$$\frac{\partial J}{\partial \theta_1} = 0 \quad ; \ldots; \quad \frac{\partial J}{\partial \theta_n} = 0 \quad .$$

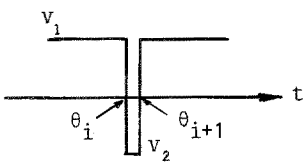
We can modify the restrictions (3) in the following form:

(3') $$0 \leq \theta_1 \leq \theta_2 \leq \ldots \leq \theta_n \leq T$$

and analyze the meaning of a point in the boundary of $\mathring{\Omega}$.

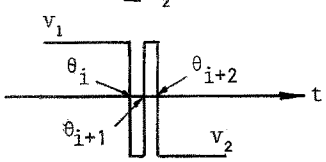
a) $\theta_1 = 0$ means that the first step has the value $u = v_2$.

b) $\theta_i = \theta_{i+1}$



there are two simultaneous commutations that could be eliminated and it remains a new control function with n-2 commutations.

c) $\theta_i = \theta_{i+1} = \theta_{i+2}$



there are three simultaneous switchings, we could eliminate two and obtain a new control with n-2 commutations.

d) $\theta_n = T$ a commutation at the end that could be eliminated and the new control has n-1 switchings.

With this meaning we can define the function $J(\theta)$ in the set $\Omega = \{\theta \in R^n / 0 \leq \leq \theta_1 \leq \theta_2 \leq \ldots \leq \theta_n \leq T\}$. In this compact set, under suitable conditions on 1, g, F,G, (it will be enough the continuity of 1 and g, and that F,G be integrable), is $J(\theta)$ continuos; then there is an optimal control in Ω that provides the minimun value of J and has $n' \leq n$ switching points.

§2. THE RELATION BETWEEN THE NECESSARY CONDITION $\frac{\partial J}{\partial \theta_i} = 0$ AND PONTRYAGIN'S MAXIMUN PRINCIPLE.

We shall see in the following number that $\frac{\partial J}{\partial \theta_i}$ has the form:

$$\frac{\partial J}{\partial \theta_i} = 1(x(\theta_i), u(\theta_i -), \theta_i) - 1(x(\theta_i), u(\theta_i +), \theta_i) + p(\theta_i) G(\theta_i) \left[u(\theta_i +) - u(\theta_i -) \right]$$

where $p(t)$ satisfies:

$$\begin{cases} \dfrac{dp}{dt} = - p\,F(t) + \dfrac{\partial l}{\partial x}(x(t), u(t), t) \\ p(T) = - \dfrac{\partial g}{\partial x}\Big|_{x(T)} \end{cases}$$

and $u(\theta_i^-) = \lim_{\substack{\varepsilon \to 0 \\ \varepsilon > 0}} u(\theta_i - \varepsilon)$; $u(\theta_i^+) = \lim_{\substack{\varepsilon \to 0 \\ \varepsilon > 0}} u(\theta_i + \varepsilon)$.

If we define:

$$H(x,p,u,t) = p(t)\left[F(t)x(t) + G(t)u(t)\right] - l(x(t),u(t),t)$$

we can write:

$$\frac{\partial J}{\partial \theta_i} = H(x(\theta_i),p(\theta_i),u(\theta_i^+),\theta_i) - H(x(\theta_i),p(\theta_i),u(\theta_i^-),\theta_i) \quad .$$

We also know that in the problem $\min_{u \in \mathcal{U}_{ad}} J(u(.))$, where $\mathcal{U}_{ad}$ is the set of step functions with values v_1, v_2, if $\bar{u}(.)$ is the optimal control in this set, it must satisfie the maximun principle:

$$H(x(t),p(t),\bar{u}(t),t) = M(x(t),p(t),t) \qquad \text{a.e.}$$

where, by definition

$$M(x,p,t) = \max_{u=\{u_1,u_2} H(x,p,u,t)$$

$M(x(t), p(t), t)$ is continuous in t , then:

$$\lim_{\varepsilon \to 0+} M(x(t+\varepsilon), p(t+\varepsilon), t) = \lim_{\varepsilon \to 0-} M(x(t+\varepsilon), p(t+\varepsilon), t+\varepsilon)$$

but

$$\lim_{\varepsilon \to 0+} M(x(t+\varepsilon),p(t+\varepsilon),t+\varepsilon) = \lim_{\varepsilon \to 0+} H(x(t+\varepsilon),p(t+\varepsilon),u(t+\varepsilon),t+\varepsilon) = H(x(t),p(t),u(t+),t)$$

and also

$$\lim_{\varepsilon \to 0-} M(x(t+\varepsilon),p(t+\varepsilon),t+\varepsilon) = \lim_{\varepsilon \to 0-} H(x(t+\varepsilon),p(t+\varepsilon),u(t+\varepsilon),t+\varepsilon) = H(x(t), p(t), u(t-), t)$$

from where it follows:

$$\frac{\partial J}{\partial \theta_i} = H(x(\theta_i),p(\theta_i),u(\theta_i^+),\theta_i) - H(x(\theta_i),p(\theta_i),u(\theta_i^-),\theta_i) = 0$$

then, the maximun principle implies the necesary conditions of optimality: $\dfrac{\partial J}{\partial \theta_i} = 0$.

§3. COMPUTATION $\dfrac{\partial J}{\partial \theta_i}$

The equation of the system's evolution is

$$x(t) = \Phi(t,0)x_o + \sum_{k=1}^{j-1} \int_{\theta_k}^{\theta_{k+1}} \Phi(t,s)G(s)u(s)ds + \int_{\theta_j}^{t} \Phi(t,s)G(s)u(s)ds$$

where $j \;/\; \theta_j \leq t \leq \theta_{j+1}$

and we can calculate $\dfrac{\partial}{\partial \theta_i} x(t)$.

We suppose $G(t)$ is continuous in the interval $[0,T]$

a) $t > \theta_i$

$$\frac{\partial}{\partial\theta_i}x(t) = \Phi(t,\theta_i)G(\theta_i)(v_2 - v_1)(-1)^i \quad .$$

It must be remembered that $\Phi(t,s)$ is the solution of the matrix differential equation

$$\frac{d}{dt}\Phi(t,s) = F(t)\Phi(t,s) \qquad \Phi\ \nu\times\nu \quad \text{matrix}$$

with initial condition: $\Phi(s,s) = I$.

b) $t < \theta_i$

$x(t)$ does not depend on θ_i , then $\frac{\partial}{\partial\theta_i}x(t) = 0$.

In this form, we can say that

$$\begin{cases} \frac{d}{dt}\left(\frac{\partial}{\partial\theta_i}x(t)\right) = F(t)\left(\frac{\partial}{\partial\theta_i}x(t)\right) & t > \theta_i \\ \text{with initial conditions:} \\ \frac{\partial}{\partial\theta_i}x(t)\Big|_{t=\theta_i} = G(\theta_i)(v_2 - v_1)(-1)^i \end{cases}$$

$$J(\theta) = g(x(T)) + \int_0^T 1(x(s), u(s), s)d\tau$$

if 1 and g are continuously differentiable, we have:

$$\frac{\partial J}{\partial\theta_i} = 1(x(\theta_i), u(\theta_i-), \theta_i) - 1(x(\theta_i), u(\theta_i+), \theta_i) + g'(x(T)).\frac{\partial}{\partial\theta_i}x(T) + \int_{\theta_i}^T \frac{\partial 1}{\partial x}(x(s),u(s),s).\frac{\partial}{\partial\theta_i}x(s)ds$$

if we introduce the adjoint vector $p(t)$ that satisfies

$$\begin{cases} -\frac{dp}{dt}(t) = p(t)F(t) - \frac{\partial 1}{\partial x}(x(t),u(t),t) \\ p(T) = -g'(x(T)) \end{cases}$$

we can write:

$$\frac{\partial J}{\partial\theta_i} = 1(x(\theta_i), u(\theta_i-), \theta_i) - 1(x(\theta_i), u(\theta_i+), \theta_i) + g'(x(T))\frac{\partial}{\partial\theta_i}x(T) + \int_{\theta_i}^T \left(p(t)F(t) + \frac{dp}{dt}(t)\right)\frac{\partial}{\partial\theta_i}x(t)dt$$

and integrating by parts we obtain:

$$\frac{\partial J}{\partial\theta_i} = 1(x(\theta_i),u(\theta_i-),\theta_i) - 1(x(\theta_i),u(\theta_i+) + p(\theta_i)G(\theta_i)(v_1 - v_2)(-1)^i$$

§4. CONTINUITY OF $\frac{\partial J}{\partial\theta_i}$

(1) $$\frac{\partial J}{\partial\theta_i} = 1(x(\theta_i),u(\theta_i-),\theta_i) - 1(x(\theta_i),u(\theta_i+),\theta_i) + p(\theta_i)G(\theta_i)(v_1 - v_2)(-1)^i \quad .$$

We suppose that $F(.), G(.)$ are continuous and $1,g$ are continuously differentiable. $\frac{\partial J}{\partial\theta_i}$ is a continuous function of $x(\theta_i), p(\theta_i)$ and $\theta_i)$. The values $u(\theta_i+)$ and $u(\theta_i-)$ are constant $(v_1$ and $v_2)$, then we must prove only the continuity of $x(\theta_i)$, $p(\theta_i)$ to obtain the continuity of $\frac{\partial J}{\partial\theta_i}$.

We can easily see that the transformation $\theta : \longrightarrow u(.)$ defined by

$$(2)\qquad \begin{cases} u(t) = v_1 & 0 \leq t < \theta_1 \\ u(t) = v_1 + (v_2 - v_1)\left[\frac{1 - (-1)^i}{2}\right] & \theta_i \leq t < \theta_{i+1} \\ u(t) = v_1 + (v_2 - v_1)\left[\frac{1 - (-1)^n}{2}\right] & \theta_n \leq t \leq T \end{cases}$$

is continuous from $\Omega \longrightarrow L_1(0,T)$.

$$(3)\qquad x(t) = \Phi(t,0)x_o + \int_0^t \Phi(t,s)G(s)u(s)ds$$

and this formula defines a continuous transformation from $L_1(0,T) \longrightarrow C(0,T;R^\nu)$ because, if $u_1(.)$, $u_2(.)$ are two controls in $L_1(0,T)$ and $x_1(.)$, $x_2(.)$ the system's evolution, it is:

$$(4)\qquad \|x_1(t) - x_2(t)\| \leq M \int_0^T |u_1(s) - u_2(s)|ds = M\|u_1 - u_2\|_{L_1(0,T)}$$

where

$$(5)\qquad M = \sup_{\substack{t \in [0,T] \\ s \in [0,t]}} \|\Phi(t,s)\| \cdot \sup_{s \in [0,T]} \|G(s)\|$$

and M is finite due to the continuity of G and Φ .

Obviously

$$(6)\qquad x(.) \longrightarrow x(\theta_i) \text{ is continuous from } C(0,T;R^\nu) \longrightarrow R^\nu \quad .$$

Then,

$$(7)\qquad \theta \longrightarrow x(\theta_i) \text{ is continuous from } R^\nu \longrightarrow R^\nu \quad .$$

From the differential equation of $p(t)$ (p is a row vector)

$$(8)\qquad -\frac{d}{dt}p(t) = p(t)F(t) - \frac{\partial l}{\partial x}$$

we obtain the complete solution using the homogeneous solution and "variation of constants" method

$$(9)\qquad p(t) = p(T)\Phi(T,t) + \int_T^t \frac{\partial l}{\partial x}(x(s),u(s),s) \cdot \Phi(s,t)ds$$

$$(10)\qquad p(T) = -g'(x(T))$$

thus, $-g'(x(T))\Phi(T,\theta_i)$ defines a continuous function from $\theta \in R^n \longrightarrow R^\nu$, because g' is continuous, $x(T)$ is a continuous function of θ and Φ is absolutely continuous in its both arguments. The transformation $R^n \longrightarrow L_1(0,T;R^\nu)$ given by:

$$\theta \longrightarrow u(.) \longrightarrow \frac{\partial l}{\partial x}(x(t),u(t),t) \in L_1(0,T;R^\nu)$$
$$\searrow\ x(.)\ \nearrow$$

is continuous from $R^n \longrightarrow L_1(0,T;R^\nu)$.

To prove it, let $\theta, \theta_\varepsilon$ be two set of switching points such that $\|\theta - \theta_\varepsilon\| \longrightarrow 0 \quad \varepsilon \to 0$.

Then, $u(t) \longrightarrow u_\varepsilon(t)$ a.e.

and $\max_{t \in [0,T]} \|x(t) - x_\varepsilon(t)\| \longrightarrow 0$

now, $\frac{\partial l}{\partial x}(x_\varepsilon(s),u_\varepsilon(s),s) \longrightarrow \frac{\partial l}{\partial x}(x(s),u(s),s)$ a.e. , and

$\int_0^T \|\frac{\partial l}{\partial x}(x_\varepsilon(s),u_\varepsilon(s),s) - \frac{\partial l}{\partial x}(x(s),u(s),s)\| ds \longrightarrow 0$ as we can see applying Lebesgue's theorem.

The formula $\int_T^t \frac{\partial I}{\partial x}(x(s),u(s),s)\Phi(s,t)ds$ defines a continuous transformation from $L_1(0,T;R^\nu) \longrightarrow C(0,T;R^\nu)$. Then, taking into account all these results we proof that $p(\theta_i)$ and also $\frac{\partial J}{\partial \theta_i}$ are continuous functions of θ.

§5. NUMERICAL SOLUTION OF THE PROBLEM BY THE APPLICATION OF THE PROJECTED GRADIENT METHOD.

We have seen that the problem of finding an optimal bang–bang policy with $n' \leq n$ switchings was reduced to the finite dimensional problem.

$$(1) \qquad \min_{\Omega} J(\theta) \qquad \Omega = \left\{\theta \in R^n \;/\; 0 \leq \theta_1 \leq \theta_2 \leq \ldots \leq \theta_n \leq T\right\}$$

Ω is a convex and compact subset where J is continuous, and this implicates the existence of a minimun.

We write the $n+1$ restrictions defining Ω in vector form.

$f(\theta) \leq 0 \qquad f \in R^{n+1}$ and $f_1 = -\theta_1, f_2 = \theta_1 - \theta_2, \ldots, f_n = \theta_{n-1} - \theta_n, f_{n+1} = \theta_n - T$.

DEFINITION.

$$I_\varepsilon(\theta) = \{i \;/\; f_i(\theta) + \varepsilon \geq 0\}.$$

DEFINITION.

Given a set of integers, $I \subset \{1,\ldots, n+1\}$, the projection of $y \in R^n$ on the subspace generated by the vectors ∇f_i, $i \in I$ is the vector $F_I \bar{\mu}$ that minimizes $\|y - F_I \mu\|^2$, where

$$F_I = \left[\nabla f_{i_1}, \ldots, \nabla f_{i_m}\right], \quad I = \left\{i_1, \ldots, i_m\right\} \text{ and } \mu \in R^m.$$

It is easily shown that

$$\bar{\mu} = \left(F_I' F_I\right)^{-1} F_I' y \qquad (F' \text{ is transpose of } F).$$

Then, the projection of y is:

$$P_I y = F_I\left(F_I' F_I\right)^{-1} F_I' y$$

and we can define the projection matrix

$$P_I' = F_I\left(F_I' F\right)^{-1} F_I'.$$

In the same form we define the projection on the subspace orthogonal to ∇f_i, $i \in I$ and the corresponding matrix is:

$$P_I^{\perp} = I - P_I.$$

In the definition of P_I, we have supposed that the vectors ∇f_i, $i \in I$ are linearly independent, and then the matrix $F_I' F_I$ is invertible.

It is known (Kuhn–Tucker's theorem) that if $\bar{\theta}$ is a solution of the problem (1), then

$$\nabla J(\bar{\theta}) = F_{I_o(\bar{\theta})}\,\bar{\mu} \qquad \bar{\mu}_1 \leq 0, \ldots, \bar{\mu}_m \leq 0$$

and

$$\bar{\mu} = \left(F_{I_o(\bar{\theta})}' F_{I_o(\bar{\theta})}\right)^{-1} F_{I_o(\bar{\theta})}' \nabla J(\bar{\theta})$$

We define a point as desirable if:

a) $\theta \in \Omega$ b) $\nabla J(\theta) = F_{I_o(\theta)}\,\mu(\theta) \qquad \mu(\theta) \leq 0$.

It is posible to apply the following algorithm: "Gradient Projected".

ALGORITHM.

Step 0 : Select $\theta_o \in \Omega$;
$\varepsilon' > 0 \ / \ \forall\, \varepsilon > 0 \ , \ \varepsilon \leq \varepsilon' \quad \left\{\nabla f_i(\theta) \ / \ i \in I_\varepsilon(\theta), \ \theta \in \Omega\right\}$ is a set of linearly independent vectors.
Choose $\beta \in (0,1)$, $\bar{\varepsilon} \in (0,\varepsilon')$, $\varepsilon'' \in (0,\bar{\varepsilon})$.
Set $i = 0$.

Step 1 : Set $\theta = \theta_i$.

Step 2 : Set $\varepsilon_o = \bar{\varepsilon}$ and $j = 0$.

Step 3 : Compute $h_{\varepsilon_j} = P^{\perp}_{I_{\varepsilon_j}(\theta)} \nabla J(\theta)$.

Step 4 : If $\|h_{\varepsilon_j}\| > \varepsilon_j$ $h(\theta) = - h_{\varepsilon_j}$ and go to step 12; else, go to 5.

Step 5 : If $\varepsilon_j \leq \varepsilon''$, compute $h_o(\theta) = P^{\perp}_{I_o(\theta)} \nabla J(\theta)$ and
$\mu_o(\theta) = \left(F'_{I_o(\theta)} F_{I_o(\theta)}\right)^{-1} F'_{I_o(\theta)} \nabla J(\theta)$ and go to 6; else, go to 7 .

Step 6 : If $\mu_o(\theta) \leq 0$ and $\|h_o(\theta)\| = 0$ set $\theta_{i+1} = \theta$ and stop, else, go to 7.

Step 7 : Compute $\mu_{\varepsilon_j}(\theta) = \left(F'_{I_{\varepsilon_j}(\theta)} F_{I_{\varepsilon_j}(\theta)}\right)^{-1} F'_{I_{\varepsilon_j}(\theta)} \nabla J(\theta)$.

Step 8 : If $\mu_{\varepsilon_j}(\theta) \leq 0$, set $\varepsilon_{j+1} = \beta \varepsilon_j$, set $j = j+1$ and go to step 3 ; else, go to step 9 .

Step 9 : Assuming that $I_{\varepsilon_j}(\theta) = \left\{k_1, \ldots, k_{m'}\right\}$ and that $k_1 < k_2 < \ldots < k_{m'}$, set $y^{k_\alpha}_{\varepsilon_j}(\theta) = \mu^{\alpha}_{\varepsilon_j}(\theta)$ for $\alpha = 1,2,\ldots, m'$ (where $\mu^{\alpha}_{\varepsilon_j}(\theta)$ is the α^{th} component of the vector $\mu_{\varepsilon_j}(\theta)$) .

Step 10 : Find the smallest $k \in I_{\varepsilon_j}(\theta)$ such that the vector $\bar{h}_{\varepsilon_j}(\theta) = P^{\perp}_{I_{\varepsilon_j}(\theta) - k} \nabla J(\theta)$ satisfies the relation

$$\|\bar{h}_{\varepsilon_j}(\theta)\| = \max\left\{\|P^{\perp}_{I_{\varepsilon_j}(\theta) - 1} \nabla J(\theta)\| \ / \ 1 \in I_{\varepsilon_j}(\theta) \ , \ y^{1}_{\varepsilon_j}(\theta) > 0\right\}$$

and set $h(\theta) = - \bar{h}_{\varepsilon_j}(\theta)$.

Step 11 : If $\|h(\theta)\| \leq \varepsilon_j$ set $\varepsilon_{j+1} = \beta \varepsilon_j$, set $j = j+1$, and go to step 3; else, go to step 12.

Step 12 : Compute $\lambda(\theta) > 0$ to be the smallest scalar satisfying $J(\theta+\lambda(\theta)h(\theta)) = \min\{J(\theta + \lambda h(\theta) \ / \ \lambda \geq 0 \ , \ (\theta + \lambda\, h(\theta)) \in \Omega\}$.

Step 13 : Set $\theta_{i+1} = \theta_i + \lambda(\theta)h(\theta)$, set $i = i+1$, and go to step 1 .

We have shown that J is continuously differentiable, then is valid the following theorem:

THEOREM. *The sequence θ_i given by the algorithm is finite and its last element is desirable or is infinite and each accumulation point of the sequence is desirable.*
(The proof of this Th. is, essentially, the same that we find in [1], pag. 195).

§6. NUMERICAL SOLUTION OF AN EXAMPLE: THE SHUT DOWN OF A NUCLEAR REACTOR.

The problem is the reduction of the power of a nuclear reactor in a fixed time. The functional to minimize is the xenon poisoning.

The model is ruled by the differential equations:

$$(1)\qquad \begin{cases} \dot{I} = -aI + b\phi \\ \dot{x} = aI + c\phi - (d + e\phi)x \\ \dot{\phi} = U\phi . \end{cases}$$

I is the iodine concentration and x the xenon concentration.

ϕ is the flux of neutrons.

U is the control and it can only assume two values.

The control is applied in the interval $[0,T]$ in such a form that $\phi(T) = \phi_f$ (a fixed value). After that $(t > T)$, the flux is held constant.

If we define $x_M = \max_{t>T} x(t)$, is possible to state the problem in the following form:

Find $U(t)$, $0 \leq t \leq T$, where $U(t)$ is a step function with n' switching ($n' \leq n$, n fixed), that takes only the values $V1$, $V2$ and such that the corresponding response of the system (1) satisfies $\phi(T) = \phi_f$ and gives the minimun value of $x_M(u(0,T))$.

This problem differs from the models studied in the nonlinearities of the equations (1) and in the fixed final condition.

In this case, it can be shown that $J(\theta) = x_M$ (where $\theta = (\theta_1,\ldots,\theta_n)$ is the set of switching points) is a continuous function both with the derivatives of J, and then the theorem remains valid.

The final condition could be introduced in the functional through a penalization function. Another method is the following, we use the propertie that $\phi(T) = \phi_f$ implies that $\forall\, u(0,T) \;/\; \phi(T) = \phi_f \qquad f^\circ(\theta) = \sum_{i=1}^{[(N+1)/2]} \left(\theta_{2i} - \theta_{2i-1}\right) = \text{constant}$

and consider this relation as an additional restriction. In the projected gradient algorithm,, the matrix $F_{I_\varepsilon(\theta)}$ is enlarged in the following form:

$$F_{I_\varepsilon(\theta)} \longrightarrow \tilde{F}_{I_\varepsilon(\theta)} = \left[\nabla f^\circ\ ,\ \nabla f_{i_1},\ldots,\ \nabla f_{i_m}\right] \quad \left\{i_1,\ldots,\ i_m\right\} = I_\varepsilon(\theta)$$

and the points constructed by the algorithm satisfie $\phi(T) = \phi_f$, provided the initial point (θ_o) satisfies that condition.

FORMULAS OF J AND ∇J:

We have defined $J = x_M$; to compute it, we solve the equations:

$$\begin{cases} \dot{I} = -I + b\phi_f \\ \dot{x} = I + c\phi_f - (d + e\phi_f)x \end{cases} \qquad \text{for} \quad t > T$$

with initial conditions $x(T)$, $I(T)$ and find the value $x_M = \max_{t\geq T} x(t)$. (We set $a=1$ making a change of variables).

If $\phi_f = 0$, it is:

$$\begin{cases} x_M = \exp(-t^*) \cdot \dfrac{I(T)}{d-1} + \exp(-dt^*)\left[x(T) - \dfrac{I(T)}{d-1}\right] & \text{if} \quad \dfrac{x(T)}{I(T)}\, d < 1 \quad \text{and} \\ x_M = x(T) \quad \text{if} \quad \dfrac{x(T)}{I(T)}\, d \geq 1 \quad \text{where} \\ t^* = \dfrac{1}{d-1} \ln\left[d + \dfrac{x(T)}{I(T)}\, d(1-d)\right] \end{cases} .$$

To compute $\dfrac{\partial J}{\partial \theta_i} = p_3(\theta_i) x_3(\theta_i)(v_2 - v_1)(-1)^i$, we integrate backwardly the adjoint equations:

$$\begin{cases} -\dfrac{dp_1}{dt} = -p_1 + p_2 \\ -\dfrac{dp_2}{dt} = -p_2(d + e\phi) \\ -\dfrac{dp_3}{dt} = p_1 + cp_2 - p_2 . e . x + p_3 U \end{cases}$$

with final conditions:

$$\begin{cases} p_1(T) = -\dfrac{\partial x_M}{\partial x(T)} \\ p_2(T) = -\dfrac{\partial x_M}{\partial I(T)} \\ p_3(T) = -\dfrac{\partial x_M}{\partial \phi(T)} \end{cases} .$$

The projected gradient algorithm (with the shown modification) was used to solve numerically the problem. The values of J and ∇J were computed integrating the differential equations of x, I, ϕ, p with a 4^{th} order Runge – Kutta method.

NUMERICAL VALUES.

$a = 0.1$	$\phi_f = 0.674 \times 10^{-2}$
$b = 1.0$	$x_o = 2.0$
$c = 1.0$	$I_o = 10.0$
$d = 0.05$	$\phi_o = 1.0$
$e = 0.95$	$T = 10.0$
$v_1 = -2.0$	
$v_2 = 0.0$	

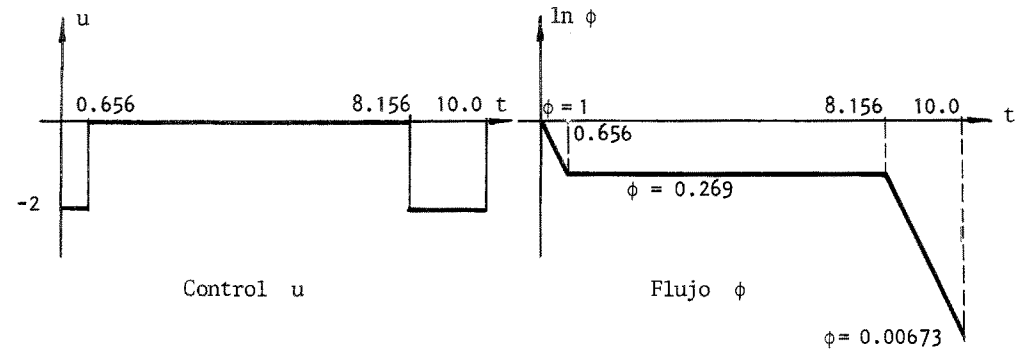
NUMERICAL RESULTS.

The optimal values obtained for $n = 2$ are:

$$\bar{\theta}_1 = 0.6561$$
$$\bar{\theta}_2 = 8.156$$
$$\bar{x}_M = 4.6351 \quad , \quad (\text{for } t_{\bar{x}_M} = 19.86)$$

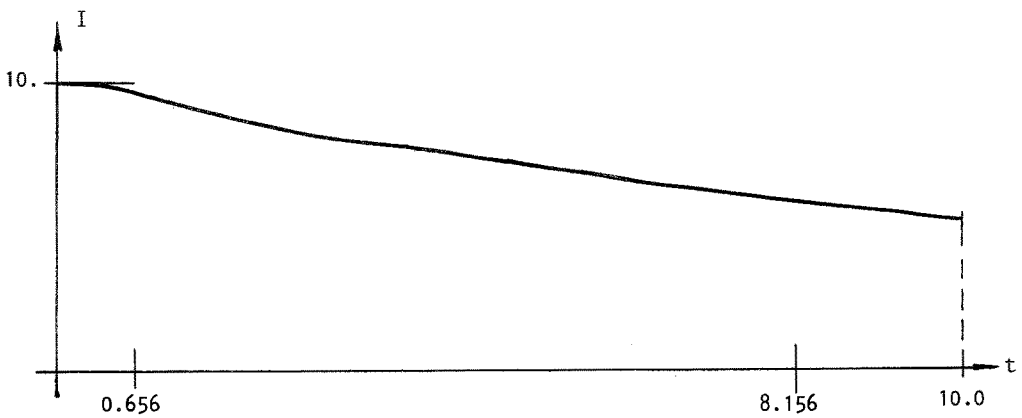
The following is a sample of the sequence produced by the algorithm, that shows rate of convergence.

θ_1	θ_2	x_M
1.000	8.500	4.82751
0.625	8.125	4.63729
0.659	8.154	4.63514
0.656	8.156	4.63513

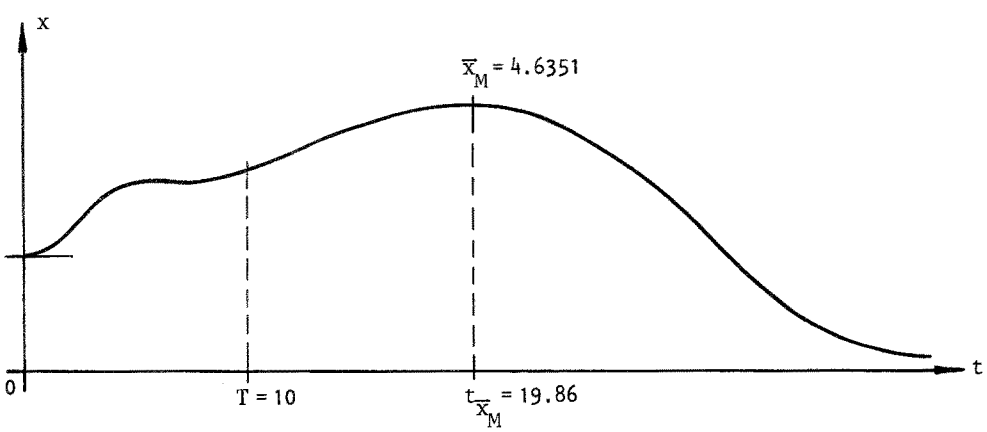


Control u

Flujo ϕ



Iodo I



Xenon: x

For $n \geq 2$ the sequences obtained were convergent to the optimal point for $n = 2$, as is shown in the case $n = 6$. Then, in the set of control with 6 or less switchings, the optimal one is a policy with two switchings.

θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	X_M
0.250	1.000	1.250	2.000	2.250	8.250	4.7485
0.401	0.999	1.213	2.115	2.115	8.115	4.6634
0.488	0.993	1.195	2.115	2.115	8.191	4.6536
0.512	1.028	1.142	2.115	2.115	8.126	4.6470
0.568	1.026	1.132	2.115	2.115	8.172	4.6427
0.585	1.050	1.102	2.115	2.115	8.137	4.6396
0.645	1.059	1.085	2.115	2.115	8.171	4.6370
0.649	1.072	1.072	2.115	2.115	8.149	4.6352
0.656	1.072	1.072	2.115	2.115	8.156	4.6351

The switchings 2-3 and 4-5 are simultaneous and could be eliminated and the new policy is the optimal one for the problem with $n = 2$.

§7. FORM AND PROPERTIES OF THE SECOND DERIVATIVES OF J.

$$\frac{\partial J}{\partial \theta_i} = 1(x(\theta_i), u(\theta_i-), \theta_i) - 1(x(\theta_i), u(\theta_i+), \theta_i) + p(\theta_i)G(\theta_i)(v_1 - v_2))(-1)^i$$

a) $\dfrac{\partial^2 J}{\partial \theta_i^2}$

First, we find the formulas of $\dfrac{\partial x(\theta_i)}{\partial \theta_i}$, $\dfrac{\partial p(\theta_i)}{\partial \theta_i}$

$$x(\theta_i) = \Phi(\theta_i, 0)x_o + \int_0^{\theta_i} \Phi(\theta_i, s)G(s)u(s)ds$$

then,

$$\frac{\partial x(\theta_i)}{\partial \theta_i} = F(\theta_i)x(\theta_i) + G(\theta_i)u(\theta_i-) \quad .$$

Also,

$$p(\theta_i) = p(T)\Phi(T, \theta_i) - \int_{\theta_i}^{T} \frac{\partial 1}{\partial x}(x(s), u(s), s)\Phi(s, \theta_i)ds$$

with

$$p(T) = -\frac{\partial g}{\partial x}(x(T))$$

$$\frac{\partial p(\theta_i)}{\partial \theta_i} = -p(T)\Phi(T, \theta_i)F(\theta_i) + \int_{\theta_i}^{T} \frac{\partial 1}{\partial x}(x(s), u(s), s)\Phi(s, \theta_i)F(\theta_i)ds +$$

$$+ \frac{\partial 1}{\partial x}(x(\theta_i), u(\theta_i+), \theta_i) - \left(\frac{\partial}{\partial \theta_i}x(T)\right)' \frac{\partial^2 g}{\partial x^2}(x(T))\,\Phi(T, \theta_i) -$$

$$- \int_{\theta_i}^{T} \left(\frac{\partial}{\partial \theta_i}x(s)\right)' \frac{\partial^2 1}{\partial x^2}(x(s), u(s), s)\Phi(s, \theta_i)ds$$

but, we know that, for $t > \theta_i$

$$\frac{\partial}{\partial \theta_i}x(t) = \Phi(t, \theta_i)G(\theta_i)(v_2 - v_1)(-1)^i$$

then,

$$\frac{\partial p(\theta_i)}{\partial \theta_i} = -p(\theta_i)F(\theta_i) + \frac{\partial 1}{\partial x}(x(\theta_i), u(\theta_i+), \theta_i) -$$

$$- (-1)^i(v_2 - v_1)G'(\theta_i)\Phi'(T, \theta_i)\frac{\partial^2 g}{\partial x^2}(x(T))\Phi(T, \theta_i) -$$

$$- \int_{\theta_i}^{T} (-1)^i (v_2 - v_1) G'(\theta_i) \Phi'(s,\theta_i) \frac{\partial^2 l}{\partial x^2}(x(s), u(s), s).\Phi(s,\theta_i) ds \quad .$$

Now, it is possible to compute $\frac{\partial^2 J}{\partial \theta_i^2}$

$$\frac{\partial^2 J}{\partial \theta_i^2} = \frac{\partial}{\partial \theta_i} l(x(\theta_i), u(\theta_i-), \theta_i) - \frac{\partial l}{\partial \theta_i}(x(\theta_i), u(\theta_i+), \theta_i) +$$

$$+ \left[\frac{\partial}{\partial x} l(x(\theta_i), u(\theta_i-), \theta_i) - \frac{\partial l}{\partial x}(x(\theta_i), u(\theta_i+), \theta_i)\right]\left[F(\theta_i)x(\theta_i) + G(\theta_i)u(\theta_i-)\right] +$$

$$+ p(\theta_i)\left(\frac{dG(\theta_i)}{d\theta_i}\right)(v_1 - v_2)(-1)^i + \left[- p(\theta_i)F(\theta_i) + \frac{\partial l}{\partial x}(x(\theta_i), u(\theta_i+), \theta_i)\right] .$$

$$.G(\theta_i)(v_1 - v_2)(-1)^i + (-1)^i (v_1 - v_2) G'(\theta_i)\Phi'(T,\theta_i)\frac{\partial^2 g}{\partial x^2}(x(T))\Phi(T,\theta_i)G(\theta_i)(v_1 - v_2)(-1)^i +$$

$$+ (v_1 - v_2)(-1)^i G'(\theta_i) \int_{\theta_i}^{T} \Phi'(s,\theta_i)\frac{\partial^2 l}{\partial x^2}(x(s), u(s), s)\Phi(s,\theta_i) ds . G(\theta_i)(v_1 - v_2)(-1)^i \quad .$$

b) If $j > i$

$$\frac{\partial}{\partial \theta_j} x(\theta_i) = 0$$

then:

$$\frac{\partial}{\partial \theta_j}\left(\frac{\partial J}{\partial \theta_i}\right) = \frac{\partial}{\partial \theta_j} p(\theta_i) \; . \; G(\theta_i)(v_1 - v_2)(-1)^i \quad ,$$

and we must only know $\frac{\partial}{\partial \theta_j} p(\theta_i)$ to find the form of $\frac{\partial^2 J}{\partial \theta_j \, \partial \theta_i}$.

From the integral formula of p we obtain:

$$\frac{\partial}{\partial \theta_j} p(\theta_i) = \left(\frac{\partial}{\partial \theta_j} x(T)\right)' . \frac{\partial^2 g}{\partial x^2}(x(T)) . \Phi(T,\theta_i) +$$

$$+ \left[\frac{\partial l}{\partial x}(x(\theta_j), u(\theta_j+), \theta_j) - \frac{\partial l}{\partial x}(x(\theta_j), u(\theta_j-), \theta_j)\right]\Phi(\theta_j,\theta_i) -$$

$$- \int_{\theta_j}^{T} (-1)^j (v_2 - v_1) G'(\theta_j)\Phi'(s,\theta_j)\frac{\partial^2 l}{\partial x^2}(x(s), u(s), s)\Phi(s,\theta_i) d\tau$$

and then,

$$\frac{\partial}{\partial \theta_j}\left(\frac{\partial J}{\partial \theta_i}\right) = (-1)^j (v_2 - v_1) G'(\theta_j)\Phi(T,\theta_j)\frac{\partial^2 g}{\partial x^2}(x(T)) .\Phi(T,\theta_i)G(\theta_i)(v_1 - v_2)(-1)^i +$$

$$+ \left[\frac{\partial l}{\partial x}(x(\theta_j), u(\theta_j+), \theta_j) - \frac{\partial l}{\partial x}(x(\theta_j), u(\theta_j-), \theta_j)\right]\Phi(\theta_j,\theta_i)G(\theta_i)(v_1 - v_2)(-1)^i -$$

$$- (-1)^j (v_2 - v_1) G'(\theta_j) \int_{\theta_j}^{T} \Phi'(s,\theta_j) \frac{\partial^2 l}{\partial x^2}(x(s), s)\Phi(s,\theta_i) ds \; . \; G(\theta_i)(v_1 - v_2)(-1)^i$$

c) If $i > j$

$$\frac{\partial}{\partial \theta_j} x(\phi_i) = \Phi(\theta_i,\theta_j)G(\theta_j)(v_2 - v_1)(-1)^j$$

and

$$\frac{\partial}{\partial \theta_j} p(\theta_i) = - \left(\frac{\partial}{\partial \theta_j} x(T)\right)' \frac{\partial^2 g}{\partial x^2}(x(T))\Phi(T,\theta_i) - \int_{\theta_i}^{T} \left(\frac{\partial}{\partial \theta_j} x(s)\right)' \frac{\partial^2 l}{\partial x^2}(x(s), u(s), s)\Phi(s,\theta_i) ds$$

then,

$$\frac{\partial}{\partial \theta_j}\left(\frac{\partial J}{\partial \theta_i}\right) = \left[\frac{\partial l}{\partial x}(x(\theta_i), u(\theta_i-), \theta_i) - \frac{\partial l}{\partial x}(x(\theta_i), u(\theta_i+), \theta_i)\right]\Phi(\theta_i,\theta_j) \quad .$$

$$.G(\theta_j)(v_2 - v_1)(-1)^j + (-1)^j (v_2 - v_1) G'(\theta_j)\Phi'(T,\theta_j)\frac{\partial^2 g}{\partial x^2}(x(T))\Phi(T,\theta_i)G(\theta_i)(v_1 - v_2)(-1)^i -$$

$$- (-1)^j (v_2 - v_1) G'(\theta_j) \int_{\theta_i}^{T} \Phi'(s,\theta_j) \frac{\partial^2 l}{\partial x^2}(x(s),u(s),s) \Phi(s,\theta_i) ds \, . \, G(\theta_i)(v_1 - v_2)(-1)^i \quad .$$

It can be proved, in the same form we have done for the first derivatives, that the second derivatives are continuous provided $g \in C^2$, $l \in C^2$, $G \in C^1$ and $F \in C$. This continuity is important to obtain superlinear convergence when it is applied the conjugate gradient method.

§8. SOLUTION OF THE PROBLEM USING A MIXED METHOD OF PENALIZATION AND CONJUGATE GRADIENTS.

The problem of minimum with restrictions:

$$(1) \qquad \min_{\Omega} J(\theta) \quad ; \quad \Omega = \{\theta \ / \ 0 \leq \theta_1 \leq \ldots \leq \theta_n \leq T \ , \ \theta \in R^n\}$$

is transformed into another that could be solved using the methods of optimization without restrictions. This is done applying penality functions.

The new problem is:

$$\min_{\mathring{\Omega}} J_\beta(\theta) \quad ; \quad \mathring{\Omega} = \left\{\theta \in R^n \ / \ 0 < \theta_1 < \theta_2 < \ldots < \theta_n < T\right\}$$

and

$$J_\beta(\theta) = J(\theta) + \beta \sum_{i=1}^{n+1} - \varphi_i^{-1}(\theta)$$

$$\varphi_1 = - \theta_1 \quad ; \quad \varphi_i = - \theta_i + \theta_{i-1} \quad i = 2,\ldots,n \quad ; \quad \varphi_{n+1} = \theta_n - T \ .$$

ALGORITHM.

Step 0: Choose $\theta_0 \in \mathring{\Omega}$; $\beta_0 > 0$; $\varepsilon_0 > 0$ and set $i = 0$.

Step 1: Apply the conjugate gradient method to the minimization of J_β until it is obtained a point θ_{i+1} such that $\|\nabla J_\beta(\theta_{i+1})\| < \varepsilon_i$

Step 2: Let $\beta_{i+1} = \frac{\beta_i}{2}$, $\varepsilon_{i+1} = \frac{\varepsilon_i}{2}$, $i = i+1$ and go to 1 .

REMARK.

The conjugate gradient method could be applied in $\mathring{\Omega}$, modifying the one dimensional search (along the conjugate directions) in such a form that the point are always chosen in $\mathring{\Omega}$.

It is known (Kuhn–Tucker's theorem) that a necessary condition for optimality of $\bar{\theta}$ in problem 1 is:

$$\nabla J(\bar{\theta}) + \sum_{i \in I(\bar{\theta})} \mu_i \nabla \varphi_i(\bar{\theta}) = 0 \qquad \mu_i \geq 0$$

$$I(\theta) = \left\{i \ / \ \varphi_i(\theta) = 0\right\} \ .$$

DEFINITION.

A point θ is desirable if:

i) $\theta \in \Omega$

ii) $\nabla J(\theta) + \sum_{i \in I(\theta)} \mu_i \nabla \varphi_i(\theta) = 0 \quad ; \quad \mu_i \geq 0 \quad i \in I(\theta)$

The algorithm has the property:

LEMA: If $J(\theta)$ is continuously differentiable, then the algorithm produces a sequence

of different points and all the accumulation points are desirable, or the sequence has a finite number of different points and the last (infinitely repeated) is desirable.

PROOF.

i) $\overset{\circ}{\Omega}$ is relatively compact, then $\{\theta_i\}$ has accumulation points.

Let be $\theta_{i_k} \longrightarrow \tilde{\theta}$ $\quad k = 1,2,\ldots..$

$$\overline{\overset{\circ}{\Omega}} = \Omega \quad , \quad \text{then} \quad \tilde{\theta} \in \Omega \quad .$$

It will be proved that $\tilde{\theta}$ satisfies the Kuhn–Tucker conditions.

From the step 1 of the algorithm it follows:

$$(1) \qquad \nabla J(\theta_{i_k}) + \sum_{j=1}^{n+1} \left[\beta_{i_k} / \varphi_j^2(\theta_{i_k})\right] \nabla\varphi_j(\theta_{i_k}) \longrightarrow 0 \quad .$$

It is easily seen that, $\forall\, \theta \in \Omega, I(\theta)$ has at most n elements and $\{\nabla\varphi_i(\theta),\ i \in I(\theta)\}$ is a set of linearly independent vectors.

Then,

$$\nabla J(\theta_{i_k}) + \sum_{j \in I(\tilde{\theta})} \left[\beta_{i_k} / \varphi_j^2(\theta_{i_k})\right] \nabla\varphi_j(\theta_{i_k}) \longrightarrow 0 \quad .$$

We define: ψ_{i_k} (matrix $n \times m$)

$$\psi_{i_k} = \left[\nabla\varphi_{j_1}, \nabla\varphi_{j_2}, \ldots, \nabla\varphi_{j_m}\right] \qquad \{j_1, \ldots, j_m\} = I(\tilde{\theta})$$

and then:

$$\begin{pmatrix} \beta_{i_k} / \varphi^2_{j_1}(\theta_{i_k}) \\ \vdots \\ \beta_{i_k} / \varphi^2_{j_m}(\theta_{i_k}) \end{pmatrix} - (\psi'_{i_k}\psi_{i_k})^{-1}\psi'_{i_k}\nabla J(\theta_{i_k}) \longrightarrow 0$$

because, $\psi_{i_k} \longrightarrow \psi = \left[\nabla\varphi_{j_1}(\tilde{\theta}), \ldots, \nabla\varphi_{j_m}(\tilde{\theta})\right]$, this matrix is of maximun rank and then $\psi'_{i_k}\psi_{i_k} \longrightarrow \psi'\psi$ invertible matrix.

Then: $\beta_{i_k} / \varphi^2_{j_s}(\theta_{i_k}) \longrightarrow \mu_{j_s} \geq 0 \qquad j_s \in I(\theta)$

and also, taking limits in (1):

$$\nabla J(\tilde{\theta}) + \sum_{j \in I(\tilde{\theta})} \mu_j \nabla\varphi_j(\tilde{\theta}) = 0 \quad .$$

Thus, $\tilde{\theta}$ is desirable.

ii) If $\exists N \,/\, \forall\, i \geq N, \theta_i = \theta_N$, from Step 1 it follows:

$$\left\| \nabla J(\theta_N) + \beta_i \nabla\left(\sum_{j=1}^{n+1} - \varphi_j^{-1}(\theta_i)\right) \right\| \longrightarrow 0 \qquad i \to \infty$$

but $\beta_i \longrightarrow 0$

then: $\nabla J(\theta_N) = 0$, $\theta_N \in \overset{\circ}{\Omega}$ and θ_N is desirable.

NUMERICAL RESULTS.

We have applied this method to the problem of the shutdown of the nuclear reactor. The results are the same obtained with the projected gradient method: the optimal policy in the set of step functions is a policy with two switchings.

The following table shows the convergence for $n = 2$

θ_1	θ_2	J_β	β
1.000	8.500	6.62755	1.0000
0.913	8.413	5.49471	0.4000
0.792	8.292	4.98835	0.1600
0.720	8.220	4.77629	0.0640
0.684	8.184	4.69171	0.0256
0.667	8.167	4.65213	0.0076
0.659	8.159	4.64021	0.0023
0.658	8.158	4.63601	0.0004
0.6570	8.157	4.63530	0.00008
0.6564	8.1564	4.63516	0.000016

§9. PROOF THAT THE SOLUTIONS OF THE PROBLEM WITH FIXED NUMBER OF SWITCHING ARE MINIMIZATING SEQUENCE FOR THE PROBLEM WITH MEASURABLE CONTROLS.

The set of problems with step function controls (as were stated in §1) could be considered as a set of approximations to the problem:

$$\min_{\mathcal{U}_{ad}} J(u(.))$$

with

$$\mathcal{U}_{ad} = \left\{ u(.) \text{ measurable in } [0,T] \;/\; u(t) = v_1 \text{ or } u(t) = v_2 \quad \text{a.e.} \right\}$$

$$J(u(.)) = \int_0^T 1(x(s), u(s), s)ds + g(x(T))$$

and $x(s)$ satisfies:

$$\begin{cases} \dfrac{dx}{dt}(t) = F(t)x(t) + G(t)u(t) \\ x(0) = x_0 \end{cases}$$

Under the assumptions that $1,g$ are continuous and F,G are integrable, it can be proved that J is a continuous functional for $u \in \mathcal{U}_{ad}$ (with the $L_1(0,T)$ topology) If we denote with $\bar{u}_n$ the optimal solution with at most n switchings, we shall prove that $\bar{u}_n$ is a minimizating sequence for the new problem.

Let w_n be a minimizating sequence:

$$\lim_{n\to\infty} J(w_n) = \inf_{\mathcal{U}_{ad}} J(u) \qquad w_n \in \mathcal{U}_{ad} \quad .$$

But, for the continuity of J, it is possible to find $\tilde{w}_n$ ($\tilde{w}_n$ a step function / $\tilde{w}_n \in \mathcal{U}_{ad}$) such that $|J(\tilde{w}_n) - J(w_n)| < \dfrac{1}{2^n}$

Let $k(n)$ be the number of switchings of $\tilde{w}_n$. By definition, $\forall\, k(n)$

$$J\left(\bar{u}_{k(n)}\right) \leq J(\tilde{w}_n) \quad .$$

Then, using the property that $J(\bar{u}_n)$ is non-incresing.

$$\lim_{m\to\infty} J(\bar{u}_m) \leq \lim_{n\to\infty} J(\tilde{w}_n) = \lim_{n\to\infty} J(w_n) = \inf_{\mathcal{U}_{ad}} J$$

but: $$J(\bar{u}_m) \geq \inf_{\mathcal{U}_{ad}} J$$

then $$\lim_{m\to\infty} J(\bar{u}_m) = \inf_{u\in\mathcal{U}_{ad}} J(u) \quad .$$

§10. OTHERS RESULTS.

The bang–bang problem with restricted number of switchings could be analysed in global form (i.e., the initial state $x_0 \in R^n$ or $x_0 \in \Omega$) and is reduced to a sequence of stopping–times problems. Also in the problem with measurable control ($u=0$ or $u=1$ a.e.) it is possible to prove existence theorems and to analyse the optimal cost function with the hamiltonian technique. These are the objects of forthcoming papers.

REFERENCES

1. E. POLAK. *"Computational methods in optimization"*.Academic Press 1971.

2. E.B. LEE, L. MARKUS."*Foundations of Optimal Control Theory*". Wiley. 1968.

3. L.S. PONTRYAGIN, V.G. BOLTYANSKII, R.V. GAMKERELIDZE, E.F. MISCHENKO. *"Mathematical Theory of Optimal Processus"*. New York, Wiley 1962.-

Los originales de este trabajo fueron preparados en el *Instituto de Matemática "Beppo Levi"* por la Sra.H.I.Warecki de MUTY.