Optimal control of differential systems with discontinuous right-hand side

DOI:
10.6100/IR111488

Document status and date:
Published: 01/01/1970

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

OPTIMAL CONTROL OF DIFFERENTIAL SYSTEMS WITH DISCONTINUOUS RIGHT-HAND SIDE

m. l. j. hautus

OPTIMAL CONTROL OF DIFFERENTIAL SYSTEMS WITH DISCONTINUOUS RIGHT-HAND SIDE

PROEFSCHRIFT

TER VERKRIJGING VAN DE GRAAD VAN DOCTOR IN DE TECHNISCHE WETENSCHAPPEN AAN DE TECHNISCHE HOGESCHOOL TE EINDHOVEN OP GEZAG VAN DE RECTOR MAGNIFICUS DR.IR. A.A.TH.M. VAN TRIER, HOOGLERAAR IN DE AFDELING DER ELEKTROTECHNIEK, VOOR EEN COMMISSIE UIT DE SENAAT TE VERDEDIGEN OP MAANDAG 29 JUNI 1970 TE 16.00 UUR.

DOOR

MATHEUS LODEWIJK JOHANNES
HAUTUS

GEBOREN TE HELDEN

DIT PROEFSCHRIFT IS GOEDGEKEURD DOOR

DE PROMOTOR PROF.DR. N.G. DE BRUIJN

Aan Hilde

Aan mijn ouders

C O N T E N T S

LIST OF SYMBOLS

SUBJECT INDEX

C H A P T E R I

INTRODUCTION

I.1. *General remarks on optimal control theory*

In the theory of optimal control of differential systems one deals with systems of differential equations of the form

$$\begin{aligned} &\dot{x}(t) = f(x(t),u(t),t) \qquad (t_1 < t < t_2) \; , \\ &x(t_1) = a \; , \end{aligned} \tag{1}$$

where a is a given n-vector, x is an n-vector valued function on $[t_1,t_2]$, and u is an m-vector valued function on the same interval. The function u is called the *control function* and x is called the *state variable*. A solution of (1) is called a *trajectory* corresponding to the control u and is denoted by x_u.

Usually, only control functions are allowed which satisfy some conditions, like $u(t) \in U$ for $t \in [t_1,t_2]$, where U is some subset of R^m. Also, there often are restrictions on the state variables, such as $x(t_2) \in S$ or $x(t) \in X(t)$ $(t_1 \le t \le t_2)$ where S is a subset of R^n and X is a mapping of $[t_1,t_2]$ into the class of subsets of R^n. A control u is called *admissible* if it satisfies the conditions mentioned above, and if x_u satisfies the conditions posed upon the state variable.

Furthermore, a *performance index* J is given, which attaches a real number to each admissible control u. For example,

$$J(u) = \int_{t_1}^{t_2} h(x_u(t),u(t),t)dt \quad \text{or} \quad J(u) = F(x_u(t_2))$$

where h, F are real valued functions.

Then an admissible control $\bar{u}$ is called *optimal* if for all admissible controls u we have $J(\bar{u}) \le J(u)$, that is, if J(u) is minimal for $u = \bar{u}$. The task of *optimal control theory* is to determine optimal controls.

Instead of (1) one can have a discrete system of the form

$$\begin{aligned} &x(k+1) = f(x(k),u(k),k) \qquad (k = 0,1,2,\dots)\ , \\ &x(0) = a\ . \end{aligned} \tag{2}$$

Here x is a sequence of n-vectors and u is a sequence of m-vectors. For (2) one can define an optimal control problem analogous to the one of (1). Discrete systems can occur as approximations for continuous systems and also they appear directly in applications. In this thesis we will consider another type of problem which gives rise to discrete systems (see section 2).

REMARK. Instead of (1) or (2) one can have more general systems, governed for example by difference-differential equations, integral equations or partial differential equations, but we will not discuss them in this thesis.

The method for finding an optimal control usually consists of two parts: First, one shows that there exists an optimal control (mostly by a non-constructive method). Then one derives necessary conditions for a control to be optimal, that is, properties that are satisfied by an optimal control. Sometimes there exists just one admissible control which satisfies the necessary conditions and then this control must be optimal. Sometimes there is a finite number of admissible controls satisfying the necessary conditions. In that case, an optimal control can be found by comparison. Anyhow, by means of the necessary conditions, one restricts the set of possibly optimal controls.

For a general class of differential systems of the form (1) necessary conditions for optimality are given by the *maximum principle* of Pontryagin. (See [1] Chapter IV, V; [2] Chapter I.) (A special case of this theorem will be given in Chapter I, section 4.) For discrete systems there exists a necessary condition which is analogous to the maximum principle and which was first given in its general form by Halkin [3]. (We will state this theorem in Chapter III, section 1.) For discrete systems there exists a second method for finding optimal controls, namely, the *dynamic programming* method (see [4]). Except for simple situations, the result of this method is a complicated functional-recurrence equation which cannot be solved explicitly. On the other hand, the discrete maximum principle results in a discrete boundary value problem.

I.2. *Summary of the thesis*

In this thesis we will study systems of the form (1.1) where the right-hand side is discontinuous for some values of (x,t). For these systems the maximum principle ceases to hold. Problems of this kind are also considered in [5] where the maximum principle is shown to be valid in a modified form. We will follow a different method, however. Loosely speaking, in order to solve optimal control problems for these systems, we will divide the trajectory into pieces on which f is continuously differentiable. For these pieces the maximum principle holds. A trajectory which is optimal on each of these pieces will be called *piecewise optimal*. In order to find a control which is optimal for the whole trajectory, we have to fit the pieces together. This results in a discrete optimal control problem which can be solved by one of the methods discussed in section 1. Instead of treating a general theory, we will restrict ourselves to a few simple examples which are carried out in detail.

In Chapter II we will discuss a simple equation, and we will solve a number of optimal control problems for this equation. We will use the dynamic programming method for the "discrete parts" of these problems. A more complete summary of Chapter II can be found in section 2 of Chapter II.

In Chapter III we will use the discrete maximum principle for the discrete part of some of the problems of Chapter II. Also, a different system of the form (1.1) will be solved by the discrete maximum principle.

We give a summary of the remaining part of this chapter. In section 3 we will develop some notations to be used throughout the thesis. In section 4 we will discuss some existence theorems for differential equations. When we require the controls to be continuous or even piecewise continuous, we cannot give existence theorems for optimal controls; hence we have to admit measurable controls. Therefore, we need a more general definition of the solution of a differential equation then the classical one. Such a definition (section 4, definition 1) was first given by Caratheodory [6] (p. 665-688). Existence and uniqueness theorems for the solutions of these generalized differential equations can be given similar to the ones for classical differential equations (for example Theorem 1 of section 4). For our purposes however, even these general existence and uniqueness theorems are not sufficient, since they require the right-hand side to be a continuous function of the state variable x. If the right-hand side does not depend continuously on x, a solution need not exist, and if we have existence, we do not necessarily have uniqueness. We will give conditions on the right-hand side which assure existence and uniqueness of the solution (section 4, Theorem 2).

In section 5 we will give a theorem which assures the existence of an optimal control for a control system of the form (1.1) where f depends linearly on x. Furthermore, we give a maximum principle for such a system. These results will be used in Chapter II, section 3 and in Chapter III, section 3.

I.3. *Notations*

A. *Formula indication*

We will number the formulas and theorems in each section independently. When referring for example to formula 1 of section 1 of Chapter I we will write (1) in section 1 of Chapter I, (1.1) in Chapter I and not in section 1, and (I.1.1) in Chapters II, III.

B. *Vectors*

The n-dimensional real euclidean space is denoted by R^n. We will distinguish between row and column vectors. Elements of R^n are considered column vectors. The set of row vectors will be denoted by R^n_T. The *norm* of x, denoted by $|x|$, is defined by

$$|x| := \left[\sum_{i=1}^{n} x_i^2\right]^{\frac{1}{2}},$$

where $x_1,\dots,x_n$ are the entries of x. In the two-dimensional case we define $\underline{x} = \binom{x}{y}$. If $S \subset R^n$, then $\bar{S}$ and ∂S are the closure and the boundary of S. The transpose of a vector or matrix will be denoted by an accent. For example, if x is column vector then $x = (x_1,\dots,x_n)'$. In the text however we will often omit the accent when there is no danger of confusion. The set of real numbers is denoted by R^1 and the set of positive real numbers by R^1_+.

C. Functions

If S and T are sets, the set of mappings from S into T is denoted by $S \to T$. If $S \subset R^n$ and $T \subset R^m$ then $C(S \to T)$ is the set of continuous mappings from S into T. If in addition S is open then $C^p(S \to T)$ denotes the set of p times continuously differentiable functions for $p = 1,2,\dots$. If I is a set of real numbers and $S \subset R^n$, $L(I \to S)$ denotes the set of all functions in $I \to S$, which are integrable on every compact subset of I. We will call these functions *locally integrable*. In particular we will encounter $L([0,\infty) \to S)$ and $L((T_1,T_2) \to S)$.

The expression "almost everywhere" (abbreviated a.e.) always refers to a one-dimensional independent variable, usually denoted by t or s.

The function which attaches the value f(x) to the variable x will be denoted by f or by $x \mapsto f(x)$. For example, if $g \in (R^n \to R^m)$ and $f \in (R^n \times R^m \to R^p)$, then $x \mapsto f(x,g(x))$ is a function in $R^n \to R^p$.

REMARK. If T is a discrete set (for example the natural numbers) and S is an open set in R^n, then we say $f \in C^p(S \times T \to U)$ if $(x \mapsto f(x,t)) \in C^p(S \to U)$ for all $t \in T$.

D. Partial derivatives

If $f \in C^1(S \to T)$ where S is an open set in R^n and $T \subset R^m$, then for a given $x \in S$ the *functional matrix* of f at the point x will be denoted by $\partial f(x)$, hence

$$(\partial f(x))_{ij} := \frac{\partial f_i}{\partial x_j} \qquad (i = 1,\dots,m;\ j = 1,\dots,n)\ .$$

In particular, if $m = 1$, then $\partial f(x)$ is the gradient of f at x. (Note that this is a row vector.)

If $f \in C^1(S \times U \to T)$ with $S \subset R^n$, $U \subset R^\ell$ open, $T \subset R^m$, and if generic elements of $S \times U$ are denoted by (s,u), then $\partial_s f(s,u)$ is the *partial functional matrix* with respect to s, at the point (s,u); that is,

$$(\partial_s f(s,u))_{ij} := \frac{\partial f_i(s,u)}{\partial s_j} \qquad (i = 1,\dots,m;\ j = 1,\dots,n)\ ,$$

and $\partial_u f(s,u)$ is defined similarly.

I.4. *Existence and uniqueness of solutions of differential equations*

In this section S is an open set in R^n and T_0, T_1, T_2 are real numbers, satisfying $T_1 < T_0 < T_2$.

DEFINITION 1. If $a \in S$, $f \in (S \times (T_1,T_2) \to R^n)$, and if t_1 and t_2 are real numbers, then $x \in ((t_1,t_2) \to S)$ is called a *solution* of the *differential equation* (with initial condition)

$$\dot{x} = f(x,t)\ ,\quad x(T_0) = a \tag{1}$$

on (t_1,t_2) if $T_1 \le t_1 < T_0 < t_2 \le T_2$,
$(t \mapsto f(x(t),t)) \in L((t_1,t_2) \to R^n)$ and

$$x(t) = a + \int_{T_0}^{t} f(x(s),s)ds \qquad (t_1 < t < t_2)\ .$$

REMARK 1. If f is continuous, it is easily seen that every solution of (1) is a solution in the classical sense. In general, a solution of (1) is an absolutely continuous function which satisfies (1) almost everywhere. Note that the phrase "solution of the differential equation (1) on (t_1,t_2)" implies conditions on the numbers t_1 and t_2. Accordingly, these conditions will not be repeated in the sequel.

REMARK 2. Sometimes solutions, defined on $[T_0,t_2)$ or $(t_1,T_0]$ (with $T_1 \le t_1 < T_0$ and $T_0 < t_2 \le T_2$) are considered. The definitions of these solutions are similar to definition 1 and they are called *one-sided solutions*.

DEFINITION 2. A solution of (1) on (t_1,t_2) is called a *unisolution* of (1) on (t_1,t_2) if for every t_3 and t_4 and every solution x^* of (1) on (t_3,t_4) we have $x^*(t) = x(t)$ $(t \in (t_1,t_2) \cap (t_3,t_4))$.

REMARK 3. If all solutions of (1) are unisolutions and if we identify functions with their graphs, then the solutions of (1) are totally ordered by inclusion. It is easily seen in that case that the union of all solutions of (1) is the maximal solution of (1) (with respect to inclusion). (Furthermore, this solution is also a unisolution.)

REMARK 4. If x is a unisolution and $\tilde{x}$ is a one-sided solution of (1), then x and $\tilde{x}$ coincide on their common domain. In order to see this, assume that $\tilde{x}$ is defined on $[T_0,t_3)$ and x on (t_1,t_2) and observe that the function x^*, defined by $x^*(t) = x(t)$ $(t_1 < t \le T_0)$, $x^*(t) = \tilde{x}(t)$ $(T_0 < t < t_3)$, is a solution of (1).

DEFINITION 3. If $f \in (S \times (T_1,T_2) \to R^n)$, we say that $f \in Li(S \times (T_1,T_2) \to R^n)$ if

i) $(t \mapsto f(x,t)) \in L((T_1,T_2) \to R^n)$ for all $x \in S$.

ii) For every compact set $K \subset S$ and every compact subinterval $[t_1,t_2]$ of (T_1,T_2) there exists $M > 0$ such that for almost all $t \in [t_1,t_2]$ we have $|f(x,t)| \le M$ $(x \in K)$ and $|f(x,t) - f(y,t)| \le M|x - y|$ $(x,y \in K)$.

Now we have:

THEOREM 1. If $a \in S$ and $f \in Li(S\times(T_1,T_2) \to R^n)$, then there exist t_1,

t_2 such that (1) has a solution on (t_1,t_2). Furthermore, every solution of (1) is a unisolution.

For a proof see [7], Chapter IX.

DEFINITION 4. A function $f \in (S \times (T_1,T_2) \to R^n)$ is called *smoothly discontinuous* if there exists $g \in C^1(S \to R^1)$ such that for almost all $t \in (T_1,T_2)$ we have

$$f(x,t) = f_j(x,t) \qquad (x \in G_j) \quad (j = 0,1,2) ,$$

where

$$f_j \in Li(G_j^* \times (T_1,T_2) \to R^n) \qquad (j = 0,1,2) ,$$

$$G_0 := \{x \in S \mid g(x) = 0\} ,$$

$$G_1 := \{x \in S \mid g(x) > 0\} ,$$

$$G_2 := \{x \in S \mid g(x) < 0\} ,$$

and where G_j^* is some neighborhood of $G_j \cup G_0$ $(j = 0,1,2)$.

For a smoothly discontinuous function f, a solution of (1) does not necessarily exist and if a solution does exist, then it is not necessarily a unisolution. This is illustrated by the following examples:

EXAMPLE 1. Consider the differential equation in R^1:

$$\dot{x} = 1 - 2 \operatorname{sgn} x,\ x(0) = 0 \qquad (-1 < t < 1) . \tag{2}$$

Suppose that x is a solution of (2) on (t_1,t_2), where $-1 \le t_1 < 0 < t_2 \le 1$. Then x cannot be identically zero on $[0,t_2)$, since otherwise we would have $\dot{x} = 1$ a.e. on $[0,t_2)$, which is contradictory. Let us assume that $x(t_0) > 0$ for some $t_0 \in (0,t_2)$. Let

$$t_3 := \max\{t \le t_0 \mid x(t) = 0\} .$$

Then we have

$$x(t_0) = x(t_3) + \int_{t_3}^{t_0} (1 - 2 \operatorname{sgn} x(s))ds = -(t_0 - t_3) < 0$$

which is also a contradiction. In a similar way $x(t_0) < 0$ can be proved to be impossible. Therefore (2) does not have a solution.

EXAMPLE 2. Consider the differential equation in R^1:

$$\dot{x} = \operatorname{sgn} x \qquad (-1 < t < 1) \ , \ \ x(0) = 0 \ . \tag{3}$$

Then the functions $\pm\,\phi(t - a)$ (restricted to $(-1,1)$), where $\phi(t) := \max\{0,t\}$ ($t \in R^1$) and $a \in [0,1)$ are solutions of (3) according to definition 1.

The following theorem gives a sufficient condition for the existence of solutions and unisolutions of (1) if f is smoothly discontinuous:

THEOREM 2. If $f \in (S \times (T_1,T_2) \to R^n)$ is smoothly discontinuous, with g, f_j, G_j, G_j^* as defined in definition 4 and if $h_j \in (G_0 \times (T_1,T_2) \to R^1)$ is defined by

$$h_j(c,t) := (\partial g(c))f_j(c,t) \qquad (j = 0,1,2) \ , \tag{4}$$

then we have the following results:

i) If for every $c \in G_0$ there exist $\delta > 0$ such that at least one of the statements:

$$\text{" } h_1(c,t) \geq \ \delta \wedge h_2(c,t) \geq \ \delta \quad \text{(a.e.) "}, \tag{5}$$

$$\text{" } h_1(c,t) \leq -\delta \wedge h_2(c,t) \leq -\delta \quad \text{(a.e.) "} \tag{6}$$

holds, then for every $a \in S$ there exist t_1, t_2 such that (1) has a solution on (t_1,t_2).

ii) If the condition of i) is satisfied and if furthermore $|h_0(c,t)| \geq \delta$ (a.e.), then every solution of (1) is a unisolution.

iii) If, in addition to the conditions of ii), there exist positive numbers M_1 and M_2 such that for almost all $t \in (t_1,t_2)$ we have $|f(x,t)| \leq M_1|x| + M_2$, then there exist t_1, t_2 such that (1) has a unisolution x on (t_1,t_2), with the following property:

If $t_2 < T_2$, we have $x(t) \to y$ $(t \to t_2)$ for some $y \in \partial S$. If $t_1 > T_1$, we have $x(t) \to y$ $(t \to t_1)$ for some $y \in \partial S$.

We need for the proof an auxiliary result:

LEMMA. If $f \in \mathrm{Li}(S \times (T_1,T_2) \to R^n)$, $g \in C^1(S \to R^1)$, $a \in S$ and $g(a) = 0$, and if for some $\delta > 0$ we have $(\partial g(a))f(a,t) \geq \delta$ (a.e.), then there exist t_1, t_2 such that (1) has a unisolution on (t_1,t_2), satisfying

$$\mathrm{sgn}(g(x(t))) = \mathrm{sgn}(t - T_0) \qquad (t_1 < t < t_2) \ . \tag{7}$$

PROOF. According to theorem 1, there exists a unisolution of (1) on some interval (t_1',t_2'). Therefore we only have to prove that (7) holds on some subinterval (t_1,t_2) of (t_1',t_2') containing T_0. Let ε_0 be a positive number such that $y \in S$ holds, whenever $|y - a| \leq \varepsilon_0$. Furthermore, assume that $T_1 < t_1'$, $t_2' < T$. Then according to definition 2 there exists $M > 0$ such that for almost all $t \in (t_1',t_2')$ we have $|f(y,t)| \leq M$, $|f(y,t) - f(a,t)| \leq M|y - a|$ for $|y - a| \leq \varepsilon_0$. Let $\varepsilon := \min\{\varepsilon_0, \frac{1}{2}\delta M^{-1}(1 + |\partial g(a)|)^{-1}\}$. Then, since x and ∂g are continuous there exist t_1, t_2 with $t_1' \leq t_1 < T_0 < t_2 \leq t_2'$ such that $|x(t) - a| < \varepsilon$ and $|\partial g(x(t)) - \partial g(a)| < \varepsilon$ for $t \in (t_1,t_2)$. Since g is continuously differentiable (and hence Lipschitz continuous) and x is absolutely continuous, it follows from ([7] p. 51) that $t \mapsto g(x(t))$ is absolutely continuous. Furthermore, for almost all $t \in (t_1,t_2)$ we have $\frac{d}{dt}\, g(x(t)) = (\partial g(x(t)))f(x(t),t)$. Hence,

$$g(x(t)) = \int_{T_0}^{t} (\partial g(x(s)))f(x(s),s)ds \qquad (t \in (t_1,t_2)) \ .$$

Now it follows from

$$(\partial g(x(s)))f(x(s),s) = (\partial g(a))f(a,s) -$$
$$- (\partial g(a) - \partial g(x(s)))f(x(s),s) - (\partial g(a))(f(a,s) - f(x(s),s))$$

that

$$(\partial g(x(s)))f(x(s),s) \geq \delta - \varepsilon M - |\partial g(a)|M\varepsilon \geq \delta/2 \quad \text{(a.e.)} \ .$$

This yields

$$g(x(t)) \geq \tfrac{1}{2}\delta(t - T_0) \qquad (t \in (T_0,t_2))$$

and

$$g(x(t)) \leq \tfrac{1}{2}\delta(t - T_0) \qquad (t \in (t_1,T_0)) .$$

REMARK 5. If it is given that the inequality

$$(\partial g(a))f(a,t) \geq \delta \tag{7a}$$

holds a.e. on $[T_0,t_2)$ instead of (t_1,t_2), we can derive (7) with $(T_0 \leq t < t_2)$ instead of $(t_1 < t < t_2)$ and a similar remark applies if (7a) holds a.e. on $(t_1,T_0]$.

PROOF of i). If $a \in G_i$ ($i = 1$ or $i = 2$), the conditions of theorem 1 are satisfied with G_i instead of S. Therefore we can assume that $a \in G_0$. Since in that case we have $a \in G_0^* \cap G_1^* \cap G_2^*$ there exist unisolutions x_i of the equations

$$\dot{x} = f_i(x,t) \ , \quad x(T_0) = a \tag{1$_i$}$$

on some interval (t_1,t_2) for $i = 1,2,3$. Assume that (5) holds for $c = a$. Then it follows from the lemma that $g(x_1(t)) > 0$ $(T_0 < t < t_2)$ and $g(x_2(t)) < 0$ $(t_1 < t < T_0)$. Therefore the function x, defined by $x(t) = x_1(t)$ $(T_0 \leq t < t_2)$, $x(t) = x_2(t)$ $(t_1 < t < T_0)$ is a solution of (1) on (t_1,t_2). If (6) holds, the argument is analogous. (Here the lemma is applied with $-g$ in place of g.) This completes the proof of i).

PROOF of ii). First we show that for every $a \in S$ there exists a unisolution of (1) on some interval (t_1,t_2). If $a \notin G_0$, the existence of a unisolution follows from theorem 1. Therefore we assume that $a \in G_0$. Let $\bar{x}$ be the solution of (1) constructed in the proof of i) and let (t_1',t_2') be the domain of $\bar{x}$. We may assume that $T_1 < t_1' < T_0 < t_2' < T_2$ holds. Let us assume that (5) is satisfied for $c = a$. Then there exists a neighborhood O of a such that for almost all $t \in (t_1',t_2')$ we have $h_1(c,t) \geq \delta \wedge h_2(c,t) \geq \delta$ $(c \in O \cap G_0)$. In fact, the function

$$c \mapsto \int_{t_1'}^{t_2'} h_1(c,t)dt$$

is continuous and therefore has constant sign in some neighborhood of a. We may assume that $\bar{0}$ is bounded and contained in S. Then by definition 2, f is bounded on $0 \times (t_1',t_2')$ and hence there exist t_1, t_2 with $t_1' \leq t_1 < T_0 < t_2 \leq t_2'$ such that $x(t) \in 0$ $(t \in (t_1,t_2) \cap (t_3,t_4))$ for all t_3, t_4 and all solutions x of (1) on (t_3,t_4). We show that the restriction of $\bar{x}$ to (t_1,t_2) is a unisolution. Let x be an arbitrary solution of (1) on (t_3,t_4). We may assume that $t_1 \leq t_3$ and $t_4 \leq t_2$ hold, since otherwise we can restrict x to $(t_1,t_2) \cap (t_3,t_4)$, and we prove that $x(t) = \bar{x}(t)$ for $t_3 < t < t_4$. It is sufficient to show that the conditions

$$g(x(t)) > 0 \qquad (T_0 < t < t_4) \ , \tag{8}$$

$$g(x(t)) < 0 \qquad (t_3 < t < T_0) \tag{9}$$

hold. In fact, from (8) and (9) it follows that x(t) satisfies $(1)_1$ a.e. on $[T_0,t_4)$ and $(1)_2$ a.e. on $(t_3,T_0]$. If we apply Remark 4 to $(1)_1$ and $(1)_2$ the result follows immediately.

In order to prove (8), suppose that $x(\bar{t}) \in G_2$ for some $\bar{t} \in (T_0,t_4)$. Let $t_0 := \max\{t \leq \bar{t} \mid x(t) \in G_0\}$ and let $c := x(t_0)$. Then $\bar{x}$ satisfies $(1)_2$ (a.e.) on $[t_0,\bar{t})$ (with a replaced by c). According to Remark 5 we therefore have $g(\bar{x}(t)) > 0$ $(t_0 < t < t')$ for some $t' \in (t_0,\bar{t})$, and this is a contradiction. It follows that we have $g(x(t)) \geq 0$ $(T_0 < t < t_4)$. Furthermore, we cannot have $g(x(t)) = 0$ on an interval. Otherwise we would have

$$\left|\frac{d}{dt}\, g(x(t))\right| = |h_0(x(t),t)| \geq \delta$$

a.e. on that interval, and this is impossible. Therefore, if for some $\bar{t} \in (T_0,t_4)$ we have $g(x(\bar{t})) = 0$, then there exist t', t" with $T_0 < t' < t'' \leq \bar{t}$ such that $g(x(t'')) = 0$ and $g(x(t)) > 0$ $(t' \leq t < t'')$. Then x satisfies $(1)_1$ a.e. on $(t',t'']$ and by Remark 5 (with T_0 replaced by t'') we have $g(x(t)) < 0$ on some interval (t_0,t''). This is a contradiction and hence (8) is proved. The statement (9) can be proved similarly, and so the existence of a unisolution is established.

We are going to prove that every solution of (1) is a unisolution. Let x be an arbitrary solution of (1) on (t_1,t_2) and suppose that x is not a unisolution. Then there exists a solution $\bar{x}$ of (1) on some interval (t_3,t_4) such that we have $x(\bar{t}) \neq \bar{x}(\bar{t})$ for some $\bar{t} \in (t_1,t_2) \cap (t_3,t_4) =: (t_5,t_6)$. Since

$$\Theta := \{t \in (t_5,t_6) \mid x(t) = \bar{x}(t)\}$$

is not empty ($T_0 \in \Theta$) and $\Theta \neq (t_5,t_6)$, there is a boundary point t_0 of Θ in (t_5,t_6). The foregoing result (with T_0 replaced by t_0) yields that there exists a unisolution x^* of (1) on some interval (t_1',t_2') with $T_1 \leq t_1' < t_0 < t_2' \leq T_2$. But then we have $x^*(t) = x(t)$ and $x^*(t) = \bar{x}(t)$ on $(t_1',t_2') \cap (t_5,t_6)$ contradicting the definition of t_0. This completes the proof of ii).

PROOF of iii). Let us first assume that f is bounded on $S \times (T_1,T_2)$ ($|f(x,t)| \leq M_0$, say). Since every solution of (1) is a unisolution, it follows by Remark 3 that there exists a unisolution x on a maximal domain (t_1,t_2). We show that $x(t) \to y$ $(t \to t_2)$ for some $y \in \bar{S}$. Indeed, we have for $t > s$:

$$|x(t) - x(s)| = \left| \int_s^t f(x(\tau),\tau)d\tau \right| \leq M_0(t - s) .$$

Hence, if $\{s_n\}$ is a sequence with $s_n \uparrow t_2$ $(n \to \infty)$, then $\{x(s_n)\}$ is a Cauchy sequence which has a limit $y \in \bar{S}$. It is easily seen (for example, by mixing two sequences) that y does not depend on the sequence $\{s_n\}$ and also that $x(t) \to y$ $(t \uparrow t_2)$. If $y \in S$ and $t_2 < T_2$, we can apply ii) with (a,T_0) replaced by (y,t_2). It follows that there exists a unisolution $\bar{x}$ on some interval (t_1',t_2') containing t_2. Then the function x^* defined by $x^*(t) = x(t)$ $(t \in (t_1,t_2))$,

$x^*(t) = \bar{x}(t)$ $(t \in [t_2, t_2'))$ is a unisolution whose domain is larger than the domain of x and this is a contradiction. In a similar way, we can prove that if $t_1 > T_1$, then we have $x(t) \to y$ $(t \downarrow t_1)$ for some $y \in \partial S$.

If we have $|f(x,t)| \le M_1|x| + M_2$ $(x \in S,\ T_1 < t < T_2)$, then a solution of (1) on some interval (t_1,t_2) satisfies

$$|x(t)| \le |a| + \int_{T_0}^{t} (M_1|x(s)| + M_2)ds \le a + M_2T_2 + M_1 \int_{T_0}^{t} |x(s)|ds .$$

Together with Gronwall's lemma (see [15] p. 19) this implies

$$|x(t)| \le (a + M_2T_2)e^{M_1(t-T_0)} \le R_2 \qquad (T_0 < t < t_2)$$

where $R_2 := (a + M_2T_2)e^{M_1(T_2-T_0)}$. Similarly we have $|x(t)| \le R_1$ $(t_1 < t < T_0)$ for every solution x, where $R_1 := (a + M_2T_1)e^{M_1(T_0-T_1)}$. Therefore we can apply the foregoing result, with S replaced by

$$S_1 := \{y \in S \mid |y| < 1 + \max(R_1,R_2)\} ,$$

since on this set f is bounded.

I.5. *A theorem for linear control systems*

In this section we consider the control system

$$\dot{x}(t) = A(t)x(t) + f(t,u(t)) , \qquad (1)$$

$$x(0) = a .$$

Here A is a continuous n × n matrix-valued function on $[0,\infty)$ and $f \in C([0,\infty) \times U \to R^n)$ where U is a compact set in R^m. The set $\Omega := L([0,\infty) \to U)$ is the set of *control functions*. A *trajectory* is a solution of (1) (in the sense of Definition 4.1) corresponding to some control u, and is denoted by x_u. Hence, $x_u \in ([0,\infty) \to R^n)$. Now let $b \in R^n$, then $u \in \Omega$ is called *admissible* if there exists $t \ge 0$ such that $x_u(t) = b$. Furthermore, an admissible control $\bar{u}$ is *optimal* (or *time-optimal*) if there exists a $\bar{t} \ge 0$ such that

$x_{\bar{u}}(\bar{t}) = b$, and such that for every $u \in \Omega$ and $t \in [0,\bar{t})$ we have $x_u(t,a) \neq b$. So, the quantity $\bar{t}$ is the minimum time at which the state variable can reach the point b. Now we have the following fundamental result:

THEOREM. With the foregoing notation we have

i) If there exists an admissible control, then there exists an optimal control.

ii) If $\bar{u} \in \Omega$ is an optimal control with minimal time $\bar{t}$, then there exists a non-trivial solution $\psi \in ([0,\bar{t}] \to R^n_T)$ of the *adjoint equation*:

$$\dot{\psi}(t) = -\psi(t)A(t) \ , \tag{2}$$

such that

$$\psi(t)f(t,\bar{u}(t)) = \max_{v \in U} \psi(t)f(t,v) \qquad \text{(a.e.)} \ .$$

The existence theorem i) is in its general form due to L.W. Neustadt (see [8] for a proof). The necessary condition ii) is the *maximum principle* of Pontryagin for this special case (see [1] Chapters IV, V; [2] Chapters I, II, III, where a proof of ii) can be found).

When applying this theorem it is not necessary in general to establish a priori the existence of an admissible control. Rather one tries to find directly admissible controls which satisfy the maximum principle. If there exists no admissible control which satisfies the maximum principle, then there does not exist an optimal control, and hence (by i)) no admissible control exists at all. If there do exist admissible controls satisfying the maximum principle, one of them has to be an optimal control.

CHAPTER II

THE YO-YO EQUATION

II.1. *Derivation of the equation*

A simple example of a system with a discontinuous right-hand side is provided by the differential equation satisfied by the idealized yo-yo.

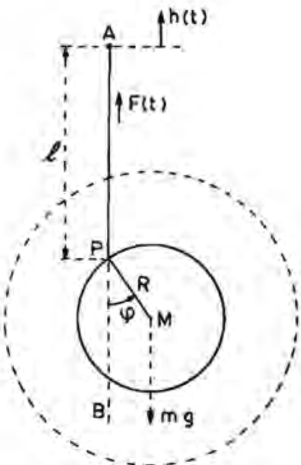


fig 1. $|\varphi| < \pi/2$

fig 2. $|\varphi| \geq \pi/2$

Let M be the center of a cross-section of the axle of the yo-yo, let R be the radius of the yo-yo and let P be the point where the string is fixed to the axle (see fig. 1). We denote by ϕ the angle between PM and the downwards directed vertical line through the point P (that is, the line PB). Here we agree that ϕ is positive if M is to the right of PB and negative if M is to the left of PB. (If $|\phi| \geq \frac{\pi}{2}$, then a piece of the string is wound on the axle.)

An upward-directed vertical force $F(t)$ is acting on the string. The upper end A of the string has the height $h(t)$.

In the following calculations we neglect the horizontal velocity of the yo-yo and we assume that the string is always vertical so that the horizontal movements do not influence the vertical movements.

First suppose that $|\phi| \le \pi/2$. Then the height of M is $h(t) - \ell - R\cos\phi(t)$, where ℓ is the length of the string. Applying Newton's law concerning the forces and accelerations in the vertical direction we obtain:

$$F(t) - mg = m[\ddot{h}(t) + R\ddot{\phi}(t)\sin\phi(t) + R\dot{\phi}^2(t)\cos\phi(t)] \, . \qquad (1)'$$

From the law concerning the relation between moment and impulse we get:

$$RF(t)\sin\phi(t) = -J\ddot{\phi}(t) \, , \qquad (2)'$$

where J is the moment of inertia of the yo-yo.

If $|\phi| \ge \pi/2$ holds, the corresponding formulas are

$$F(t) - mg = m[\ddot{h}(t) + R\ddot{\phi}\,\mathrm{sgn}\,\phi] \, , \qquad (1)$$

$$RF(t)\,\mathrm{sgn}\,\phi(t) = -J\ddot{\phi}(t) \, . \qquad (2)$$

If $\dot{\phi}$ is sufficiently large for $|\phi| \le \pi/2$, the time intervals on which we have $|\phi| \le \pi/2$ are very small. We will neglect these intervals so that we only have to deal with equations (1) and (2).

We make this more precise in the following way: Let the sets A, B, C be defined by

$$A := \{t \ge 0 \mid |\phi(t)| \le \pi/2\} \, ,$$
$$B := \{t \ge 0 \mid \phi(t) > \pi/2\} \, ,$$
$$C := \{t \ge 0 \mid \phi(t) < -\pi/2\} \, .$$

Assume that the yo-yo crosses the origin infinitely often with $\dot{\phi}$ large. Then there exist monotonic sequences $\{t_i\}$ and $\{s_i\}$ such that A, B and C have the following form (assuming $\phi(0) = \pi/2$, $\phi(0) < 0$)

$$A = \bigcup_{n=1}^{\infty} [s_i, t_i] \ , \quad B = \bigcup_{n=0}^{\infty} (t_{2n}, s_{2n+1}) \ , \quad C = \bigcup_{n=1}^{\infty} (t_{2n-1}, s_{2n}) \ .$$

Now define a function θ as follows:

$$\theta(t) := t - \mu(\{s \in A \mid s \le t\}) \ .$$

Hence

$$\theta(t) = t - \sum_{t_i < t} (t_i - s_i) \qquad (t \in B \cup C)$$

$$\theta(t) = s(t) - \sum_{t_i < t} (t_i - s_i) \qquad (t \in A)$$

where μ is the Lebesgue-measure and $s(t) := \max\{s_i \mid s_i \le t\}$. We see that θ is strictly increasing on $B \cup C$ and constant on each interval of A. In particular, we have $\theta(s_i) = \theta(t_i) =: \tau_i$. We introduce the new independent variable $\tau := \theta(t)$ on $B \cup C$, and define the functions ψ and F^*, h^* by

$$\begin{aligned} \psi(\tau) &= \phi(t) - \pi/2 && (t \in B) \\ &= \phi(t) + \pi/2 && (t \in C) \ , \\ F^*(\tau) &= F(t) && (t \in B \cup C) \ , \\ h^*(\tau) &= h(t) && (t \in B \cup C) \ . \end{aligned}$$

These functions are well-defined and satisfy the equations

$$F^*(\tau) - mg = m\left[\frac{d^2h^*}{d\tau^2} + r\,\frac{d^2\psi}{d\tau^2}\,\operatorname{sgn}\,\psi(\tau)\right] \ ,$$

$$rF^*(\tau)\ \operatorname{sgn}\,\psi(\tau) = -J\,\frac{d^2\psi}{d\tau^2} \ ,$$

for $\tau > 0$ except at the points $\tau_1, \tau_2, \ldots$. It is easily seen that the function ψ is continuous for all $\tau > 0$, if we define $\psi(\tau_i) := 0$ for $i = 1,2,\ldots$. The behaviour of $d\psi/d\tau$ at τ_i depends on $F(t)$ $(s_i \le t \le t_i)$. In general, $d\psi/d\tau$ has a jump at τ_i equal to $\dot\phi(t_i) - \dot\phi(s_i)$. It follows from (2)' that

$$J[\dot{\phi}^2(t_i) - \dot{\phi}^2(s_i)] = \pm R \int_{-\pi/2}^{\pi/2} F(t) \sin\phi \, d\phi .$$

Therefore, if F is constant on $[s_i,t_i]$, we have $\dot{\phi}(t_i) = \dot{\phi}(s_i)$. We will assume henceforth that on the intervals where $|\phi| \le \pi/2$ holds, F is constant. Then it follows that $d\psi/d\tau$ is continuous at the junction points τ_i. Thus, we have the equation:

$$\frac{d^2\psi}{d\tau^2} = - v(\tau) \operatorname{sgn} \psi(\tau) \tag{3}$$

where $v(\tau) := J^{-1}RF^*(\tau)$.

We can consider v a control function here. It is obvious that $v \ge 0$ must hold. If we impose the further restriction $v(\tau) \le M$ for some positive M on the control function, we are led to optimal control problems of the type studied in this chapter. With this constraint we can normalize (3) by the substitutions:

$$v =: \tfrac{1}{2} M(1 + u) \ ; \quad \psi =: \tfrac{1}{2} Mx .$$

Writing t instead of τ we then get what we call the *yo-yo equation*:

$$\ddot{x} + (1 + u) \operatorname{sgn} x = 0 \tag{4}$$

where u satisfies $|u(t)| \le 1 \quad (t \ge 0)$. In this chapter we pose the somewhat more general restriction $|u(t)| \le \alpha$ where α is some number in $(0,1]$.

Instead of F^* (or equally v) we can consider h^* as a control function. This does not alter the situation, however, because eliminating $\ddot{\phi}$ in (1) and (2) gives

$$F^* = Jm(J + mR^2)^{-1}(g + \ddot{h}^*) .$$

By substituting this relation into (3) we obtain an equation which also can be normalized to (4).

We give two other examples which give rise to equation (4).

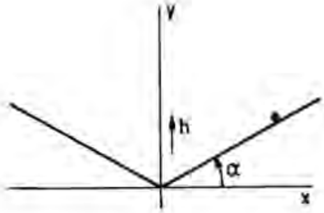


fig 3.

i) Consider a particle which moves on a V-shaped configuration shown in fig. 3. Here the y-axis indicates the vertical direction. The configuration can be translated upwards and downwards. Equivalent to this system is the system at rest, with variable gravitation. Then the gravitation acting on the particle has a vertical direction and equals

$$F = mg + m\ddot{h} .$$

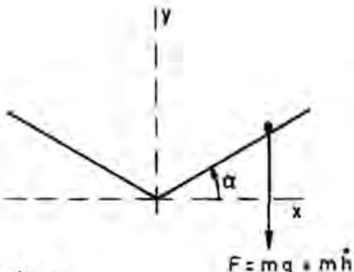


fig 4.

Let (x,y) denote the coordinates of the particle. The component of the force in the direction of the line ℓ towards the origin is F sin α. The component orthogonal to ℓ is compensated by the normal force on ℓ provided $F \geq 0$. Now according to Newton's law, the equation of motion of the particle is given by:

$$m\ddot{x} + F \sin\alpha \cos\alpha = 0 \qquad (x > 0)$$

$$m\ddot{x} - F \sin\alpha \cos\alpha = 0 \qquad (x < 0) .$$

Hence,

$$\ddot{x} + F \sin\alpha \cos\alpha \operatorname{sgn} x = 0 . \tag{5}$$

If we pose the condition that $\ddot{h} \le M$ holds for some $M \ge 0$ equation (5) can easily be transformed to (4) with $|u| \le 1$ and $u = c_1 + c_2\ddot{h}$ for some constants c_1, c_2.

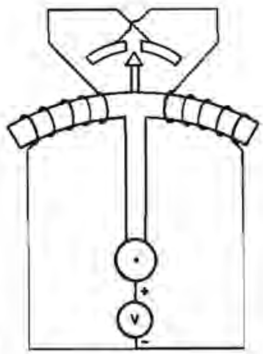


fig 5.

ii) As a final example of a system which gives rise to equation (4), we mention (without derivation) the electrical relay system given in fig. 5, where the voltage v acts as a control (see [9] p. 56).

II.2. *Behaviour of the trajectories, summary of Chapter II*

We write equation (1.4) in the form of a system

$$\begin{aligned} \dot{x} &= y , \\ \dot{y} &= -(1 + u) \operatorname{sgn} x . \end{aligned} \tag{1}$$

In vector form we have $\underline{\dot{x}} = \underline{f}(\underline{x},u)$ where $\underline{x} := (x,y)'$ and $\underline{f}(\underline{x},u) := (y, -(1 + u) \operatorname{sgn} x)'$. Control functions of (1) are elements of $L([0,\infty) \to [-\alpha,\alpha])$ where α is some number in $(0,1]$. For a given point $\underline{a} \in R^2$ and control functions u we denote by $t \mapsto \underline{x}_u(t,\underline{a})$ the solution of (1) corresponding to u with $\underline{x}(0) = \underline{a}$. In order to prove the existence and uniqueness of the solution we use theorem (I.4.2, iii). Let the control u be fixed now. Then we apply theorem (I.4.2, iii) with $S := \{\underline{x} \in R^2 \mid \underline{x} \neq \underline{0}\}$, $f := ((\underline{x},t) \mapsto \underline{f}(\underline{x},u(t)))$ and $g := (\underline{x} \mapsto x)$. Then the functions f_i $(i = 0,1,2)$ are defined by

$$f_0(\underline{x},t) := (y,0)' \;;\;\; f_1(\underline{x},t) := (y,-(1+u))' \;;\;\; f_2(\underline{x},t) := (y,1+u)' \,.$$

It follows that $h_i(\underline{c},t) := \partial g(\underline{c}) f_i(\underline{c},t) = c_2$ $(i = 0,1,2)$ where $\underline{c} = (c_1,c_2)$. We have $c_1 = 0$, $c_2 \neq 0$ on G_0 (since $\underline{0} \notin S$). We conclude from theorem (I.4.2,iii) that for every $T > 0$ and each $\underline{a} \in S$ there exists $T_0 \leq T$ such that there is a solution of (1) with $\underline{x}(0) = \underline{a}$ on $[0,T_0)$ and where either $T_0 = T$ or $\underline{x}(T_0) = \underline{0}$. However, in this last case we can extend the trajectory by setting $\underline{x}(t) = \underline{0}$ $(T_0 \leq t < T)$. Thus it is easily seen that we obtain a solution of (1) on $[0,T)$. Since this can be done for every $T > 0$ and since the corresponding trajectories (with the same control u and initial value $\underline{a}$) coincide on common intervals, we see that there exists a solution of (1) on $[0,\infty)$ for every $(\underline{a},u)$. It should be remarked that the trajectory given here is not necessarily a unisolution (definition I.4.2). If $\underline{x}(t_1) = \underline{0}$ for some $t_1 > 0$ we sometimes can extend the trajectory on $[t_1,\infty)$ in a way which differs from the trivial one given above. We will give an example of such a trajectory in section 7. However, if we have $\underline{x}(t) \neq \underline{0}$ $(t \geq 0)$ then it follows from theorem I.4.3,iii that $\underline{x}$ is a unisolution.

We are going to describe now the general behaviour of the trajectories. First suppose that $0 < \alpha < 1$. Let u be an arbitrary control. In the first quadrant x is increasing and y is decreasing $(\dot{y} \leq -1+\alpha)$ and each trajectory will intersect the positive x-axis within a finite time interval. In the fourth quadrant, y is still decreasing, now x is also decreasing, and the negative y-axis will be crossed.

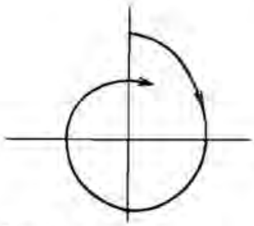


fig 1.

In the left half plane the situation is analogous. We see that every trajectory starting at a point $\neq (0,0)$ winds around $(0,0)$ infinitely often (as we will see later, possibly in a finite time interval). If $\alpha = 1$, then for an arbitrary control u a trajectory $\underline{x} = (x,y)$ will have a non-increasing y and a strictly increasing x coordinate in the first quadrant, and a non-increasing y and a strictly decreasing x in the fourth quadrant. The behaviour in the second and third quadrant can be obtained again by reflection with respect to the origin. It is possible that y is constant on a trajectory or on a piece of it. It is even possible that $\underline{x}$ is constant on a trajectory (for $y = 0$, $u = 1$).

REMARK 1. The properties of the trajectories given here are intuitively obvious and can be proved rigorously using theorem (I.4.2, iii).

It follows from these considerations that (for an arbitrary $\alpha \in (0,1]$) the origin cannot be attained by a trajectory which remains wholly in either the right or the left half plane. Yet, the origin can be reached from each point in the plane by some trajectory as will follow from the results of sections 5 and 6.

System (1) has some symmetry and homogeneity properties which will be important in the sequel. These properties are expressed in the following theorem:

THEOREM 1. If $\underline{x} = (t \mapsto (x(t),y(t))')$ is a trajectory on $[0,T]$, then the following functions are also trajectories on $[0,T]$:

i) $t \mapsto (-x(t),-y(t))'$;

ii) $t \mapsto (x(T-t),-y(T-t))'$.

Furthermore, if $\rho > 0$, then $t \mapsto (\rho^2 x(\rho t),\rho y(\rho t))'$ is a trajectory on

$[0,\rho T]$, and if t_0 is a real number, then $t \mapsto (x(t_0+t),y(t_0+t))'$ is a trajectory on $[t_0,t_0+T]$.

The theorem can be verified by substituting the functions defined here into (1).

Now we give a summary of the contents of the remaining part of this chapter. We will restrict ourselves to the case $0 < \alpha < 1$ in sections 3 to 9, and we consider the case $\alpha = 1$ in section 10. We will investigate several problems in this chapter, and define several optimality criteria which all give rise to different kinds of optimal controls. Since sometimes we have to refer to other sections we must distinguish carefully between several kinds of optimality. Therefore we start every section by stating a certain problem P_k containing a number of parameters $(\lambda_1,\ldots,\lambda_r)$ and we will call the corresponding optimal controls and trajectories "$P_k(\lambda_1,\ldots,\lambda_r)$-optimal".

In section 3 we calculate the control which transfers a point $(0,p)$ on the positive y-axis in minimal time to a point $(0,-q)$ on the negative y-axis via the right half-plane. Since sgn x does not change in the right half-plane, the problem is one of the type described in section I.5. We will see in section 3 that it is possible to attain the point $(0,-q)$ from $(0,p)$ via the right half-plane if and only if $\beta^{\frac{1}{2}} \le r \le \beta^{-\frac{1}{2}}$ where $r := q/p$ and $\beta := (1-\alpha)/(1+\alpha)$, and that the minimal time is equal to $T = f(r)p$, where $f \in C([\beta^{\frac{1}{2}},\beta^{-\frac{1}{2}}] \to R^1)$ is a positive, strictly convex function. By theorem 1 it is evident that the same results hold for the left half-plane (from the negative to the positive y-axis). The results of section 3 are used in all subsequent sections of this chapter.

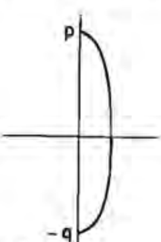


fig 2.

DEFINITION 1. If $p > 0$ and N is natural, a trajectory $\underline{x}$ is called a *p-start, N-step trajectory* on $[0,T]$ if $y(0) = p$ and if the function x has exactly N+1 zeros on $[0,T]$, denoted by $t_0,t_1,\ldots,t_N$, and satisfying $0 = t_0 < t_1 < \ldots < t_N = T$. The numbers t_k will be called *alteration times*, the points $(0,y(t_k))$ *alteration points* and the numbers $p_k := (-1)^k\, y(t_k)$ *alteration moduli*. Obviously, we have $p_0 = p$, $p_k > 0$ $(k = 0,\ldots,N)$. The quantity T is called the *final time*. If we do not want to specify p, we will omit the term "p-start".

We will call (see Definition 3.1) an N-step, p-start trajectory *piecewise optimal*, if the parts of it between two consecutive intersection points are optimal in the sense of section 3. Therefore, if we denote the ratio p_{k+1}/p_k by r_k, then for a piecewise optimal control the time used for the k-th step $(k = 0,\ldots,N-1)$, is equal to $f(r_k)p_k$ and the final time is given by $\sum_{k=0}^{N-1} f(r_k)p_k$.

In section 4 we determine a p-start, N-step control with minimal final time. It is shown there that such a control exists and is piecewise optimal. The differential optimization problem is reduced by this result to a discrete optimization problem. In fact, now the problem is: "Find a sequence $\{r_0,\ldots,r_{N-1}\}$ with $\beta^{\frac{1}{2}} \le r_k \le \beta^{-\frac{1}{2}}$ $(k = 0,\ldots,N-1)$ such that $\sum_{k=0}^{N-1} f(r_k)p_k$ is minimal, where $\{p_k\}_0^N$ is given by $p_0 := p$, $p_{k+1} := r_kp_k$ $(k = 0,\ldots,N-1)$". This problem is solved in the following way:

We assume that the problem with N-1 steps is already solved. Then we perform one step and for the remaining part we use the optimal sequence of the problem with N-1 steps. Finally, we choose the first step in an optimal way. This method is called by R. Bellman the *dynamic programming method* (see [4]).

In section 5 the problem is to find a control u such that the origin is reached from a point on the positive y-axis in minimal time. Since the origin cannot be attained from either the right half-plane or the left half-plane, the desired trajectory has to intersect the y-axis infinitely often. (We call such a trajectory a *p-start,∞-step trajectory*.) It is not directly clear in this case that an optimal control exists. But, it is shown in section 5 that we can restrict our attention to piecewise optimal controls also in this case. Therefore, we now have the following discrete problem:

"Find a sequence $\{r_0,r_1,\ldots\}$ such that $r_k \in [\beta^{\frac{1}{2}},\beta^{-\frac{1}{2}}]$ and such that for the sequence $\{p_0,p_1,\ldots\}$ defined by $p_0 = p$, $p_{k+1} = r_k p_k$ $(k = 0,1,\ldots)$ we have $p_k \to 0$ $(k \to \infty)$ and $\sum_{k=0}^{\infty} f(r_k)p_k$ is convergent and minimal". Assuming the existence of an optimal control we can easily find the solution by the same method as in section 4. Afterwards it is proved that the solution is actually an optimal control. It also turns out that the optimal control is unique. (This is also the case for the optimal control of section 4.)

As is shown in section 6, the optimal control of section 5 can conveniently be described graphically in the (x,y)-plane. Then it turns out that we can generalize the result of section 5 to the case of an arbitrary starting point (a,b) in the phase plane instead of a point on the y-axis as was required in section 5. The graphical solution makes it clear, that the optimal control is a feedback control, that is, u is given as a function of x and y instead of t, a, and b. A similar, though more complicated, graphical solution can be found for the problem treated in section 4.

In section 7 we will consider the problem of finding a control which maximizes $|\underline{x}(t)/t|$. For definiteness we will restrict ourselves to p-start, N-step controls. Then the problem can be stated as follows: "Find a p-start, N-step control which maximizes the quantity p_N/T, where p_N is the final alteration modulus and T the final time". In mechanical terms, the optimal control is the one which maximizes the average acceleration (note that p has the dimension of velocity). It turns out that the optimal control is strongly related to the one of section 4.

In section 8 we are going to consider the following problem: "Given positive real numbers p, q and a natural number N, find a p-start, N-step control for which we have $p_N = q$ and such that T is minimal". We will see that the fact that the endpoint of the trajectory is prescribed, complicates the problem a great deal. The dynamic programming method used so far in this chapter yields a complicated recurrence relation for the solution. A more useful result can be obtained, either by the discrete maximum principle or by means of an auxiliary problem treated in section 9.

The problem treated in section 9 is the following one: "Given real numbers p, ρ, with p > 0, and a natural number N, find a p-start, N-step control which maximizes the quantity $p_N + \rho T$ where p_N is the final alteration modulus and T is the final time". The problem can be solved in the same way as the problem of section 4. It is shown in section 9 that the solution of this problem also yields the solutions of the problems of sections 4, 7, 8.

In section 10 we will discuss the results of this chapter for the case $\alpha = 1$.

II.3. *A basic optimal control problem for the yo-yo equation*

$P_1(p,q)$: "Given positive numbers p and q, find a control $\bar{u}$ and a number $\bar{t}$ such that

$$\underline{x}_u(t_1,(0,p)') = (0,-q)' , \tag{1}$$

$$x_u(t,(0,p)') > 0 \qquad (0 < t < t_1) \tag{2}$$

holds for $u = \bar{u}$ and $t_1 = \bar{t}$ and such that for all controls u and numbers t_1 satisfying (1) and (2) we have $t_1 \geq \bar{t}$."

Here, as defined in section 2, $t \mapsto \underline{x}_u(t,\underline{a})$ is the trajectory of system (2.1) corresponding to the control u and the initial value $\underline{a}$. Because of condition (2) we can write (instead of (2.1)):

$$\begin{aligned} \dot{x} &= y , \\ \dot{y} &= -(1+u) . \end{aligned} \tag{3}$$

Hence, except for condition (2), problem P_1 is of the type discussed in section I.5. Therefore, denoting by $t \mapsto \underline{x}_u(t,\underline{a})$ the trajectory of (3) satisfying $\underline{x}(0) = \underline{a}$, we replace P_1 by the following problem:

$P_1'(p,q)$: "Given positive numbers p and q, find a control $\bar{u}$ and a number $\bar{t}$ such that $x_{\bar{u}}(\bar{t},(0,p)') = (0,-q)'$, whereas for every control u and every t_1 with $\underline{x}_u(t_1,(0,p)') = (0,-q)'$ we have $t_1 \geq \bar{t}$."

It will turn out afterwards that the solutions of P_1 and P_1' coincide. In order to apply the theorem of section I.5 we write (3) in vector notation:

$$\dot{\underline{x}} = A\underline{x} + \underline{f}(u)$$

with

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \underline{f}(u) = \begin{bmatrix} 0 \\ -(1+u) \end{bmatrix}.$$

If $\bar{u}$ is a $P_1'(p,q)$-optimal control, then according to the theorem there exists a non-trivial solution $\underline{\psi} = (\phi,\psi)$ of the adjoint equation $\dot{\underline{\psi}} = -\underline{\psi}A$ (or $\dot{\phi} = 0$, $\dot{\psi} = -\phi$) such that

$$\underline{\phi}f(\bar{u}) = \max_{|v|\le\alpha} \underline{\phi}f(v)\ ,$$

that is, $\bar{u} = \alpha \operatorname{sgn} \psi$ for almost all t with $\psi(t) \neq 0$. Now ψ is linear and not identically zero (since $\underline{\psi}$ is non-trivial), and therefore ψ vanishes at no more than one point. It follows that we may assume

$$\bar{u}(t) = \alpha \operatorname{sgn} \psi(t) \qquad (0 \le t \le \bar{t})$$

since the values of $\bar{u}$ on a set of measure zero do not influence the trajectory. We conclude that $u(t) = \pm\, \alpha$, and that there is at most one alteration of sign. On the other hand, if u has this form, it satisfies the necessary conditions of theorem (I.5). We will show that there exist at most two admissible controls satisfying this property. Let us first determine the trajectories in the phase plane corresponding to constant controls u. For constant u these trajectories satisfy $y\,\frac{dy}{dx} + 1 + u = 0$, and hence

$$\tfrac{1}{2}\, y^2 + (1 + u)x = C \tag{4}$$

for some C. Moreover, for all C there is a trajectory of (3) corresponding to the constant control u for which (4) holds. Let us denote the parabola of the form (4) which intersects the y-axis in the points (0,a) and (0,-a) by $\pi(u,a)$. According to the foregoing we are interested in $\pi(\pm\,\alpha,p)$ and $\pi(\pm\,\alpha,q)$. The trajectory corresponding to the optimal control (which will be called the optimal trajectory henceforth) consists of one or two arcs of these parabolas.

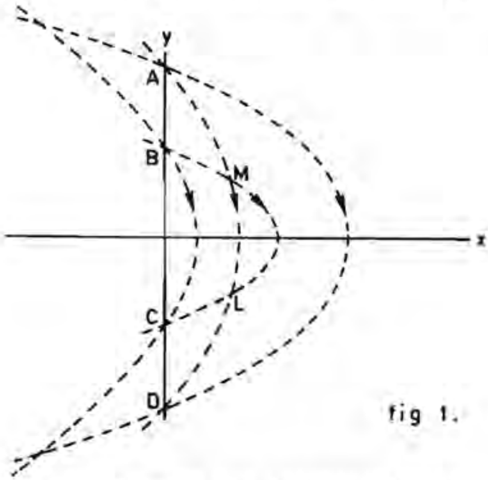


fig 1.

Therefore, we have to determine the intersection points of $\pi(\alpha,p)$ and $\pi(-\alpha,q)$ and the ones of $\pi(-\alpha,p)$ and $\pi(\alpha,q)$. But, since y is decreasing ($\dot{y} \le -1+\alpha$) we can only use intersection points $(\bar{x},\bar{y})$ with $-q \le \bar{y} \le p$, or equivalently, intersection points with $\bar{x} \ge 0$. We distinguish three cases:

i) $q < p$. Here we have to consider the intersections of the parabolas $\pi(\alpha,p)$ and $\pi(-\alpha,q)$. (In fig. 1: A = (0,p), C = (0,-q).) If there are intersection points they satisfy the equations:

$$\begin{aligned} \tfrac{1}{2}\,y^2 + (1+\alpha)x &= \tfrac{1}{2}\,p^2\,, \\ \tfrac{1}{2}\,y^2 + (1-\alpha)x &= \tfrac{1}{2}\,q^2\,. \end{aligned} \tag{5}$$

If $(\bar{x},\bar{y})$ is a solution of (5), we have

$$\begin{aligned} \bar{x} &= \frac{p^2 - q^2}{4\alpha}\,, \\ \bar{y}^2 &= \frac{(1+\alpha)q^2 - (1-\alpha)p^2}{2\alpha}\,. \end{aligned} \tag{6}$$

If we put

$$r := q/p \ , \tag{7}$$

$$\beta := \frac{1-\alpha}{1+\alpha} \ , \tag{8}$$

it follows that there is no intersection point if $r < \beta^{\frac{1}{2}}$. If $r = \beta^{\frac{1}{2}}$ there is exactly one intersection point (lying on the x-axis), and if $r > \beta^{\frac{1}{2}}$, there are two intersection points.

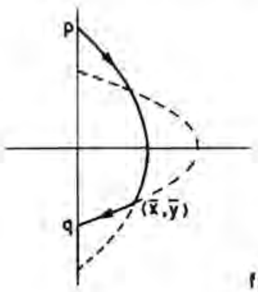


fig.2

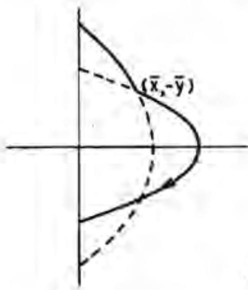


fig 3.

Using the notation of (7) and (8), the intersection points are $(\bar{x},\bar{y})$ and $(\bar{x},-\bar{y})$, where

$$\bar{x} := \tfrac{1}{4}\,\frac{1+\beta}{1-\beta}\,(1-r^2)p^2 \ , \tag{9}$$

$$\bar{y} := -p\left(\frac{r^2-\beta}{1-\beta}\right)^{\frac{1}{2}} \tag{10}$$

(note that $\beta^{\frac{1}{2}} < r < 1$).

Thus, we have found two admissible trajectories which satisfy the necessary conditions of theorem (I.5). In order to decide which one of them is an optimal trajectory we have to compare the corresponding numbers t_1 (see (1), (2)). It is easily seen that the trajectory with intersection point $(\bar{x},\bar{y})$ has final time t_1 which is minimal; indeed, we have

$$t_1 = t_1^+ := \frac{p-\bar{y}}{1+\alpha} + \frac{\bar{y}+q}{1-\alpha} \ , \tag{11}$$

whereas the other trajectory has the final time

$$t_1^- := \frac{p + \bar{y}}{1 + \alpha} + \frac{q - \bar{y}}{1 - \alpha} = t_1^+ - \frac{4\alpha\bar{y}}{1 - \alpha^2} > t_1^+$$

since $\bar{y} < 0$ and $0 < \alpha < 1$. Thus, the trajectory with intersection point $(\bar{x},\bar{y})$ is the optimal trajectory.

REMARK 1. The point $(\bar{x},\bar{y})$ is called the *switching point* of the trajectory.

ii) $q = p$. In this case it is easily seen from figure 1 that we have $\bar{u} = \alpha$ on the whole trajectory.

iii) $q > p$. It can be shown by an argument analogous to the one of case i) that the optimal trajectory now starts on $\pi(-\alpha,p)$ and ends on $\pi(\alpha,q)$. Furthermore, the switching point is the intersection point $(\bar{x},\bar{y})$ of these parabolas with $\bar{y} > 0$. The switching point is given by

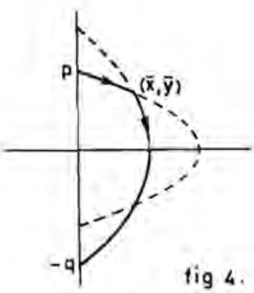


fig 4.

$$\bar{x} = \tfrac{1}{4}\,\frac{1 + \beta}{1 - \beta}\,(r^2 - 1)p^2 \ ,$$

$$\bar{y} = p\left(\frac{1 - \beta r^2}{1 - \beta}\right)^{\frac{1}{2}} \ , \tag{12}$$

where it is supposed that $1 < r \le \beta^{-\frac{1}{2}}$. It follows that there are admissible controls if and only if $r \in J$ where J is defined by

$$J := [\beta^{\frac{1}{2}},\beta^{-\frac{1}{2}}] \ . \tag{13}$$

For the same values of r there exists a unique optimal control $\bar{u}$, which is of the "bang-bang" type, that is, $\bar{u}$ assumes only the extremal values $\pm\,\alpha$. We calculate the minimal final time, which will be denoted by $T(p,q)$. Using (7) and (8) it follows that

$$T(p,q) = f(r)p \ , \tag{14}$$

where

$$f(r) = \frac{1+\beta}{2\beta}\,[r + \beta - \sqrt{(1-\beta)(r^2-\beta)}] \quad (\beta^{\frac{1}{2}} \le r \le 1) \tag{15}$$

and

$$f(r) := \frac{1+\beta}{2\beta}\,[1 + \beta r - \sqrt{(1-\beta)(1-\beta r^2)}] \quad (1 \le r \le \beta^{-\frac{1}{2}}), \tag{16}$$

where (15) follows from (11) and the latter formula can be obtained in a way analogous to the way formula (15) was obtained.

We have found now the solution of problem P_1' and since $x \ge 0$ holds on the P_1'-optimal trajectory, it follows that the same control furnishes a solution of P_1.

Some properties of f, which will be used in the sequel, will be derived now. It is easily seen from Theorem 2.1 that $T(p,q) = T(q,p)$. According to (7) and (14) this implies

$$f(r) = rf(1/r) \qquad (r \in J) \;. \tag{17}$$

It should be remarked that this symmetry property can be used for deriving (16) from (15). Furthermore, f is continuous on J and twice continuously differentiable on J except at the points $\beta^{-\frac{1}{2}}$, 1, $\beta^{\frac{1}{2}}$, and we have:

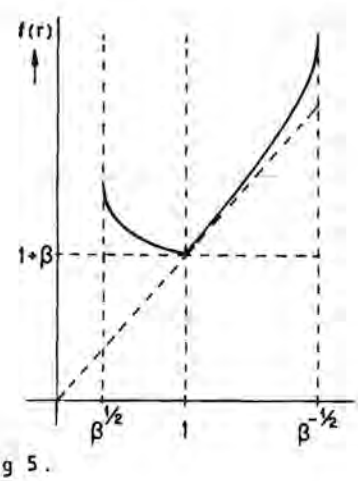


fig 5.

$$\left.\begin{array}{ll} f'(r) \to -\infty & (r \to \beta^{\frac{1}{2}}), \\ f'(r) < 0 & (\beta^{\frac{1}{2}} < r < 1), \\ f'(r) \to 0 & (r \uparrow 1), \\ f'(r) \to 1+\beta & (r \downarrow 1), \\ f'(r) > 0 & (1 < r < \beta^{-\frac{1}{2}}), \\ f'(r) \to \infty & (r \to \beta^{-\frac{1}{2}}). \end{array}\right\} \tag{18}$$

The function f is positive and has a unique minimum $1 + \beta$ at $r = 1$. Furthermore, $f''(r) > 0$ on $(\beta^{\frac{1}{2}},1)$ and $(1,\beta^{-\frac{1}{2}})$, and hence f is strictly convex.

REMARK 2. If one wants to find the control which transfers the state variable from $(0,-p)$ to $(0,q)$ via the half-plane $x < 0$ in a minimal time, then the optimal control one gets is equal to the $P_1(p,q)$-optimal control. The optimal trajectory is obtained by reflecting the $P_1(p,q)$-optimal trajectory with respect to the origin. It is obvious then, that there is an admissible control again if and only if $r := q/p \in J$, and that the minimal final time is given by $T = f(r)p$ (see also Theorem 2.1).

REMARK 3. It is clear by the autonomity of the problem, that an optimal control (in the sense of $P_1(p,q)$) with $\underline{x}(t_0) = (0,p)$ (instead of $x(0) = (0,p)$) is obtained from the $P_1(p,q)$-optimal control by a translation of time.

REMARK 4. It follows from the uniqueness of the optimal control $\bar{u}$ that $\bar{u}$ is strictly optimal; that is, every admissible control which is not equal to $\bar{u}$ almost everywhere has a larger final time.

Let $\underline{x}$ be a p-start, N-step trajectory (see Definition 2.1), and let t_k and p_k for $k = 0,\dots,N$ be defined as in Definition 2.1. Then, according to the foregoing, we have $r_k := p_{k+1}/p_k \in J$, and $t_{k+1} - t_k \geq f(r_k)p_k \quad (k = 0,\dots,N-1)$.

DEFINITION 1. A p-start, N-step control (trajectory) is called *piecewise optimal* if on each interval $[t_k, t_{k+1}]$ the control (trajectory) is optimal in the sense of P_1.

Hence, for a piecewise optimal control we have $t_{k+1} - t_k = f(r_k)p_k$. A piecewise optimal control u is for a given p completely determined by the sequence $\{r_0,\ldots,r_{N-1}\}$ where $r_k := p_{k+1}/p_k$ $(k = 0,\ldots,N-1)$. The sequence $\Lambda := \{r_0,\ldots,r_{N-1}\}$ is an element of J^N and is called a *strategy* (or, if we want to specify N it is called an "N-step strategy"). The functions $(p,\Lambda) \mapsto p_k(p,\Lambda)$, $(p,\Lambda) \mapsto t_k(p,\Lambda)$ for $p > 0$, $\Lambda \in J^N$ and $k = 0,\ldots,N$ are defined by:

$$\begin{aligned} &p_0(p,\Lambda) := p \ , \\ &p_{k+1}(p,\Lambda) := r_k p_k(p,\Lambda) \qquad (k = 0,\ldots,N-1) \ , \end{aligned} \tag{19}$$

and

$$\begin{aligned} &t_0(p,\Lambda) = 0 \ , \\ &t_{k+1}(p,\Lambda) = t_k(p,\Lambda) + f(r_k)p_k(p,\Lambda) \quad (k = 0,\ldots,N-1) \ . \end{aligned} \tag{20}$$

Furthermore, for $N > 1$ the mapping $\Lambda \mapsto \Lambda'$ of $(J^N \to J^{N-1})$ is defined by:

$$\text{"If } \Lambda = \{r_0,\ldots,r_{N-1}\} \text{ then } \Lambda' = \{r_1,\ldots,r_{N-1}\}\text{"} \ . \tag{21}$$

COROLLARY 1. Given the sequence $\{p_0,\ldots,p_{N-1}\}$ with $p_{k+1}/p_k \in J$ $(k = 0,\ldots,N-1)$, the piecewise optimal control with alteration moduli p_k $(k = 0,\ldots,N-1)$ minimizes t_N among all the p -start, N-step controls with the same alteration moduli.

COROLLARY 2. We have the following fundamental equality:

$$t_N(p,\Lambda) = f(r_0)p + t_{N-1}(r_0 p,\Lambda') \ , \qquad (N > 1) \tag{22}$$

where $\Lambda = \{r_0,\ldots,r_{N-1}\}$. In fact:

$$t_N(p,\Lambda) = f(r_0)p + \sum_{k=1}^{N-1} f(r_k)p_k$$

and

$$t_{N-1}(r_0 p,\Lambda') = \sum_{k=0}^{N-2} f(r_{k+1})p_{k+1} .$$

REMARK 5. If Λ has length 1 (that is, $\Lambda = \{r_0\}$), we say that Λ' is the empty sequence, and we write $\Lambda' = \emptyset$. We call $\emptyset$ a 0-step strategy and define $t_0(p,\emptyset) = 0$, $p_0(p,\emptyset) = p$. With this convention (22) is also true for $N = 1$.

II.4. *Time-optimal* p-*start,* N-*step control*

$P_2(p,N)$: "Given a real positive number p and a natural number N, find a p-start, N-step control u such that the final time t_N is minimal."

From Corollary 1 of the previous section we have the following result:

THEOREM 1. If $\bar{u}$ is $P_2(p,N)$-optimal, then u is piecewise optimal. On the other hand, if $\bar{u}$ is piecewise optimal and minimizes the final time among the piecewise optimal controls (with $p_0 = p$), then $\bar{u}$ is $P_2(p,N)$-optimal.

PROOF. If $\bar{u}$ is not piecewise optimal, and if $\hat{u}$ is the piecewise optimal control with the same alteration points, then the final time corresponding to $\hat{u}$ is less than the one of $\bar{u}$. The second part of the theorem is proved similarly.

Theorem 1 reduces P_2 to a discrete optimization problem:

$P_2'(p,N)$: "Find an N-step strategy Λ such that $t_N(p,\Lambda)$ is minimal."

Since J^N is compact and $\Lambda \mapsto t_N(p,\Lambda)$ is continuous, the existence of an optimal strategy is clear. The minimal time, which depends on p is denoted by $T_N(p)$. Hence:

$$T_N(p) := \min\{t_N(p,\Lambda) \mid \Lambda \in J^N\} . \qquad (1)$$

In particular, we have $T_0(p) = 0$ for every $p > 0$ (see Remark 3.5). Now let for a given N and p, $\bar{\Lambda} = \{\bar{r}_0,\dots,\bar{r}_{N-1}\}$ be an optimal strategy, so that $t_N(p,\bar{\Lambda}) = T_N(p)$. According to (3.22) we have:

$$t_N(p,\bar{\Lambda}) = f(\bar{r}_0)p + t_{N-1}(\bar{r}_0 p,\bar{\Lambda}') \ . \tag{2}$$

This implies that $\bar{\Lambda}'$ is an $P_2'(\bar{r}_0 p, N-1)$-optimal strategy. Otherwise we would be able to improve on $t_N(p,\bar{\Lambda})$. It follows that we have:

$$T_N(p) = f(\bar{r}_0)p + T_{N-1}(\bar{r}_0 p) \ . \tag{3}$$

Also, it follows from the optimality of $\bar{\Lambda}$, that $f(\bar{r}_0)p + T_{N-1}(\bar{r}_0 p) = \min_{r\in J} [f(r)p + T_{N-1}(rp)]$. Therefore we get the following recurrence relation

$$T_N(p) = \min_{r\in J} [f(r)p + T_{N-1}(rp)] \tag{4}$$

for $T_N(p)$, whereas $\bar{\Lambda}$ satisfies the following properties:

i) $\bar{r}_0$ is the value of r for which the minimum in (4) is assumed. (4a)

ii) $\bar{\Lambda}'$ is $P_2'(\bar{r}\ p, N-1)$-optimal. (4b)

It follows from (3.19) and (3.20) that $p \mapsto p_k(p,\Lambda)$ and $p \mapsto t_k(p,\Lambda)$ are homogeneous with degree 1 for $k = 0,1,\dots,N$ and for every $\Lambda \in J^N$. This implies that $p \mapsto T_N(p)$ is homogeneous. In fact, if $\rho > 0$, we have:

$$T_N(\rho p) = \min_{\Lambda\in J^N} t_N(\rho p,\Lambda) = \rho \min t_N(p,\Lambda) = \rho T_N(p) \ .$$

Therefore, if we introduce

$$S_k := T_k(1) \ , \qquad (k = 0,\dots,N) \tag{5}$$

we have $T_N(p) = S_N p$. Now (4) implies

$$S_N = \mu(S_{N-1}) \ , \tag{6}$$

where the function $\mu \in (R^1 \to R^1)$ is defined by

$$\mu(s) := \min_{r \in J} [f(r) + rs] \ . \tag{7}$$

Now we derive some properties of the function μ. Since f is strictly convex, the minimum in (7) is assumed at exactly one value of r. This value will be denoted by h(s), so that we have

$$\mu(s) = f(h(s)) + sh(s) \ . \tag{8}$$

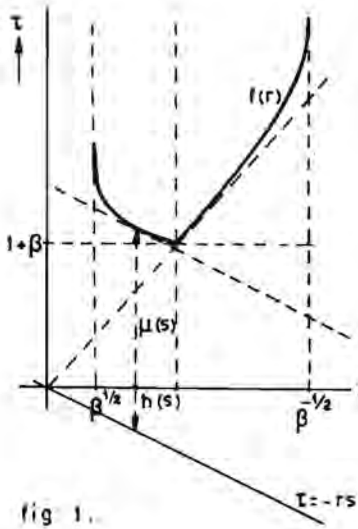


fig 1.

It follows from fig. 1, that we have $h(s) = 1 \quad (-1-\beta \le s \le 0)$, whereas for other values of s the values of the function h is determined by $f'(h(s)) = -s$. Hence

$$h(s) = \frac{1 + \beta + 2\beta s}{\sqrt{(\beta+1)^2 + 4(\beta+1)s + 4\beta s^2}} \qquad (s > 0) \ ,$$

$$h(s) = 1 \qquad (-1-\beta \le s \le 0) \ , \tag{9}$$

$$h(s) = \frac{-1 - \beta - 2s}{\sqrt{(\beta+1)^2 + 4\beta(\beta+1)s + 4\beta s^2}} \qquad (s < -1-\beta) \ .$$

And by equation (8) this yields

$$\left.\begin{aligned}
\mu(s) &= \tfrac{1}{2}\,[1+\beta+\sqrt{(\beta+1)^2+4(\beta+1)s+4\beta s^2}\,] && (s>0)\ ,\\
\mu(s) &= 1+\beta+s && (-1-\beta \le s \le 0)\ ,\\
\mu(s) &= \frac{1}{2\beta}\,[1+\beta-\sqrt{(\beta+1)^2+4\beta(\beta+1)s+4\beta s^2}\,] && (s<-1-\beta)\ .
\end{aligned}\right\} \quad (10)$$

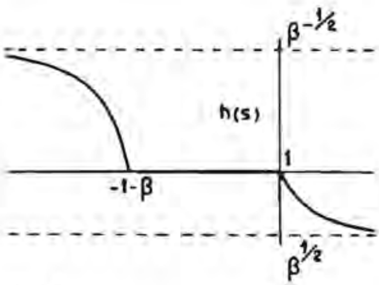


fig 2.

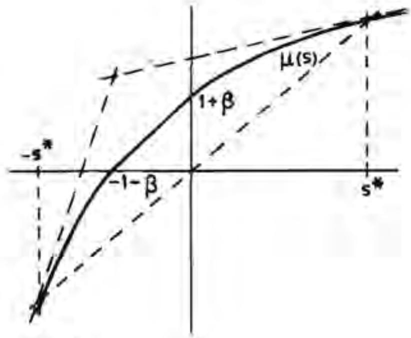


fig 3.

We have the following asymptotic formulas for μ:

$$\begin{aligned}
\mu(s) &= \beta^{+\frac{1}{2}}s+\tfrac{1}{2}(1+\beta)(1+\beta^{-\frac{1}{2}})+O(1/s) && (s\to\infty)\ ,\\
\mu(s) &= \beta^{-\frac{1}{2}}s+\tfrac{1}{2}\beta^{-1}(1+\beta)(1+\beta^{+\frac{1}{2}})+O(1/s) && (s\to-\infty)\ .
\end{aligned} \quad (11)$$

REMARK 1. Since these asymptotic formulas are starting terms of a convergent series, the asymptotic formulas of $h(s) = \mu'(s)$ (see formula (15)) can be derived by formal differentiation. Calculating one more term of the asymptotic series for μ and differentiating we obtain

$$h(s) = \beta^{\frac{1}{2}}+\frac{1}{8}\,(1-\beta)(1+\beta)^2\beta^{-3/2}s^{-2}+O(s^{-3}) \qquad (s\to\infty)\ ,$$

$$h(s) = \beta^{-\frac{1}{2}}-\frac{1}{8}\,(1-\beta)(1+\beta)^2\beta^{-3/2}s^{-2}+O(s^{-3}) \qquad (s\to-\infty)\ .$$

(We will use these formulas in section 9.)

There are exactly two points s for which $\mu(s) = s$ holds, namely,

s^* and $-s^*$ where

$$s^* := \frac{2(1+\beta)}{1-\beta} = \frac{2}{\alpha} \tag{12}$$

Let us return to problem P_2'. It follows from (4a) that $\bar{r}_0 = h(S_{N-1})$ and from (4b): $\bar{r}_1 = h(S_{N-2})$. It is easily proved by induction that

$$\bar{r}_k = h(S_{N-k-1}) \qquad (k = 0,\dots,N-1) \tag{13}$$

holds for a $P_2'(-p,N)$-optimal strategy $\bar{\Lambda} = \{\bar{r}_0,\dots,\bar{r}_{N-1}\}$. This provides the solution of problem P_2. In particular, it follows that the $P_2'(p,N)$-optimal strategy and hence the $P_2(p,N)$-optimal control is unique. The solution is indicated graphically in fig. 4.

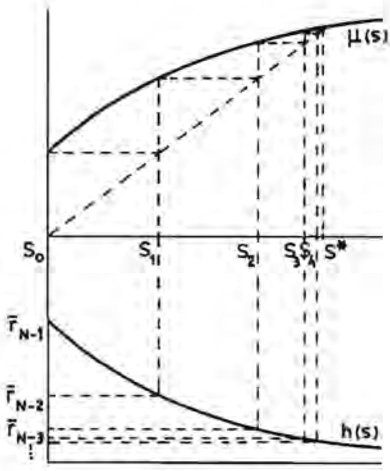


fig 4.

In section 6 we will discuss the graphical constructions in the (x,y)-plane. From the fact that μ is increasing, $\mu(s) > s$ $(0 \le s < s^*)$ and $\mu(s^*) = s^*$ it follows that

$$0 = S_0 < S_1 < S_2 < \dots < s^*$$

and

$$S_N \to s^* \qquad (N \to \infty) . \tag{14}$$

This implies that $1 = \bar{r}_{N-1} > \bar{r}_{N-2} > \dots > r^* := h(s^*)$.

REMARK 2. A number of properties of μ and h can be derived using only the following properties of f:

i) $f \in C(J \to R^1)$,

ii) for all $r \in J$ we have $1/r \in J$,

iii) f is strictly convex,

iv) $f(r) = rf(1/r) \quad (r \in J)$.

Again μ is defined by (7). Using iii), we see that the minimum in (7) is attained for exactly one value of r. This value of r is denoted by $h(s)$. It follows that we have equation (8). Now we have the following results:

LEMMA 1. $\mu'(s) = h(s)$. (15)

PROOF. For real s and δ we have

$$\mu(s+\delta) = \min_r[f(r) + r(s+\delta)] \le f(h(s)) + (s+\delta)h(s) = \mu(s) + \delta h(s)$$

and

$$\mu(s) = \min_r[f(r) + rs] \le f(h(s+\delta)) + sh(s+\delta) = \mu(s+\delta) - \delta h(s+\delta) .$$

Hence, if δ is positive

$$h(s+\delta) \le \frac{\mu(s+\delta) - \mu(s)}{\delta} \le h(s) ,$$

and if $\delta < 0$,

$$h(s) \le \frac{\mu(s+\delta) - \mu(s)}{\delta} \le h(s+\delta) .$$

Since f is strictly convex, the function h is continuous, hence (15).

From this proof it follows also that h is non-increasing and hence μ is concave.

LEMMA 2. $\mu(-\mu(-s)) = s$, (16)

that is, the graph of μ is symmetric with respect to the line $\mu = -s$.

PROOF. Let $\nu(s) := \mu(-s) = \min_{r\in J} [f(r) - rs]$. Then we have

$$\nu(\nu(s)) = \min_{r\in J} \max_{\rho\in J} [f(r) - rf(\rho) + r\rho s] = \min_{r\in J} \max_{\rho\in J} F(r,\rho) ,$$

where $F(r,\rho) := f(r) - rf(\rho) + r\rho s$. Since for all $r \in J$ we have

$$\max_{\rho\in J} F(r,\rho) \geq F(r,1/r) = s ,$$

it follows that $\nu(\nu(s)) \geq s$. Furthermore, it is well known (see [10] p. 78-80), since F is continuous on $J \times J$, $r \mapsto F(r,\rho)$ convex for all ρ and $\rho \mapsto F(r,\rho)$ concave for all r, that

$$\nu(\nu(s)) = \max_{\rho\in J} \min_{r\in J} F(r,\rho) \leq \max_{\rho\in J} F(\tfrac{1}{\rho},\rho) = s ,$$

which yields (16).

LEMMA 3. $h(-\mu(s)) = 1/h(s)$. (17)

This result is obtained by differentiating (16) and replacing s by $-s$.

II.5. *Time optimal null-control*

$P_3(p)$: "Given a real number $p > 0$, find the control u such that if $\underline{x} := (t \mapsto \underline{x}_n(t,(0,p)'))$, we have $\underline{x}(T) = \underline{0}$ for a minimal T."

Since, as we have seen in section 2, the origin cannot be reached from either the right or the left half-plane, the trajectory $\underline{x}$ has to intersect the y-axis infinitely often. We will call such a trajectory with initial point $(0,p)$, $p > 0$, a p-*start*, ∞-*step trajectory* (and the corresponding control a p-*start*, ∞-*step control*). The alteration times $\{t_0,t_1,\ldots\}$ form an infinite sequence and so do the alteration points $\{(0,y(t_k))\}_0^\infty$ and the alteration moduli $\{p_k\}_0^\infty$ with $p_k := (-1)^k y(t_k)$. Again we have $r_k := p_{k+1}/p_k \in J$ and

$t_{k+1} - t_k \geq f(r_k)p_k$. If $\lim_{k\to\infty} t_k =: T$ exists, it is called the final time, otherwise we have $t_k \to \infty$ $(k \to \infty)$ and then we set $T = \infty$. Piecewise optimal controls are defined in a similar way as in section 3. For piecewise optimal controls we have $t_{k+1} - t_k = f(r_k)p_k$. Just as in section 4, it is easily seen that we can restrict our attention to piecewise optimal controls when looking for $P_3(p)$-optimal control. Therefore we define an *∞-step strategy*, as an infinite sequence $\{r_k\}_0^\infty$ of numbers $r_k \in J$. The set of ∞-step strategies will be denoted by J^∞. Furthermore, the functions t_k, p_k, in $(R_+^1 \times J^\infty \to R^1)$ are defined by

$$p_0(p,\Lambda) = p \ ,$$
$$p_{k+1}(p,\Lambda) = r_k p_k(p,\Lambda) \ , \tag{1}$$

and

$$t_0(p,\Lambda) = 0 \ ,$$
$$t_{k+1}(p,\Lambda) = t_k(p,\Lambda) + f(r_k)p_k(p,\Lambda) \ . \tag{2}$$

for $k = 0,1,\ldots$.

Furthermore, the function $T \in (R_+^1 \times J^\infty \to R^1 \cup \{\infty\})$ is defined by: If $\lim_{k\to\infty} t_k(p,\Lambda)$ exists, then

$$T(p,\Lambda) = \lim_{k\to\infty} t_k(p,\Lambda) \ , \tag{3}$$

otherwise (if $t_k(p,\Lambda) \to \infty$ $(k \to \infty)$), we define $T(p,\Lambda) = \infty$. The problem $P_3(p)$ is equivalent to

$P_3'(p)$: "Given a real positive number p, find an ∞-step strategy Λ such that $T(p,\Lambda)$ is finite and minimal."

We have omitted the condition $p_k(p,\Lambda) \to 0$ $(k \to \infty)$ in P_3'. But this condition is evidently satisfied if

$$T(p,\Lambda) = \sum_{k=0}^{\infty} f(r_k)p_k(p,\Lambda)$$

is finite, since $f(r_k) \geq 1+\beta$. Examples of strategies Λ with $T(p,\Lambda) < \infty$ are easily given. For instance, if $\Lambda := \{r,r,r,\dots\}$ with $\beta^{\frac{1}{2}} \leq r < 1$, we have

$$T(p,\Lambda) = pf(r)(1 + r + r^2 + \dots) = pf(r)/(1-r) .$$

It is not immediately clear that a P_3'-optimal strategy exists. We will assume for the moment the existence of a $P_3(p)$-optimal strategy for every $p > 0$. Given $p > 0$ we denote this optimal strategy by $\bar{\Lambda} = \{\bar{r}_0,\bar{r}_1,\dots\}$. Let $\bar{\Lambda}' := \{\bar{r}_1,\bar{r}_2,\dots\}$. Then in a similar way as in (3.22), or by taking limits in (3.22), we can prove

$$T(p,\bar{\Lambda}) = f(\bar{r}_0)p + T(\bar{r}_0 p,\bar{\Lambda}') . \tag{4}$$

In a similar way as in section 4 it can be proved that $\bar{\Lambda}'$ is a $P_3'(\bar{r}_0 p)$-optimal strategy. If we define

$$\bar{T}(p) := \inf\{T(p,\Lambda) \mid \Lambda \in J^\infty\} \qquad (p > 0) , \tag{5}$$

then it follows that

$$\bar{T}(p) = f(\bar{r}_0)p + \bar{T}(\bar{r}_0 p) = \min_{r\in J} [f(r)p + \bar{T}(rp)] . \tag{6}$$

Also, $\bar{T}$ is easily seen to be homogeneous. Furthermore, defining

$$S := \bar{T}(1) , \tag{7}$$

it follows from (6) that:

$$S = \mu(S) , \tag{8}$$

where μ is the function defined by (4.7). Since clearly $S > 0$ holds, we see that

$$S = s^\star = 2/\alpha \tag{9}$$

(cf. (4.12)). Also we have $\bar{r}_k = r^\star$ $(k = 0,1,\dots)$ where

$$r^\star := h(s^\star) = \frac{1 + 3\beta}{3 + \beta} = \frac{2 - \alpha}{2 + \alpha} . \tag{10}$$

This constitutes the solution of problem P_3, provided the optimal control exists. We show the existence of the optimal strategy by proving that $\bar{\Lambda} := \{r^*, r^*, \ldots\}$ actually is a $P_3(p)$-optimal strategy. We have

$$T(p,\bar{\Lambda}) = \frac{pf(r^*)}{1-r^*} = p\,\frac{f(h(s^*))}{1-h(s^*)} = p\,\frac{\mu(s^*) - s^*h(s^*)}{1-h(s^*)} = ps^* ,$$

where we have used (4.8) and the fact that $\mu(s^*) = s^*$.

Suppose that there exists a strategy $\Lambda = \{r_0, r_1, \ldots\}$ such that $T(p,\Lambda) < s^*p$. We have seen in section 4, that for the $P_2'(p,N)$-optimal strategy we have $T_N(p) = S_N p \to s^*p \quad (N \to \infty)$. Hence, there exists N_0 such that $S_{N_0}p > T(p,\Lambda)$. But

$$T(p,\Lambda) = \sum_{k=0}^{\infty} f(r_k)p_k(p,\Lambda) \geq \sum_{k=0}^{N_0-1} f(r_k)p_k(p,\Lambda^*) \geq T_{N_0}(p) ,$$

where $\Lambda^* = \{r_0, \ldots, r_{N_0-1}\} \in J^{N_0}$, so that we have a contradiction. So we have obtained the following result: The strategy $\bar{\Lambda} := \{r^*, r^*, \ldots\}$ where r^* is given by (10), is the unique $P_3'(p)$-optimal strategy for every $p > 0$. The $P_3(p)$-optimal control $\bar{u}$ is the piecewise optimal control with $p_k(p,\bar{\Lambda})$ as alteration moduli for $k = 0,1,\ldots$. Hence the $P_3(p)$-optimal control is unique also.

REMARK. The existence of a $P_3'(p)$-optimal strategy may also be proved (without using the results of P_2) by Tychonoffs theorem, according to which J^∞ is compact, with respect to the Tychonoff topology (see [11] p. 5). Unfortunately, the function $\Lambda \mapsto T(p,\Lambda)$ is not continuous (not even finite) on all of J^∞. Therefore we introduce an auxiliary system in the following way: The function $g_\delta \in (J \to R^1)$ is defined by $g_\delta(r) := \min\{r, 1-\delta\}$, where δ is some positive number. Then we define functions $p_{k,\delta}$ and $t_{k,\delta}$ in $(R^1_+ \times J^\infty \to R^1)$ by:

$$p_{0,\delta}(p,\Lambda) = p ,$$

$$p_{k+1,\delta}(p,\Lambda) = g(r_k)p_{k,\delta}(p,\Lambda) ,$$

$$t_{0,\delta}(p,\Lambda) = 0 ,$$

$$t_{k+1,\delta}(p,\Lambda) = t_k(p,\Lambda) + f(r_k)p_{k,\delta}(p,\Lambda) ,$$

where $\Lambda = \{r_0,r_1,\ldots\}$. Furthermore, we define $T_\delta \in (R^1_+ \times J^\infty \to R^1)$ by

$$T_\delta(p,\Lambda) = \lim_{k\to\infty} t_{k,\delta}(p,\Lambda) = \sum_{k=0}^{\infty} f(r_k)p_{k,\delta}(p,\Lambda) .$$

T_δ is well defined since $|p_{k,\delta}(p,\Lambda)| \le (1-\delta)^k$. In the Tychonoff topology, a function $F \in (J^\infty \to R^1)$ is continuous, if for any finite sequence $I = \{i_1,\ldots,i_\nu\}$ the mapping $\{r_{i_1},\ldots,r_{i_\nu}\} \mapsto F(\Lambda)$ is continuous for each sequence $\{r_k\}_{k=0,k\notin I}^{\infty}$. Therefore, $\Lambda \mapsto T_\delta(p,\Lambda)$ is easily seen to be continuous, so that there exists $\bar\Lambda \in J^\infty$, such that

$$T_\delta(p,\bar\Lambda) = \min\{T_\delta(p,\Lambda) \mid \Lambda \in J^\infty\} =: \bar T_\delta(p) .$$

The optimal strategy $\bar\Lambda$ can be obtained in the same way as the one for $P_3'(p)$. It follows that $\bar T_\delta(p) = S_\delta p$, where S_δ satisfies $S_\delta = \mu_\delta(S_\delta)$ and where μ_δ is given by

$$\mu_\delta(s) := \min_{r\in J} [f(r) + sg_\delta(r)] = \min\{\mu(s), 1+\beta+(1-\delta)s\} .$$

Now, if $\delta > 0$ is sufficiently small, it follows that $S_\delta = s^*$ and $\bar\Lambda = \{r^*,r^*,\ldots\}$. But we have also $T(p,\bar\Lambda) = s^*$. Therefore we have for arbitrary $\Lambda \in J^\infty$:

$$T(p,\Lambda) \ge T_\delta(p,\Lambda) \ge T_\delta(p,\bar\Lambda) = T(p,\bar\Lambda)$$

which proves the $P_3(p)$-optimality of $\bar\Lambda$.

II.6. *Switching curves*

In this section we give a description in terms of graphs of the solution found in the previous section. As is shown in section 3, for a $P_1(p,rp)$-optimal control (with $r > 1$) the control assumes the values α and $-\alpha$ consecutively, and the point where u changes sign, called the *switching point*, is given by (3.9) and (3.10). As we saw in the previous section, a $P_3(p)$-optimal trajectory consists of an infinity of pieces, each of which is a $P_1(p,r^*p)$-optimal trajectory with a translated initial time. The switching points of these pieces in the right half-plane lie on a curve which is given in parameter representation by

$$x = \frac{2}{(2+\alpha)^2}\bar{p}^2 , \quad y = -\frac{\alpha}{2+\alpha}\bar{p} \tag{1}$$

where $\bar{p}$ is the parameter (see formulas (3.9), (3.10)). The curve is also given by the equations $x = 2y^2/\alpha^2$, $y < 0$. For the pieces in the left half-plane, the switching points lie on the curve $x = -2y^2/\alpha^2$, $y > 0$. In general, the set of switching points is given by the equation:

$$x = -2\alpha^{-2}y|y| . \tag{2}$$

This set is called the *switching curve* and will be denoted by C. If we introduce the function $g \in C^1(R^2 \to R^1)$ by

$$g(\underline{x}) = x^2 + 2\alpha^{-2}xy|y| , \tag{3}$$

then it follows from the foregoing that if $\bar{u}$ is $P_3(p)$-optimal control and $\bar{\underline{x}}$ is the corresponding trajectory, we have $\bar{u}(t) =$ $= \alpha \operatorname{sgn} g(\underline{x}(t))$ a.e. On the other hand, by theorem (I.4.2) it is easily seen that the system

$$\begin{aligned} \dot{x} &= y \\ \dot{y} &= -(1 + \alpha \operatorname{sgn} g(x,y)) \operatorname{sgn} x \end{aligned} \tag{4}$$

has a unisolution for initial value $\underline{a} \in R^2$, $\underline{a} \neq \underline{0}$, and that this solution can be continued, either for all $t > 0$ or to the boundary of $S := \{\underline{x} \in R^2 \mid \underline{x} \neq \underline{0}\}$; that is, until we have $\underline{x} = \underline{0}$.

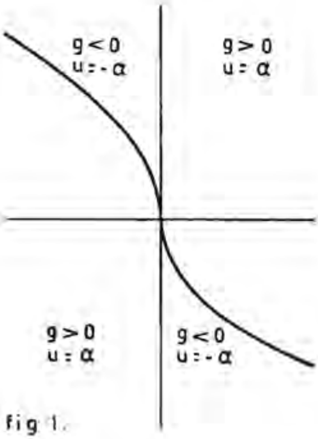


fig 1.

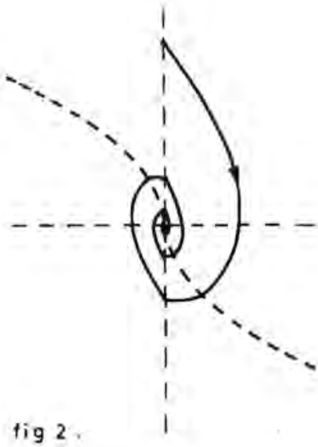


fig 2.

This implies that trajectories of (4) are P_3-optimal trajectories (see fig. 2).

An optimal control which is given as a function of the state variable is called a *feedback control*. Therefore, by the equation $u = \alpha \operatorname{sgn} g(\underline{x})$ we have given the optimal feedback.

The feedback formulation of the solution of P_3 enables us to find the optimal trajectory for a more general problem than P_3. In fact, now we can solve the problem of time optimal null-control with arbitrary value $\underline{a} = (a,b) \in R^2$ instead of the point $(0,p)$ on the y-axis. For this problem, the optimal control is also given by the feedback $u = \alpha \operatorname{sgn} g(\underline{x})$.

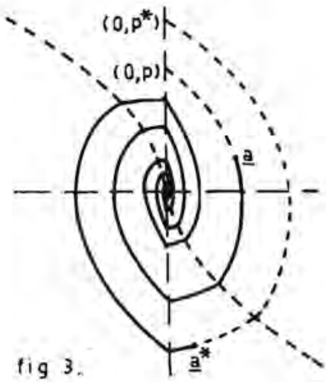


fig 3.

In order to see this, consider for definiteness a point $\underline{a} = (a,b)$ with $a > 0$. There exists exactly one real $p > 0$ such that the $P_3(p)$-optimal trajectory passes through the point $\underline{a}$ on the first step. Now the part of this trajectory, starting at $\underline{a}$ is the time optimal null-control with initial value $\underline{a}$. For, if there exists a control $\hat{u}$, which transfers $\underline{a}$ to the origin in a shorter time interval, then we can construct a trajectory which transfers $(0,p)$ to the origin in a shorter time; this is done by concatenating the part between $(0,p)$ and (a,b) of the $P_3(p)$-optimal control and the trajectory corresponding to $\hat{u}$ from $\underline{a}$ to $\underline{0}$. The same argument shows the optimality of the feedback $u = \alpha \operatorname{sgn} g(\underline{x})$ for initial values $\underline{a} = (a,b)$ with $a < 0$.

We find a surprisingly simple result for the switching time θ_k in the time interval (t_k, t_{k+1}). In fact, we have

$$\theta_k = \tfrac{1}{2}(t_k + t_{k+1}) \ . \tag{5}$$

Hence, the control assumes the values $u = \alpha$ and $u = -\alpha$ on intervals of equal length in (t_k, t_{k+1}). We can obtain formula (5) by substituting $r = r^* \ (= (2-\alpha)/(2+\alpha))$ in (3.10) and computing the first and second term of the right-hand side of (3.11) with $q = r^* p$. Now we describe an independent method for finding the $P_3(p)$-optimal control (in particular formula (5)), without the results of sections 3, 4, 5 and even without the maximum principle. Admittedly, we assume in this derivation that the existence of a $P_3(p)$-optimal control has already been proved, but, if we wish to do so, we can establish the existence of an optimal control independently by means of the solution of the problem P_3^* below. Then it is a consequence of Theorem 2.1 that the optimal final time $T(p)$ depends homogeneously on p, so that we can write $T(p) = Sp$ for some $S > 0$. Let $\underline{x} = (x,y)$ be an optimal trajectory and let t_1 be the smallest value of $t > 0$ for which we have $y(t_1) = 0$. Then we have $T(|y(0)|) = t_1 + T(|y(t_1)|)$ and hence $S = t_1/(|y(0)| - |y(t_1)|)$. Because of the homogeneity of the problem, it is no loss of generality to assume that $y(0) = 1$. Then we have $y(t_1) < 0$ and hence

$$S = t_1/(1 + y(t_1)) = 1/\tan \phi \tag{6}$$

(see fig. 4, where $A = (0,1)$, $B = (t_1,1)$, $A' = (0,-1)$, $B' = (t_1,-1)$, $O = (0,0)$, $P = (t_1,0)$, $C = (0,y(t_1))$, $D = (t_1,y(t_1))$).

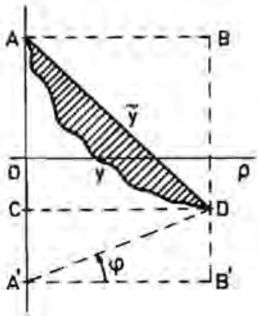


fig 4.

Furthermore, it follows from $x(0) = x(t_1) = 0$ that

$$\int_0^{t_1} y(t)dt = 0 . \tag{7}$$

Also (according to (2.1)), the function y satisfies $|\dot{y}+1| \le \alpha$. We will consider the somewhat more general problem, where y has to satisfy $-\rho \le \dot{y} \le -\sigma$, and σ, ρ are arbitrary numbers satisfying $\rho > \sigma \ge 0$. We show that, if (7) is satisfied, the area of the triangle A'B'D equals the area of the shaded figure (in fig. 4), that is,

$$\int_0^{t_1} (\tilde{y}(t) - y(t))dt ,$$

where $\tilde{y}$ is the function of which the line segment AD is the graph. Denoting the area of a polygon XYZ... by < XYZ... > we have

$$\int_0^{t_1} (\bar{y}(t) - y(t))dt = \int_0^{t_1} \bar{y}(t)dt = < ADC > - < OPCD > =$$

$$= \tfrac{1}{2} < ABCD > - < OPCD > = \tfrac{1}{2}(< ABPO > - < OPCD >) = < A'B'D > .$$

Since $\tan \phi = < A'B'D >/(OP)^2$ we have reduced P_3 to the following problem:

P_3^* : "Find a number t_1 and an absolutely continuous function $y \in ([0,t_1] \to R^1)$, satisfying (7), $y(0) = 1$, and $-\rho \le \dot{y} \le -\sigma$ (a.e.), such that

$$I(y,t_1) := t_1^{-2} \int_0^{t_1} (\bar{y}(t) - y(t))dt$$

is maximal, where $\bar{y}(t) := 1 + (y(t_1) - 1)t/t_1$."

Let y be a function in some interval $[0,t_1]$ satisfying $y(0) = 1$, $-\rho \le \dot{y} \le -\sigma$ (a.e.), and $\int_0^{t_1} y(t)dt < t_1$. We define $y^\lambda \in ([0,\lambda t_1] \to R^1)$ by

$$y^\lambda(t) := 1 + \lambda(y(t/\lambda) - 1) , \tag{8}$$

for $\lambda > 0$. Then we have $y^\lambda(0) = 1$, $-\rho \le \dot{y}^\lambda(t) \le -\sigma$ (a.e.). Note that the graph of y^λ is obtained by multiplying the graph of y by λ with respect to (0,1). There exists $\lambda > 0$ such that $\int_0^{\lambda t_1} y^\lambda(t)dt = 0$. In fact, it follows from

$$J_\lambda := \int_0^{\lambda t_1} y^\lambda(t)dt = \lambda^2\left[(\lambda^{-1} - 1)t_1 + \int_0^{t_1} y(t)dt\right] \tag{9}$$

that $J_\lambda \sim \lambda^2(J_1 - t_1) < 0$ $(\lambda \to \infty)$ and $J_\lambda \sim \lambda t_1 > 0$ $(\lambda \to 0)$. Also, it follows from (9) that $I(y^\lambda,\lambda t_1) = I(y,t_1)$. This implies that we can restate P_3^* as follows:

P_3^{**} : "Find an absolutely continuous function $y \in ([0,1] \to \mathbb{R}^1)$, satisfying:

$$y(0) = 1, \int_0^1 y(t)dt < 1, -\rho \le \dot{y}(t) \le -\sigma \text{ (a.e.) }, \tag{10}$$

such that

$$J(y) := \int_0^1 (\tilde{y}(t) - y(t))dt$$

is maximal, where $\tilde{y}(t) := 1 + t(y(1) - 1) \quad (0 \le t \le 1)$."

We will call functions y, satisfying (10) *admissible*. Let y be admissible. Because of the restriction $-\rho \le \dot{y} \le -\sigma$ we have $y(t) \ge 1-\rho t$ and $y(t) \ge y(1) + \sigma(1-t)$. Hence

$$y(t) \ge y^*(t) := \max\{1-\rho t, y(1) + \sigma(1-t)\} .$$

But it is easily seen that also y^* is admissible and that $J(y^*) > J(y)$.

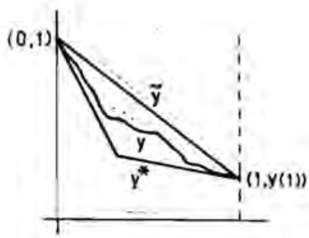


fig 5:

It follows that if y is optimal, then there exists $\theta \in [0,1]$ such that y is linear on each of the intervals $[0,\theta]$, $[\theta,1]$ and such that $\dot{y} = -\rho \ (0 < t < \theta)$, $\dot{y} = -\sigma \ (\theta < t < 1)$. Hence y is of the form

$$y(t) = 1 - \rho t \qquad (0 < t < \theta) ,$$
$$y(t) = 1 - \rho\theta + \tfrac{1}{2}(\rho - \sigma)\theta^2 \qquad (\theta < t < 1) .$$

A short calculation yields that in this case we have

$$J(y) = \tfrac{1}{2}(\rho - \sigma)\theta(1 - \theta)$$

so that $J(y)$ is maximal for $\theta = \frac{1}{2}$. This formula is equivalent to (5).

REMARK 1. Mechanically the switching point is determined by a fixed ratio of the potential and kinetic energy of the system. Apart from constant factors, the potential energy is given by $E_p = |x|$ and kinetic energy by $E_k = y^2$. Hence for the switching point we have $E_p = 2\alpha^{-2}E_k$. Furthermore, E_p decreases (considered as function of t) at the switching time (i.e. $dE_p/dt < 0$).

We are going to give now an analogous description in terms of graphs of the P_2-optimal trajectories. Consider the set of $P_1(p,rp)$-trajectories for fixed $r \in J$, $r \leq 1$ and varying $p > 0$. The switching points of these trajectories are given by (3.9) and (3.10) and form together a semi-parabola: $x = \theta(r)y^2$, $y < 0$, where

$$\theta(r) = \frac{1 + \beta}{4}\,\frac{1 - r^2}{r^2 - \beta} \quad , \tag{11}$$

provided $r > \beta^{\frac{1}{2}}$. If $r = \beta^{\frac{1}{2}}$ the switching points are given by the equation $y = 0$. If $r = 1$, the y-axis is the switching curve. We will denote this curve together with its mirror image with respect to the origin by $\Gamma(r)$. The $P_2(p,N)$-optimal trajectory is piecewise optimal corresponding to the strategy $\bar{\Lambda} = \{\bar{r}_0,\ldots,\bar{r}_{N-1}\}$ given in (4.13). The switching curve for the piece between the k-th and the (k+1)-st intersection point is $\Gamma(\bar{r}_k)$. As a result we have a set of N switching curves $\Gamma(\bar{r}_0),\ldots,\Gamma(\bar{r}_{N-1})$. We denote the curve $\Gamma(\bar{r}_{N-k})$ by Γ_k, so that

$$\Gamma_k = \Gamma(h(S_{k-1})) \qquad (k = 1,2,\ldots)\ .$$

We have $\Gamma_k \to \Gamma(r^\star)$ $(k \to \infty)$ (in an appropriate sense) and Γ_1 is the y-axis.

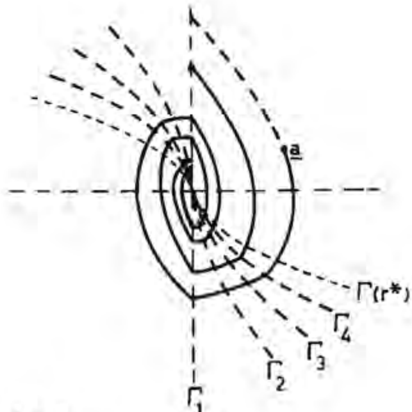


fig 5.

At each piece of the P_2-optimal trajectory we have to choose one of the curves as a switching curve. In fact, if the trajectory still has to intersect the y-axis ℓ times (irrespective the numbers of intersections before), then we have to choose Γ_ℓ as the switching curve. Exactly in the same way as for time optimal null control, in this case we can find optimal trajectories for the generalized version of P_2, where the initial value is arbitrary (instead of on the positive y-axis).

II.7. *Maximal average acceleration*

$P_4(p,N)$: "Given a natural number N and a positive real number p, find the p-start, N-step control which maximizes the ratio p_N/T where p_N is the last alteration modulus and T is the final time."

Just as in the problems P_2 and P_3 it is easy to show here that we can restrict our attention to piecewise optimal controls. Therefore, equivalent to P_4 is the following problem:

$P_4'(p,N)$: "Given $p > 0$, N a natural number, find an N-step strategy $\Lambda = \{r_0,\dots,r_{N-1}\}$ such that $V_N(p,\Lambda) := t_N(p,\Lambda)/p_N(p,\Lambda)$ is minimal."

Just as in problem P_2', the existence of an optimal strategy is obvious. It follows from Remark 3.5 that $V_0(p,\emptyset) = 0 \quad (p > 0)$.

Let for $N > 1$ and an arbitrary $\Lambda = \{r_0,\ldots,r_{N-1}\} \in J^N$, the $(N-1)$-step strategy $\tilde{\Lambda}$ be defined by $\tilde{\Lambda} = \{r_0,\ldots,r_{N-2}\}$ and for $N = 1$, $\tilde{\Lambda} = \emptyset$. Then we have for $N \geq 1$:

$$V_N(p,\Lambda) = \frac{t_{N-1}(p,\tilde{\Lambda}) + f(r_{N-1})p_{N-1}(p,\tilde{\Lambda})}{r_{N-1}p_{N-1}(p,\tilde{\Lambda})} =$$

$$= \frac{1}{r_{N-1}} [V_{N-1}(p,\tilde{\Lambda}) + f(r_{N-1})] \ . \tag{1}$$

It follows from this formula, that if $\bar{\Lambda}$ is $P_4'(p,N)$-optimal, $\tilde{\bar{\Lambda}}$ is $P_4'(p,N-1)$-optimal. If $\bar{V}_N$ is defined by

$$\bar{V}_N := \min_{\Lambda \in J^N} \{V_N(p,\Lambda)\} \ , \tag{2}$$

(so that $\bar{V}_0 = 0$), we infer from (1) that

$$\bar{V}_N = \min_{r \in J} [(\bar{V}_{N-1} + f(r))/r] \ ,$$

whereas r_{N-1} is a value of r for which the minimum in (2) is assumed. But since $1/r \in J$ if and only if $r \in J$, it follows by the substitution $\rho = 1/r$, that

$$\bar{V}_N = \min_{\rho \in J} [\rho(\bar{V}_{N-1} + f(1/\rho))] = \min_{\rho \in J} [\rho\bar{V}_{N-1} + f(\rho)] = \mu(\bar{V}_{N-1}) \ ,$$

with μ as defined by (4.7). The value of ρ for which the minimum is assumed is $h(\bar{V}_{N-1})$, and hence $\bar{r}_{N-1} = 1/h(\bar{V}_N)$. Since $\tilde{\bar{\Lambda}}$ is $P_4(p,N-1)$-optimal, we have $\bar{r}_{N-2} = 1/h(\bar{V}_{N-2})$. It follows by induction that we have

$$\bar{V}_k = \mu(\bar{V}_{k-1}) \quad (k = 1,\ldots,N) \ , \quad \bar{V}_0 = 0 \ , \tag{3}$$

and the $P_4'(p,N)$-optimal strategy $\bar{\Lambda} = \{\bar{r}_0,\ldots,\bar{r}_{N-1}\}$ is given by

$$\bar{r}_k = 1/h(\bar{V}_k) \qquad (k = 0,\ldots,N-1) \; . \tag{4}$$

In particular, we see that the optimal strategy (and hence the $P_4(p,N)$-optimal control) is unique.

REMARK. It is somewhat surprising that $\bar{r}_0 = 1$, but in the case $N = 1$ it is easily seen to be correct from fig. 3.5. Note also that V_N does not depend on p_0.

It follows from (3) that $\bar{V}_N = S_N$, in particular, we have $\bar{V}_N \to s^*$ $(N \to \infty)$ where the convergence is monotonic.

If we use the N-step strategy

$$\Lambda^* = \{1/r^*, 1/r^*, \ldots\} \; , \tag{5}$$

with r^* as defined in (5.10), we have $V_k(p,\Lambda^*) = s^*$ for $k = 1,\ldots,N$. Although this strategy is not optimal we see that $V_N(p,\Lambda^*) - \bar{V}_N \to 0$ $(N \to \infty)$. Therefore we will call this strategy "quasi-optimal". The advantage of the quasi-optimal strategy is that the corresponding piecewise optimal control is simple to describe. We shall do this in a graphical way just as in the previous section. The switching curve corresponding to (5) is given by (3.12) with $r = 1/r^*$ and where p is a parameter. Eliminating p we obtain

$$x = 2y|y|/\alpha \tag{6}$$

for the switching curve (see fig. 1). For the actual optimal control we have a number of switching curves similar to the ones of the P_2-optimal controls. In this case we get for the set of switching curves:

$$x = \theta(\bar{r}_k)y|y| \qquad (k = 0,1,\ldots) \; , \tag{7}$$

where θ is defined by

$$\theta(r) = \frac{1+\beta}{4}\,\frac{r^2 - 1}{1 - \beta r^2} \qquad (1 \le r < \beta^{-\frac{1}{2}})$$

and $\bar{r}_k$ by (4) (see fig. 2).

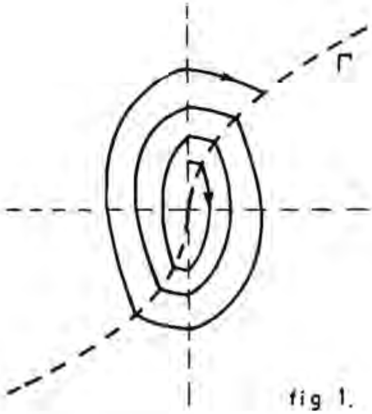


fig 1.

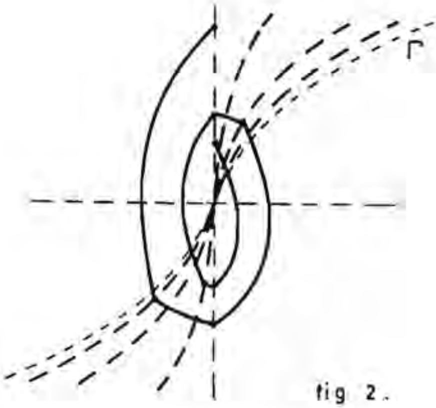


fig 2.

REMARK 1. Unlike the situation of section 5 we cannot use this feedback control for optimal trajectories starting at an arbitrary point of the plane.

REMARK 2. We are now in the position to give the example (promised in section 2) of a situation where the trajectories are not unique. For that aim we use the feedback control indicated in fig. 1; that is, $u = \alpha \operatorname{sgn} g(\underline{x})$, where $g(\underline{x}) = x^2 - 2y|y|/\alpha$. If we reverse the time in the system

$$\dot{x} = y,\ \dot{y} = -(1 + \alpha \operatorname{sgn} g(\underline{x}))x\ , \tag{8}$$

then the trajectory attains the origin in a finite time interval T. Therefore, the solution of (8) starting at $\underline{0}$ are not unisolutions.

II.8. *Time optimal* p-*start,* N-*step control with fixed endpoint*

$P_5(p,q,N)$: "Given real positive numbers p and q and a natural number N, determine the p-start, N-step control with final alteration modulus $p_N = q$ and minimal final time t_N."

Again, we can confine ourselves to piecewise optimal controls. Hence equivalent to P_5 we have:

$P_5'(p,q,N)$: "Given real numbers $p,q > 0$ and a natural number N, find the N-step strategy Λ such that $p_N(p,\Lambda) = q$ and $t_N(p,\Lambda)$ is minimal, where the functions p_N and t_N are defined in (3.19) and (3.20)."

In order that there exists $\Lambda \in J^N$ for which $p_N(p,\Lambda) = q$ holds, it is necessary and sufficient that $q/p \in J_N$, where

$$J_N := [\beta^{N/2}, \beta^{-N/2}] \ . \tag{1}$$

Since J^N is compact and $\Lambda \mapsto p_N(p,\Lambda)$, $\Lambda \mapsto t_N(p,\Lambda)$ are continuous, it is obvious that there exists an optimal strategy whenever the condition $q/p \in J_N$ is satisfied. We will denote the $P_5(p,q,N)$-optimal time by $T_N(p,q)$. Hence

$$T_N(p,q) := \min\{t_N(p,\Lambda) \mid \Lambda \in J^N \wedge p_N(\Lambda) = q\} \ . \tag{2}$$

Here the minimum of an empty set is defined to be infinity. Hence $T_N(p,q) = \infty$ if $q/p \notin J_N$. Let $\bar{\Lambda} = \{\bar{r}_0,\dots,\bar{r}_{N-1}\}$ be a $P_5'(p,q,N)$-optimal strategy. By (3.22) we have

$$t_N(p,\bar{\Lambda}) = f(\bar{r}_0)p + t_{N-1}(\bar{r}_0 p,\bar{\Lambda}') \ . \tag{3}$$

Since $p_N(p,\bar{\Lambda}) = q$, it follows that $p_{N-1}(\bar{r}_0 p,\bar{\Lambda}') = q$. Therefore, it is easily seen that $\bar{\Lambda}'$ is $P_5'(\bar{r}_0 p,q,N-1)$-optimal. It follows then that $T_N(p,q) = f(\bar{r}_0)p + T_N(\bar{r}_0 p,q)$ and because of the optimality of $\bar{\Lambda}$ this yields the equation

$$T_N(p,q) = \min_{r\in J} [f(r)p + T_{N-1}(rp,q)] \ . \tag{4}$$

Properly speaking the minimum should be over those $r \in J$ for which $q/(rp) \in J_{N-1}$ holds. However, as $T_N(p,q) = \infty$ $(q/p \notin J_N)$, we can use the simpler formulation (4). Now it is easily seen that for $\rho > 0$

we have $T_N(\rho p,\rho q) = \rho T_N(p,q)$; that is, T_N is homogeneous. Therefore, if we set

$$S_N(q) := T_N(1,q) \ . \tag{5}$$

we obtain the following recurrence relation for $S_N(q)$:

$$\left.\begin{aligned} S_N(q) &= \min_{r\in J} [f(r) + rS_{N-1}(q/r)] \ , \\ S_1(q) &= f(q) \qquad (q \in J) \ , \\ S_1(q) &= \infty \qquad (q \notin J) \ . \end{aligned}\right\} \tag{6}$$

From this recurrence relation the functions $S_N(q)$ can be calculated successively for $N = 1,2,\ldots$. However, these calculations cannot be performed analytically, but only by a numerical method. We shall not go into this subject here, since we will obtain a much more tractable result in the next section.

We conclude this section with some simple properties of S_N. At first, it is easily shown by induction that S_N is continuous. Furthermore, we have

$$S_N(1/q) = q^{-1}S_N(q) \ . \tag{7}$$

This is a consequence of the relation $T_N(p,q) = T_N(q,p)$. The latter equality is an easy consequence of Theorem 2.1. By means of (7) we can write instead of the first equation of (6):

$$S_N(q) = \min_{r\in J} [f(r) + qS_{N-1}(r/q)] \ . \tag{8}$$

Furthermore, it follows from section 4 that $\min_{q\in J_N} S_N(q) = S_N$ (defined by (4.6)) and that the minimum is assumed for $q = p_N(1,\bar{\Lambda})$, where $\bar{\Lambda} := \{\bar{r}_0,\ldots,\bar{r}_{N-1}\}$ and $\bar{r}_k$ $(k = 0,\ldots,N-1)$ is defined in (4.13).

More properties of S_N will be obtained in section 9. It will be shown there that S_N is continuously differentiable in the interior of J_N for $N \geq 2$, S_N is strictly convex, and that the graph of S_N has vertical slope at the endpoints of J_N. Furthermore, it will be shown that $S_N(q) \to s^*(1 + q)$ $(N \to \infty)$ for every $q > 0$, where s^* is given by (4.12).

II.9. *An auxiliary optimization problem*

$P_6(p,\rho,N)$: "Given $p > 0$, ρ real and N natural, find a p-start, N-step control for which $\theta_N := t_N + \rho p_N$ is minimal, where t_N is the final time and p_N the final alteration modulus."

Note that $P_6(p,0,N) = P_2(p,N)$. Again we have to consider only piecewise optimal control. Hence:

$P_6'(p,\rho,N)$: "Given $p > 0$, ρ real, N natural, find a N-step strategy Λ which minimizes $\theta_N(p,\rho,\Lambda) := t_N(p,\Lambda) + \rho p_N(p,\Lambda)$ ".

The existence of an optimal strategy is clear. From (3.22) it follows that for $\Lambda = \{r_0,\ldots,r_{N-1}\} \in J^N$ we have:

$$\theta_N(p,\rho,\Lambda) = f(r_0)p + \theta_{N-1}(r_0p,\rho,\Lambda') . \tag{1}$$

If we introduce $\Theta_N \in (R^1_+ \times R^1 \to R^1)$ by

$$\Theta_N(p,\rho) := \min_{\Lambda\in J^N} \theta_N(p,\rho,\Lambda) , \tag{2}$$

then the relation

$$\Theta_N(p,\rho) = \min_{r\in J} [f(r)p + \Theta_{N-1}(rp,\rho)] \tag{3}$$

can be derived from (1) in a way analogous to the way (4.4) is derived from (4.1). Now it is easily seen that $p \mapsto \Theta_N(p,\rho)$ is homogeneous of degree 1. Hence, if we define $\Phi_N \in (R^1 \to R^1)$ by

$$\Phi_N(\rho) := \Theta_N(1,\rho) , \tag{4}$$

we have $\Theta_N(p,\rho) = p\Phi_N(\rho)$. It follows then that

$$\Phi_N(\rho) = \mu(\Phi_{N-1}(\rho)) \ ,$$
$$\Phi_0(\rho) = \rho \ , \qquad (5)$$

with μ as defined in (4.7). By this recurrence relation $\Phi_n(\rho)$ can be calculated successively. Furthermore, $\bar{r}_k$ can be found from the equalities

$$\bar{r}_k = \omega_{N-1-k}(\rho) \qquad (k = 0,\dots,N-1) \qquad (6)$$

where the functions $\omega_k \in (R^1 \to J)$ are defined by

$$\omega_k(\rho) = h(\Phi_k(\rho)) \qquad (k = 0,\dots,N-1) \qquad (7)$$

and h is defined in (4.9).

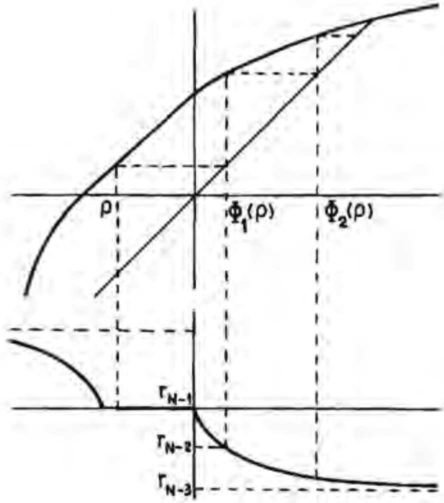


fig 1.

We see that the $P_6'(p,\rho,N)$-optimal strategy (and hence the $P_6(p,\rho,N)$-optimal control) is unique. The optimal strategy as well as Θ_N depends continuously on ρ. Furthermore, Φ_k is strictly increasing, ω_k is non-increasing and ω_k is strictly decreasing except at the points ρ for which $\Phi_k(\rho) \in [-1-\beta,0]$. We will denote $p_N(p,\bar{\Lambda})$ by $\pi_N(\rho)p$. Hence

$$\pi_N(\rho) = p_N(1,\bar{\Lambda}) \ . \tag{7a}$$

Then π_N is continuous. Furthermore, since $\pi_N(\rho) = \omega_{N-1}(\rho) \ \dots \ \omega_0(\rho)$ and since $-1-\beta \le \Phi_k(\rho) \le 0$ holds for at most one k, it follows that π_N is strictly decreasing if $N \ge 2$.

REMARK 1. Note that, since $\pi_1(\rho) = \omega_0(\rho) = h(\rho)$, the function π_1 is not strictly decreasing (see fig. 4.2).

Finally it is obvious that we have

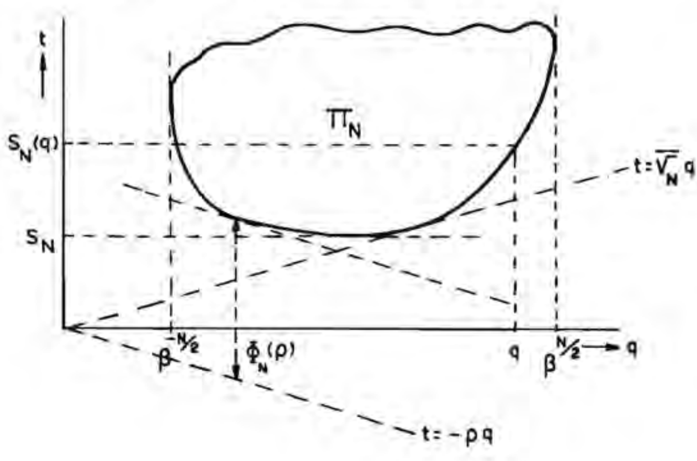
$$\Phi_k(\rho) \to \infty \quad (\rho \to \infty) \ , \quad \Phi_k(\rho) \to -\infty \quad (\rho \to -\infty) \ , \tag{8a}$$

$$\omega_k(\rho) \to \beta^{\frac{1}{2}} \quad (\rho \to \infty) \ , \quad \omega_k(\rho) \to \beta^{-\frac{1}{2}} \quad (\rho \to -\infty) \ , \tag{8b}$$

$$\pi_N(\rho) \to \beta^{N/2} \ (\rho \to \infty) \ , \quad \pi_N(\rho) \to \beta^{-N/2} \ (\rho \to -\infty) \ . \tag{8c}$$

Problem P_6 is related to problems P_2, P_4, P_5. We will give a geometric interpretation of these problems which will make this relationship clearer. Because of homogeneity it is no loss of generality to assume that $p = 1$. We introduce the *attainable set*:

$$\Pi_N := \{p_N(1,\Lambda),t_N(1,\Lambda) \mid \Lambda \in J^N\} \ . \tag{9}$$



We have the following relations:

$$J_N = \{q \mid \exists_t \ (q,t) \in \Pi_N\} \ , \tag{10}$$

$$S_N = \min\{t \mid \exists_q \ (q,t) \in \Pi_N\} \ , \tag{11}$$

$$S_N(q) = \min\{t \mid (q,t) \in \Pi_N\} \ , \tag{12}$$

$$\Phi_N(\rho) = \min\{t + \rho q \mid (q,t) \in \Pi_N\} \ . \tag{13}$$

Here S_N is defined in (4.1), $S_N(q)$ in (8.2) and J_N by (8.1).

Now we can solve the problems P_2, P_4, P_5 using the solutions of P_6.

I. As we have observed already, $P_2(p,N)$ is equivalent to $P_6(p,0,N)$.

II. We shall show that $-\bar{V}_N$ (defined in (7.2)) is the unique value of ρ for which $\Phi_N(\rho) = 0$ holds, and that the $P_4'(1,N)$-optimal strategy is the $P_6(1,\rho,N)$-optimal strategy with this value of ρ. The existence and uniqueness of the zero of Φ_N is a consequence of (8a) and the monotonicity of Φ_N. Let us denote the zero of Φ_N by $\bar{\rho}$ and let $(\bar{q},\bar{t}) \in \Pi_N$ be a point for which $\Phi_N(\bar{\rho}) = \bar{t} + \bar{\rho}\bar{q}$ (see formula (13)). Then for all $(q,t) \in \Pi_N$ we have $\bar{t} + \bar{\rho}\bar{q} \leq t + \bar{\rho}q$. Since $\bar{t} + \bar{\rho}\bar{q} = 0$ it follows that $\bar{\rho} = -\bar{t}/\bar{q}$, and $t + \bar{\rho}q \geq 0$ for all $(t,q) \in \Pi_N$. Hence, $t/q \geq -\bar{\rho} = \bar{t}/\bar{q}$ $((t,q) \in \Pi_N)$, and this implies $\bar{V}_N = -\bar{\rho}$. Therefore, we have the following result for P_4: "The P_4-optimal strategy $\bar{\Lambda} = \{\bar{r}_0,\ldots,\bar{r}_{N-1}\}$ satisfies the equations $\bar{r}_k = h(\phi_{N-k-1})$, where ϕ_k is determined by $\phi_N = 0$, $\phi_{k+1} = \mu(\phi_k)$ $(k = N-1,\ldots,0)$." Note that ϕ_k is uniquely determined by these relations, since μ is strictly increasing.

REMARK 2. An argument, similar to the foregoing one, is used in [12].

Now we will prove that the result obtained here coincides with the one given in section 7. For that aim we express ϕ_k in terms of ϕ_{k+1} by means of the formula $\mu(-\mu(-s)) = s$ (see (4.16)). Then it follows that $\phi_k = -\mu(-\phi_{k+1})$ $(k = 0,1,\ldots,N-1)$. Furthermore, introducing the notation $v_k := -\phi_{N-k}$, we obtain $v_0 = 0$, $v_{N-k} = \mu(v_{N-k-1})$

$(k = N-1,\ldots,0)$, $\bar{r}_k = h(-v_{k+1})$. Finally, using formula (4.17) we have: $v_0 = 0$, $v_{k+1} = \mu(v_k)$ $(k = 0,\ldots,N-1)$, $\bar{r}_k = 1/h(v_k)$, and this result is equivalent to (7.3) and (7.4).

III. Consider problem $P_5(1,q,N)$ for $N \geq 2$. (Note that $P_5(1,q,1)$ is equivalent to $P_1(1,q)$.) If $q = \beta^{N/2}$ or $q = \beta^{-N/2}$ there exists only one N-step strategy Λ for which $p_N(1,\Lambda) = q$ holds. This is necessarily the optimal strategy. We assume henceforth that $\beta^{N/2} < q < \beta^{-N/2}$. Then we will show that there exists a unique $P_5'(1,q,N)$-optimal strategy and that this strategy is $P_6'(1,\rho,N)$-optimal, where ρ is uniquely determined by the equation

$$\pi_N(\rho) = q \ . \tag{14}$$

The existence of a solution of (14) is a consequence of (8c) and the continuity of π_N. The solution is unique since π_N is strictly decreasing. We denote the solution ρ of (14) by $R_N(q)$. Hence:

$$R_N(q) = \rho \iff \pi_N(\rho) = q \quad (\beta^{N/2} < q < \beta^{-N/2}) \ . \tag{15}$$

It follows that R_N is continuous and strictly decreasing. For a given q, let the quantity $\bar{\rho}$ be defined by $\bar{\rho} = R_N(q)$ and let $\bar{\Lambda}$ denote the $P_6'(1,\bar{\rho},N)$-optimal strategy. Then $p_N(1,\bar{\Lambda}) = \pi_N(\bar{\rho}) = q$ (see (7a)). Also, if $\Lambda \neq \bar{\Lambda}$ is an N-step strategy with $p_N(1,\Lambda) = q$, we have

$$t_N(1,\bar{\Lambda}) = \theta_N(1,\bar{\rho},\bar{\Lambda}) - \bar{\rho}p_N(1,\bar{\Lambda}) = \Phi_N(\bar{\rho}) - \bar{\rho}q <$$

$$< \theta_N(1,\bar{\rho},\Lambda) - \bar{\rho}q = t_N(p,\Lambda) \ .$$

This proves that $\bar{\Lambda}$ is the unique $P_5'(1,q,N)$-optimal strategy.

REMARK 3. It follows from the foregoing that we have the following formula

$$\Phi_N(R_N(q)) = S_N(q) + qR_N(q) \ . \tag{15a}$$

In order to find the $P_5'(1,q,N)$-optimal strategy we have to solve equation (14). It seems impossible to find an explicit expression for the solution. However, since π_N decreases, the equation is easily solved by a numerical method (e.g. regula falsi).

Let us now derive some properties of $S_N(q)$ for $N \geq 2$. (Note that $S_1(q) = f(q)$.) Let

$$q_1, q_2 \in \operatorname{int}(J_N) := \{q \mid \beta^{N/2} < q < \beta^{-N/2}\} .$$

Since $(q_1, S_N(q_1)) \in \Pi_N$, we have by (13): $S_N(q_1) + \rho q_1 \geq \Phi_N(\rho)$ for every real ρ. Specializing to $\rho = R_N(q_2)$ and using (15a) for $q = q_2$ we obtain

$$S_N(q_1) \geq \Phi_N(R_N(q_2)) - q_2 R_N(q_2) + (q_2 - q_1)R_N(q_2) =$$

$$= S_N(q_2) + (q_2 - q_1)R_N(q_2) .$$

Hence

$$S_N(q_1) - S_N(q_2) \geq -R_N(q_2)(q_1 - q_2) .$$

Interchanging the role of q_1 and q_2 and combining the two results we obtain under the assumption $q_2 > q_1$:

$$- R_N(q_1) \leq \frac{S_N(q_2) - S_N(q_1)}{q_2 - q_1} \leq -R_N(q_2) .$$

By the continuity of R_N this implies

$$S_N'(q) = - R_N(q) . \tag{16}$$

As a consequence we see that S_N is strictly convex.

REMARK 4. It is surprising that S_N is continuously differentiable for $N \geq 2$, whereas $S_1 = f$ is not differentiable at $q = 1$. This circumstance is due to the fact that π_N is strictly decreasing for $N \geq 2$ but not strictly decreasing if $N = 1$ (see Remark 1).

Now we will derive some properties of Φ_N. Since $\mu \in C^1$ it follows from (5) that $\Phi_k \in C^1$ $(k = 0,\dots,N)$ and from $\mu'(s) = h(s)$ (formula (4.15)) we infer

$$\Phi_0'(\rho) = 1 \ , \quad \Phi_{k+1}'(\rho) = \omega_k(\rho)\Phi_k'(\rho) \ . \tag{17}$$

Especially it follows that

$$\Phi_N'(\rho) = \pi_N(\rho) \ . \tag{18}$$

Since for $N \geq 2$ the function π_N is strictly decreasing we conclude that Φ_N is strictly concave. Furthermore, the following relation can be derived from (12) and (13):

$$\Phi_N(\rho) = \min_{q \in J_N} \{S_N(q) + \rho q\} \ . \tag{19}$$

In fact, if we define $\tilde{\Phi}_N(\rho) := \min_{q \in J_N} \{S_N(q) + \rho q\}$, we have

$$\Phi_N(\rho) = \min \{t + \rho q \mid (q,t) \in \Pi_N\} \leq S_N(q) + \rho q$$

for all $q \in J_N$, since $(q, S_N(q)) \in \Pi_N$. Hence, $\Phi_N(\rho) \leq \tilde{\Phi}_N(\rho)$. On the other hand, if $(\bar{q},\bar{t}) \in \Pi_N$ satisfies $\bar{t} + \rho\bar{q} = \Phi_N(\rho)$, we have

$$\Phi_N(\rho) = \bar{t} + \rho\bar{q} \geq S_N(\bar{q}) + \rho\bar{q} \geq \tilde{\Phi}_N(\rho) \ .$$

This proves (19).

Since S_N is continuous and strictly convex, and since $S_N(q) = qS_N(1/q)$ (see formula (8.7)), we can use Remark (4.2) to obtain

$$\Phi_N(-\Phi_N(-\rho)) = \rho \ , \tag{20}$$

(that is, the graph of Φ_N is symmetric with respect to the line $\Phi = -\rho$), and furthermore

$$\pi_N(-\Phi_N(\rho)) = 1/\pi_N(\rho) \ . \tag{21}$$

REMARK 5. Equation (20) can also be derived by induction from (4.16). In fact, $\Phi_1 = \mu$ and $\Phi_N(\rho) = \mu(\Phi_{N-1}(\rho)) = \Phi_{N-1}(\mu(\rho))$. Hence (4.16) is equivalent to (20) for $N = 1$. Assuming that (20) is proved with $N-1$ instead of N, we have

$$\Phi_N(-\Phi_N(-\rho)) = \mu(\Phi_{N-1}(-\Phi_{N-1}(\mu(-\rho))) = \mu(-\mu(-\rho)) = \rho \ .$$

We now derive some asymptotic formulas for the functions discussed in this section.

A. Suppose that N is fixed. Then we derive asymptotic formulas for $\Phi_N(\rho)$, $\omega_N(\rho)$, $\pi_N(\rho)$ for $\rho \to \pm \infty$ (a first estimate is given in (8)). In (4.11) we have seen that $\mu(s) = \beta^{\frac{1}{2}}s + a_1 + O(s^{-1})$ $(s \to \infty)$, where

$$a_1 := \tfrac{1}{2}(1 + \beta)(1 + \beta^{-\frac{1}{2}}) = f(\beta^{\frac{1}{2}}) .$$

It is easily shown by induction that this implies

$$\Phi_N(\rho) = \beta^{N/2}\rho + a_N + O(\rho^{-1}) \qquad (\rho \to \infty) , \tag{22}$$

(not uniformly with respect to N) for N = 1,2,..., where

$$a_N := \frac{1 - \beta^{N/2}}{1 - \beta^{\frac{1}{2}}} a_1 = S_N(\beta^{N/2}) , \tag{23}$$

and similarly

$$\Phi_N(\rho) = \beta^{-N/2}\rho + b_N + O(\rho^{-1}) \quad (\rho \to -\infty) , \tag{24}$$

where

$$b_N := \frac{\beta^{-N/2} - 1}{\beta^{-\frac{1}{2}} - 1} b_1 = S_N(\beta^{-N/2}) , \tag{25}$$

and

$$b_1 := (2\beta)^{-1}(1 + \beta)(1 + \beta^{\frac{1}{2}}) = f(\beta^{-\frac{1}{2}}) . \tag{26}$$

The asymptotic formulas for $\omega_k(\rho)$ follow from remark (4.1): We have

$$h(s) = \beta^{\frac{1}{2}} + cs^{-2} + O(s^{-3}) \qquad (s \to \infty) ,$$

where

$$c := \frac{1}{8} (1 - \beta)(1 + \beta)^2\beta^{-3/2} . \tag{27}$$

Therefore the following formulas obtain for $\omega_k(\rho)$:

$$\omega_k(\rho) = \beta^{\frac{1}{2}} + c\beta^{-k}\rho^{-2} + O(\rho^{-3}) \qquad (\rho \to \infty) , \tag{28}$$

and similarly

$$\omega_k(\rho) = \beta^{-\frac{1}{2}} - c\beta^{k}\rho^{-2} + O(\rho^{-3}) \qquad (\rho \to -\infty) . \tag{29}$$

Finally, from $\pi_N(\rho) = \omega_0(\rho) \cdots \omega_{N-1}(\rho)$ it follows that

$$\pi_N(\rho) = \beta^{N/2} + d_N\rho^{-2} + O(\rho^{-3}) \qquad (\rho \to \infty) , \tag{30}$$

and

$$\pi_N(\rho) = \beta^{-N/2} - d_N\rho^{-2} + O(\rho^{-3}) \qquad (\rho \to \infty) , \tag{31}$$

where

$$d_N := \frac{1}{8}\,(1+\beta)^2\beta^{-1-N/2}(1-\beta^N) . \tag{32}$$

B. Again, we assume that N is fixed and now we derive formulas for $R_N(q)$ and $S_N(q)$ for $q \to \beta^{\pm N/2}$ (R_N is defined by (15a) and S_N is defined by (8.5)). Denoting $(q - \beta^{N/2})/d_N$ by ε and $R_N(q)$ by ρ we have

$$q = \pi_N(\rho) = \beta^{N/2} + d_N\rho^{-2} + O(\rho^{-3}) \qquad (q \to \beta^{N/2}) ,$$

since $\rho \to \infty$ if $q \to \beta^{N/2}$. Hence, $\varepsilon = \rho^{-2} + O(\rho^{-3})$ $(\varepsilon \to 0)$, which yields $\rho = \varepsilon^{-\frac{1}{2}} + O(1)$ $(\varepsilon \to 0)$. Thus, we have found

$$R_N(q) = d_N^{\frac{1}{2}}(q - \beta^{N/2})^{-\frac{1}{2}} + O(1) \qquad (q \downarrow \beta^{N/2}) , \tag{33}$$

and analogously

$$R_N(q) = -\,d_N^{\frac{1}{2}}(\beta^{-N/2} - q)^{-\frac{1}{2}} + O(1) \qquad (q \uparrow \beta^{-N/2}) . \tag{34}$$

Formulas for $S_N(q)$ can be obtained by quadrature using (23) and (25). This yields

$$S_N(q) = a_N - 2d_N^{\frac{1}{2}}(q - \beta^{N/2})^{\frac{1}{2}} + O(q - \beta^{N/2}) \qquad (q \downarrow \beta^{N/2}) , \tag{35}$$

and

$$S_N(q) = b_N - 2d_N^{\frac{1}{2}}(\beta^{-N/2} - q)^{\frac{1}{2}} + O(\beta^{-N/2} - q) \qquad (q \uparrow \beta^{-N/2}) . \tag{36}$$

These formulas imply in particular that the graph of S_N has vertical slope at the endpoints.

C. Now let $\rho \in (-s^*, s^*)$ be fixed, let $N \to \infty$, and consider $\Phi_N(\rho)$, $\omega_N(\rho)$, $\pi_N(\rho)$ (s^* is defined by (4.12)). Then we get (cf. [13] Ch. 8)

$$\Phi_N(\rho) = s^* + O((r^*)^N) \qquad (N \to \infty) , \tag{37}$$

$$\omega_k(\rho) = r^* + O((r^*)^N) \qquad (N \to \infty) , \tag{38}$$

$$\pi_N(\rho) = C(\rho)(r^*)^N(1 + O((r^*)^N)) \qquad (N \to \infty) . \tag{39}$$

Here r^* is defined by (5.10): $r^* = h(s^*) = \mu'(s^*)$ and $C(\rho)$ is some constant not depending on N. The last formula is obtained from

$$\log \pi_N(\rho) = \sum_{k=0}^{N-1} \log \omega_k(\rho) = N \log r^* + \sum_{k=0}^{N-1} O((r^*)^k) =$$

$$= N \log r^* + C_1 + O((r^*)^N)$$

where C_1 does not depend on N.

D. Let $q > 0$ be fixed and $N \to \infty$, and consider $R_N(q)$, $S_N(q)$. We first observe that $-s^* < R_N(q) < \Phi_N(R_N(q)) < s^*$ holds for sufficiently large N. In fact, if $R_N(q) \le -s^*$, then $\Phi_k(R_N(q)) \le -s^*$ $(k = 0,\dots,N-1)$, and hence $\omega_k(R_N(q)) \ge (r^*)^{-1}$ $(k = 0,\dots,N-1)$. But then we have $q = \pi_N(R_N(q)) \ge (r^*)^{-N}$ which is impossible for sufficiently large N. In a similar way we get a contradiction from the assumption that $R_N(q) \ge s^*$. Hence, $-s^* < R_N(q) < s^*$. But then it follows from the monotonicity of μ and from $\mu(s^*) = s^*$, that $R_N(q) < \Phi_N(R_N(q)) < s^*$.

Furthermore, we will show that

$$R_N(q) \to -s^* \qquad (N \to \infty) , \tag{40}$$

and

$$S_N(q) \to s^*(1 + q) \qquad (N \to \infty) . \tag{41}$$

Suppose that (40) does not hold. Then there exists $\delta > 0$ with $\delta < 2s^*$ and an increasing sequence $\{N_1, N_2, \ldots\}$ such that $R_{N_k}(q) \geq -s^* + \delta$ $(k = 1,2,\ldots)$. But then we have

$$q = \pi_{N_k}(R_{N_k}(q)) \leq \pi_{N_k}(-s^* + \delta) \to 0 \qquad (k \to \infty)$$

(see (39)) which is a contradiction. We also have $\Phi_N(R_N(q)) \to s^*$. If this were not the case, there would be $\delta > 0$ and strictly increasing sequence $\{N_1, N_2, \ldots\}$ such that $\Phi_{N_k}(R_{N_k}(q)) \leq s^* - \delta$. But then it would follow from (21) that

$$q = \pi_{N_k}(R_{N_k}(q)) = 1/\pi_{N_k}(-\Phi_{N_k}(R_{N_k}(q))) \geq 1/\pi_{N_k}(-s^* + \delta) \to \infty \quad (k \to \infty)$$

which is again contradictory. Formula (41) now follows from (15b).

REMARK 6. Formula (41) can be made intuitive in the following way. If for fixed q and very large N one has to find an N-step trajectory x_u with $x_u(t_N) = q$, the final time t_N will be reasonable small if $|x_u(t)|$ is small on a large number of steps. Therefore the optimal trajectory will have the following nature: First we go from (0,1) to a point very close to the origin. By the result of section 5 we can do this in a time interval of approximate length s^*. Having done this, we go back to (0,q). This takes us as much time as from (0,q) to (0,0), which can be done in a time interval of approximate length qs^* (solution of problem P_3 with reversed time). Therefore the total time is about $s^*(1 + q)$.

REMARK 7. By the symmetry properties of the problem it is possible to calculate $S_N(1)$ and $R_N(1)$ explicitly. In fact, we have: If $N = 2K$, then

$$R_N(1) = -\Phi_K(0) \ , \quad S_N(1) = 2\Phi_K(0) \ . \tag{42}$$

If $N = 2K + 1$, then

$$R_N(1) = -\Phi_K(\sigma) \ , \quad S_N(1) = 2\Phi_K(\sigma) \ , \tag{43}$$

where $\sigma = \frac{1}{2}(1 + \beta)$.

PROOF. Let $\bar{\rho} := R_N(1)$, hence $\pi_N(\bar{\rho}) = 1$. Because of (21) we have $\pi_N(-\Phi_N(\bar{\rho})) = 1$, so $-\Phi_N(\bar{\rho})$ is also a solution of (14). But the solution of (14) is unique and therefore we must have

$$\Phi_N(\bar{\rho}) = -\bar{\rho} \ . \tag{44}$$

Since Φ_N is strictly increasing, $\bar{\rho}$ is uniquely determined by (44). Furthermore, by (15a) we have

$$S_N(1) = -2\bar{\rho} = -2R_N(1) \ . \tag{45}$$

If N is even, say $N = 2K$, we have

$$\Phi_K(\Phi_K(\bar{\rho})) = \Phi_{2K}(\bar{\rho}) = -\bar{\rho} = \Phi_K(-\Phi_K(\bar{\rho}))$$

(see (20)), which implies $\Phi_K(\bar{\rho}) = -\Phi_K(\bar{\rho})$, and hence $\Phi_K(\bar{\rho}) = 0$. Again by (20) we obtain $\bar{\rho} = -\Phi_K(-\Phi_K(\bar{\rho})) = -\Phi_K(0)$. Together with (45) this implies (42).

If N is odd, say $N = 2K + 1$, define $\sigma := -\Phi_K(\bar{\rho})$. Then we have

$$\bar{\rho} = -\Phi_K(-\Phi_K(\bar{\rho})) = -\Phi_K(\sigma) \ .$$

and furthermore

$$\Phi_K(\mu(-\sigma)) = \Phi_{K+1}(-\sigma) = \Phi_{2K+1}(\bar{\rho}) = -\bar{\rho} = \Phi_K(\sigma) \ .$$

Hence $\mu(-\sigma) = \sigma$. It is easily seen that this implies $\sigma = \frac{1}{2}(1 + \beta)$ (see formula (4.10) and fig. 4.3). This proves (43).

Note that the relation (45) also can be proved by differentiating (8.7) with respect to q, using (16) and setting $q = 1$.

Finally we remark that the formulas (42) and (43) can be derived from Theorem 2.1 and the observation that for the $P_5(1,1,N)$-optimal strategy $\bar{x}$ we have $\bar{x}(T - t) = \bar{x}(t)$.

II.10. *The case* $\alpha = 1$

We consider in this section the problems $P_1,\dots,P_6$ for the case in which $\alpha = 1$.

$P_1(p,q)$: The analysis here is very similar to the one of section 3. Therefore we will indicate only the results, with special emphasis on the points where these results deviate from the ones in which $\alpha < 1$. Just as in section 3, the optimal control is unique and of the "bang-bang" type, taking on only the values ± 1. But the parabolas $\pi(-1,p)$ and $\pi(-1,q)$ (for the notation see section 3) degenerate to straight lines parallel to the x-axis. If $q = rp$, the optimal control exists for $r \in J := (0,\infty)$. The formulas for the switching point $(\bar{x},\bar{y})$ and the function $f(r)$ can be obtained by taking the limit $\alpha \to 1$ (or $\beta \to 0$) or by direct calculation.

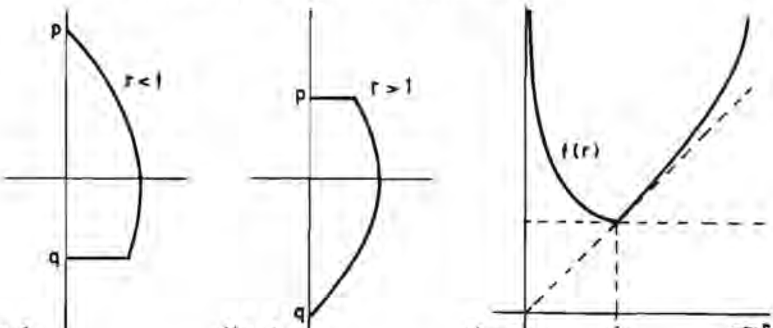


fig 1 fig 2. fig 3.

Thus, we obtain: If $r \le 1$, we have $(\bar{x},\bar{y}) = (\frac{1}{4}(1-r^2)p^2,-rp)$ and $f(r) = (4r)^{-1}(1+r)^2$. If $r > 1$, then $(\bar{x},\bar{y}) = (\frac{1}{4}(r^2-1)p^2,rp)$ and $f(r) = \frac{1}{4}(1+r)^2$. The minimum time here is also given by $T(p,q) = f(r)p$. The function f is strictly convex and $f(r)/r \to \infty$ $(r \to \infty)$, $f(r) \to \infty$ $(r \to 0)$.

The concepts "p-start, N-step" and "piecewise optimal" and the functions p_N, t_N are defined like in section 3. For problems $P_2,\dots,P_6$ we can restrict ourselves to piecewise optimal controls. However, since J is no longer compact, the existence of optimal strategies is not obvious. Let us start with problem P_6. We will proceed just as in section 5. First we assume the existence of an optimal strategy. Defining θ_N, Θ_N, Φ_N, ω_N, π_N just as in section 9, with μ and h defined as in section 4 (it is easily seen that these functions exist), we can introduce the strategy $\bar{\Lambda}_N = \{\bar{r}_0,\ldots,\bar{r}_{N-1}\}$, with $\bar{r}_k$ as defined in (9.6). We show by induction that $\bar{\Lambda}_N$ actually constitutes the unique $P_6'(p,\rho,N)$-optimal strategy. For N = 1, we have

$$\Phi_1(\rho) = \min_{r\in J} [f(r) + r\rho] = f(\bar{r}_0) + \bar{r}_0\rho \le f(r) + r\rho$$

for all $r \in J$ with equality only if $r = \bar{r}_0$. Hence,

$$\theta_1(p,\rho,\bar{\Lambda}_1) = t_1(p,\bar{\Lambda}_1) + \rho p_1(p,\bar{\Lambda}_1) = (f(\bar{r}_0) + \bar{r}_0\rho)p \le$$

$$\le (f(r_0) + r_0\rho)p = \theta_1(p,\rho,\Lambda)$$

for all $\Lambda = \{r_0\} \in J_1$, with equality only if $\Lambda = \bar{\Lambda}_1$. Now consider an arbitrary N, and assume that $\bar{\Lambda}_{N-1}$ has been shown to be $P_6'(p,\rho,N-1)$-optimal for all $p > 0$. Then we observe from (9.6) that if $\bar{\Lambda}_N = \{\bar{r}_0,\ldots,\bar{r}_N\}$, the strategy $\bar{\Lambda}_N' = \{\bar{r}_1,\ldots,\bar{r}_N\}$ equals $\bar{\Lambda}_{N-1}$. This implies:

$$\theta_N(p,\rho,\bar{\Lambda}_N) = f(\bar{r}_0)p + \bar{r}_0\theta_{N-1}(p,\rho,\bar{\Lambda}_{N-1}) \le$$

$$\le f(r)p + r\theta_{N-1}(p,\rho,\bar{\Lambda}_{N-1}) \le f(r)p + r\theta_{N-1}(p,\rho,\Lambda)$$

for all $r \in J$ and all $\Lambda \in J^{N-1}$, with equality only if $r = \bar{r}_0$ and $\Lambda = \bar{\Lambda}_{N-1}$. The first inequality follows from the definition of $\bar{r}_0$ (since $\theta_{N-1}(p,\rho,\bar{\Lambda}_{N-1}) = \Theta_{N-1}(p,\rho) = p\Phi_N(\rho)$), and the second inequality is a consequence of the strict optimality of $\bar{\Lambda}_{N-1}$. Hence, $\bar{\Lambda}_N$ is $P_6'(p,\rho,N)$-optimal.

Problem P_2 is a special case of P_6. As we saw in section 9 the solutions of problems P_4 and P_5 can be constructed from the one of P_6 without presupposing the existence of P_4- and P_5-optimal strategies. Problem P_3 can be dealt with exactly as in section 5.

We give some formulas for $\alpha = 1$:

$$\mu(s) = \tfrac{1}{2}(1 + \sqrt{1+4s}) \qquad (s > 0) ,$$

$$\mu(s) = 1 + s \qquad (-1 \le s \le 0) ,$$

$$\mu(s) = - s(1 + s) \qquad (s < -1) ,$$

and

$$h(s) = (1 + 4s)^{-\frac{1}{2}} \qquad (s > 0) ,$$

$$h(s) = 1 \qquad (-1 \le s \le 0) ,$$

$$h(s) = - 1 - 2s \qquad (s < -1) .$$

Furthermore we have $s^* = 2$, $r^* = \frac{1}{3}$ and the switching curve of problem P_3 is given by $x = 2y|y|$.

The asymptotic formulas given in section 9 have to be modified here. After some calculations one finds

$$S_N(q) = a_N q^{-b_N} + O(1) \qquad (q \to 0) ,$$

$$S_N(q) = c_N q^{d_N} + O(q) \qquad (q \to \infty) ,$$

where

$$a_N = (2^N - 1)^{-1}\, 2^{-N/(2^N-1)} ,$$

$$b_N = (2^N - 1)^{-1} ,$$

$$c_N = (2^N - 1)^{-1}\, 2^{N/(2^N-1)} ,$$

$$d_N = 2^N(2^N - 1)^{-1} .$$

Furthermore, formula (9.41):

$$S_N(q) \to 2(1 + q) \qquad (N \to \infty)$$

is also valid in this case.

CHAPTER III

APPLICATION OF THE DISCRETE MAXIMUM PRINCIPLE

III.1. *The discrete maximum principle*

Although the dynamic programming method used in the previous chapter is an elementary method, it usually results in a functional-recurrence relation which is not easily solved explicitly. The special character of the problems of the previous chapter which made it possible to apply successfully the dynamic programming method, is the homogeneous dependence on p of the functions p_k and t_k defined in(3.19), (3.20). But even for this simple situation we have seen in section II.8 an instance of the difficulties one encounters in the dynamic programming method. The equation (II.8.6) is a typical example of such a functional-recurrence relation. Fortunately, we could solve the problem of section 8 by an auxiliary problem discussed in section 9. However, for more general situations such an artifice usually is not possible, and we are left with an equation of the type (II.8.6).

In this chapter we discuss a different method for solving discrete optimal control problems which is based on a rather complicated theorem, called the *discrete maximum principle*. This theorem is the discrete analogon of the well-known maximum principle of Pontryagin for continuous (that is, differential) systems (sometimes called the *continuous maximum principle*), a special case of which we have met in section I.5. In its general form the discrete maximum principle is due to H. Halkin (see [3]). A simpler proof can be found in [14].

DEFINITION 1. If N is a natural number, then

$$\hat{N} := \{0,\ldots,N-1\}\,. \tag{1}$$

The set of natural numbers is denoted by Nt.

DEFINITION 2. If $N \in Nt$, $U \subset R^m$ for some positive integer m, then

$H(R^n \times U \times \hat{N} \to R^n) :=$

$$:= \{f \in C(R^n \times U \times \hat{N} \to R^n) \mid (x \mapsto f(x,v,k)) \in C^2(R^n \to R^n)$$

$$(v \in U,\ k \in \hat{N})\} .$$

Now suppose we are given a vector $a \in R^n$, a set $U \subset R^m$, an element $N \in Nt$ and the following functions

α) $f \in H(R^n \times U \times \hat{N} \to R^n)$,

β) $g \in C^2(R^n \to R^p)$,

γ) $w \in C^2(R^n \to R^1)$.

Then we will call the elements of $\hat{N} \to U$ *controls*. For a given control u, the function $x_u \in (\widehat{N+1} \to R^n)$, defined by

$$x_u(0) = a ,$$
$$x_u(k+1) = f(x_u(k),u(k),k) \qquad (k \in \hat{N}) , \tag{2}$$

is called the *response* of u. A control u is called *admissible* if $g(x_u(N)) = 0$. The set of admissible controls is denoted by Ω:

$$\Omega := \{u \in (\hat{N} \to U) \mid g(x_u(N)) = 0\} . \tag{3}$$

An admissible control $\bar{u}$ is called *optimal* if for every $u \in \Omega$ we have $w(x_u(N)) \geq w(x_{\bar{u}}(N))$. Hence, an optimal control minimizes $w(x_u(N))$ under the restriction $g(x_u(N)) = 0$. With these definitions we have the following theorem:

THEOREM 1 (Halkin). Let $\bar{u}$ be an optimal control and let $\bar{x} := x_{\bar{u}}$ be the corresponding response. If for every $x \in R^n$, $k \in \hat{N}$ the set $f(x,U,k) := \{f(x,v,k) \mid v \in U\}$ is convex and if

$$\operatorname{rank}\begin{pmatrix}\partial g(\bar{x}(N))\\ \partial w(\bar{x}(N))\end{pmatrix} = p + 1 , \tag{4}$$

then there exists a non-trivial $\psi \in (\widehat{N+1} \to R^n_T)$ with the properties

i) $\psi(k) = \psi(k+1)\partial_x f(\bar{x}(k),\bar{u}(k),k) \quad (k \in \hat{N})$,

ii) $\psi(N) = \lambda\partial w(\bar{x}(N)) + \mu\partial g(\bar{x}(N))$ for some $\lambda \geq 0$, $\mu \in R^p_T$,

iii) $\psi(k+1)f(\bar{x}(k),\bar{u}(k),k) = \min_{v\in U} \psi(k+1)f(\bar{x}(k),v,k) \quad (k \in \hat{N})$.

REMARK 1. In [3] the additional condition "$\partial_x f(x(k),u(k),k)$ is non-singular" is posed, but as is shown in [14] this condition is superfluous. Note that also we have formulated the problem as a minimum problem, whereas in [3] the problem is considered where a function has to be maximized. Of course these problems are equivalent.

We will derive a modified version of Theorem 1 which will be useful for the problems discussed in this chapter. Suppose that we are given functions f and g just as in α) and β), but that instead of w we have a function $f_0 \in H(R^n \times U \times \hat{N} \to R)$. We define the response x_u and the set of admissible controls Ω as before, but now we call $\bar{u} \in \Omega$ *optimal* if

$$\sum_{k=0}^{N-1} f_0(x_{\bar{u}}(k)\bar{u}(k),k) \leq \sum_{k=0}^{N-1} f_0(x_u(k),u(k),k) \tag{5}$$

holds for all $u \in \Omega$. Then we have:

THEOREM 2. Let $\bar{u}$ be an optimal control, and let $\bar{x} := x_{\bar{u}}$ be the response of u. If for every $x \in R^n$, $k \in \hat{N}$, the set

$$V(x,k) := \{(z_0,z')' \in R^{n+1} \mid z = f(x,v,k) \wedge z_0 \geq f_0(x,v,k) \wedge v \in U\}$$

is convex, and if rank $\partial g(\bar{x}(N)) = p$, then there exists $\lambda \geq 0$ and $\psi \in (\widehat{N+1} \to R^n_T)$ with the properties:

i) $(\lambda, \psi(N)) \neq 0$,

ii) $\psi(k) = \psi(k+1)\partial_x f(\bar{x}(k), \bar{u}(k), k) + \lambda \partial_x f_0(\bar{x}(k), \bar{u}(k), k) \quad (k \in \hat{N})$,

iii) $\psi(N) = \mu \partial g(\bar{x}(N))$ for some $\mu \in R^p_T$,

iv) $H(\bar{u}(k), k) = \min_{v \in U} H(v, k) \quad (k \in \hat{N})$,

where $H \in (U \times \hat{N} \to R^1)$ is defined by

$$H(v,k) := \psi(k+1) f(\bar{x}(k), v, k) + \lambda f_0(\bar{x}(k), v, k) \ . \tag{6}$$

PROOF. We introduce the following notation:

$$\underline{x} := \begin{pmatrix} x_0 \\ x \end{pmatrix} \in R^{n+1} \text{ , where } x_0 \in R^1, \ x \in R^n \ ;$$

$$\underline{u} := \begin{pmatrix} u_0 \\ u \end{pmatrix} \in R^{m+1} \text{ , where } u_0 \in R \ , \ u \in R^m \ ;$$

$$\underline{U} := \{\underline{u} \in R^{n+1} \mid u \in U \wedge u_0 \geq 0\} \ ;$$

$$\underline{\Omega} := \{\underline{u} \in (\hat{N} \to U) \mid g(x_u(N)) = 0\} \ .$$

$\underline{f} \in (R^{n+1} \times \underline{U} \times \hat{N} \to R^{n+1})$ is defined by

$$f(\underline{x}, \underline{u}, k) := \begin{pmatrix} x_0 + f_0(x,u,k) + u_0 \\ f(x,u,k) \end{pmatrix};$$

$$\underline{w}(\underline{x}) := x_0 \ ; \quad \underline{a} := \begin{pmatrix} 0 \\ a \end{pmatrix} \ ; \quad \underline{g}(\underline{x}) := g(x) \ .$$

Then we consider the optimal control problem of theorem 1, with $n+1$, $m+1$, $\underline{x}$, $\underline{u}$, $\underline{f}$, $\underline{g}$, $\underline{w}$, $\underline{a}$, $\underline{U}$, $\underline{\Omega}$ instead of n, m, x, u, f, g, w, a, U, Ω. It is easily seen that $\underline{f}$, $\underline{g}$, $\underline{w}$ satisfy the conditions α), β) and γ). Also we observe that $\underline{\bar{u}} := \begin{pmatrix} 0 \\ \bar{u} \end{pmatrix}$ is an optimal control. In fact, if we denote $\underline{x}_{\underline{\bar{u}}}$ by $\underline{\bar{x}}$, then $\underline{g}(\underline{\bar{x}}(N)) = g(\bar{x}(N)) = 0$. And for every $\underline{u} = \begin{pmatrix} u_0 \\ u \end{pmatrix} \in \underline{\Omega}$ we have

$$\underline{w}(\underline{x}_{\underline{u}}(N)) = \sum_{k=0}^{N-1} \{f_0(x_u(k),u(k),k) + u_0(k)\} \geq$$

$$\geq \sum_{k=0}^{N-1} f_0(\bar{x}(k),\bar{u}(k),k) = \underline{w}(\bar{\underline{x}}(N)) \ ,$$

since $\bar{u}$ is optimal and $u_0(k) \geq 0$. Furthermore, we see that the conditions of Theorem 1 are satisfied. In fact

$$\text{rank} \begin{pmatrix} \partial \underline{g}(\bar{\underline{x}}(N)) \\ \partial \underline{w}(\bar{\underline{x}}(N)) \end{pmatrix} = \text{rank} \begin{pmatrix} 0 & \partial g(\bar{x}(N)) \\ 1 & 0 \end{pmatrix} = p + 1 \ ,$$

and $\underline{f}(\underline{x},\underline{U},k) = \begin{pmatrix} x_0 \\ 0 \end{pmatrix} + V(x,k)$ is convex for every $(\underline{x},k) \in R^{n+1} \times \hat{N}$.

It follows then from Theorem 1 that there exists a non-trivial $\underline{\psi} \in (\widehat{N+1} \to R_T^{n+1})$, which (if we set $\underline{\psi} = (\psi_0,\psi)$) satisfies:

i) $\psi_0(k) = \psi_0(k+1)$,

$$\psi(k) = \psi(k+1)\partial_x f(\bar{x}(k),\bar{u}(k),k) + \psi_0(k+1)\partial_x f_0(\bar{x}(k),\bar{u}(k),k) \quad (k \in \hat{N}) \ ;$$

ii) $\psi_0(N) = \lambda$, $\psi(N) = \mu\partial g(\bar{x}(N))$ for some $\lambda \geq 0$, $\mu \in R_T^p$;

iii) $\psi(k+1)f(\bar{x}(k),\bar{u}(k),k) + \psi_0(k+1)f_0(\bar{x}(k),\bar{u}(k),k) =$

$$= \min_{v \in U} \{\psi(k+1)f(\bar{x}(k),v,k) + \psi_0(k+1)f_0(\bar{x}(k),v,k)\} \ ,$$

and this yields the result.

REMARK 2. If there is no restriction on the final value of x_u (that is, if g does not appear in theorem 2, and hence all $u \in (\hat{N} \to U)$ are admissible), then $p = 0$ and the rank condition is satisfied. In this case iii) has to be interpreted as $\psi(N) = 0$. It follows then from i) that $\lambda > 0$. Since (λ,ψ) appears homogeneous in Theorem 2 we may assume in this case that $\lambda = 1$.

III.2. *Application of the maximum principle to problems of Chapter II*

We apply Theorem (1.2) to the problems P_2' and P_5' treated in chapter II.

First consider problem P_2'. In the notation used in the formulation of Theorem (1.2) we have as a special case of the recurrence (1.2):

$$\begin{aligned} &x_u(0) = p > 0 \;, \\ &x_u(k+1) = u(k)x_u(k) \qquad (k \in \hat{N}) \;, \end{aligned} \tag{1}$$

where x and u are scalar valued $(n = m = 1)$ and $U = J = [\beta^{\frac{1}{2}},\beta^{-\frac{1}{2}}]$. The function u should not be confused with the control function u of problem P_2. Here u is a function defined on $\hat{N}$ instead of on an interval, and we have $u(k) = r_k$ $(k = 0,\dots,N-1)$, with r_k as defined in problem P_2'. Furthermore we have $f_0(x,u,k) = f(u)x$. Since f is convex and $\partial g(\bar{x}(N)) = 0$ (cf. remark (1.2)), we can apply Theorem (1.2) and set $\lambda = 1$ in order to find necessary conditions for the optimal control: If $\bar{u}$ is optimal there exists $\psi \in (\widehat{N+1} \to R^1)$ such that

i) $\qquad \psi(k) = \psi(k+1)\bar{u}(k) + f(\bar{u}(k)) \qquad (k \in \hat{N})$,

ii) $\qquad \psi(N) = 0$,

iii) $\qquad \psi(k+1)\bar{u}(k)\bar{x}(k) + f(\bar{u}(k))\bar{x}(k) =$

$$= \min_{v \in U} [v\psi(k+1)\bar{x}(k) + f(v)\bar{x}(k)] \qquad (k \in \hat{N}) .$$

Since $p > 0$ and $u(k) > 0$ $(k \in \hat{N})$ if $u \in (\hat{N} \to J)$, we see that $x(k) > 0$ $(k \in \widehat{N+1})$. Hence, if we use the function μ defined in (II.4.7), we have:

i) $\qquad \psi(N) = 0,\ \psi(k) = \mu(\psi(k+1)) \qquad (k \in \hat{N})$;

ii) $\qquad \psi(k+1)\bar{u}(k) + f(\bar{u}(k)) = \mu(\psi(k+1))$.

It follows from ii) that $\bar{u}(k) = h(\psi(k+1))$ where h is the function defined in (II.4.9). Put $\psi(k) = S_{N-k}$ $(k \in \hat{N})$, where S_k is defined by (II.4.5); then we have obtained the same result as in section (II.4) if we have shown that

$$T_N(a) := \sum_{k=0}^{N-1} f(\bar{u}(k))\bar{x}(k) = \psi(0)a \ .$$

But this is an immediate consequence of the equalities

$$\psi(k)\bar{x}(k) = (\psi(k+1)\bar{u}(k) + f(\bar{u}(k)))\bar{x}(k) = \psi(k+1)\bar{x}(k+1) + f(\bar{u}(k))\bar{x}(k)$$

and $\qquad \psi(N) = 0,\ x(0) = a$.

Problem P_5 differs from problem P_2 only by the condition $x(N) = q$, where b is some positive number. Hence, we have $g(x) = x - q$ (in the notation of Theorem (1.2)). Since rank $\partial g(x) = 1$, we can apply Theorem (1.2) again. Let us assume for simplicity that $p = 1$. This is no loss of generality because of the homogeneity of the problem. Furthermore, just as in section (II.8) we observe that there exists an admissible control if and only if $b \in J_N := [\beta^{-N/2}, \beta^{N/2}]$. Also, as we have seen in section (II.9), we may assume that $\beta^{-N/2} < b < \beta^{N/2}$, since otherwise the admissible control is unique. Now it follows from Theorem (1.2) that if $\bar{u}$ is optimal, there exist $\lambda \geq 0$, $\psi \in (\widehat{N+1} \to R^n_T)$ such that

i) $\qquad \psi(k) = \psi(k+1)\bar{u}(k) + \lambda f(\bar{u}(k)) \qquad (k \in \hat{N})$,

ii) $\qquad \psi(N) = \rho \quad$ for some $\rho \in R^1$,

iii) $\qquad \psi(k+1)\bar{u}(k) + \lambda f(\bar{u}(k)) = \min_{v \in J} [\psi(k+1)v + \lambda f(v)] \quad (k \in \hat{N})$,

iv) $\qquad (\lambda, \rho) \neq 0$.

It follows from these relations that $\lambda > 0$, since otherwise $\operatorname{sgn} \psi(k) = \operatorname{sgn} \rho \quad (k \in \widehat{N+1})$ and hence (by iii)) $\bar{u}(k) = \beta^{\frac{1}{2}} \quad (k \in \hat{N})$ if $\rho > 0$ and $\bar{u}(k) = \beta^{-\frac{1}{2}} \quad (k \in \hat{N})$ if $\rho < 0$. But this would imply $b = \beta^{\pm N/2}$, contradicting our original assumption. Hence, we may assume that $\lambda = 1$ and we obtain

i) $\qquad \psi(k) = \mu(\psi(k+1)) \quad (k \in \hat{N})$, $\ \psi(N) = \rho$,

ii) $\bar{u}(k) = h(\psi(k+1)) \qquad (k \in \hat{N})$,

and this result coincides with the one obtained in section II.9, setting $\psi(k) = \Phi_{N-k}(\rho)$, $\bar{u}(k) = \bar{r}_k = \omega_{N-k-1}(\rho)$. That is, just as in section II.9 the problem P_5 is solved if we have found the value of ρ for which we have $x_{\bar{u}}(N) = b$, where $\bar{u}$ is defined by i) and ii). The fact that

$$S_N(b) := \Sigma\, f(\bar{u}(k))\bar{x}(k) = \psi(0) - \rho b$$

can be proved in a way analogous to the one used in the case of problem P_2.

III.3. *A controlled system with Coulomb friction*

Consider the following control system:

$$\ddot{x} + \operatorname{sgn} \dot{x} + x = u \ , \tag{1}$$

with the functions $u \in L([0,\infty) \to [-1,1])$ as controls. We rewrite (1) as

$$\begin{aligned} \dot{x} &= y \ , \\ \dot{y} &= -x - \operatorname{sgn} y + u \ , \end{aligned} \tag{2}$$

or in vector form $\dot{\underline{x}} = \underline{f}(\underline{x},u)$ where $\underline{x} = (x,y)'$ and $\underline{f}(\underline{x},u) = (y, -x - \operatorname{sgn} y + u)'$. We can apply Theorem (I.4.2) to this system in the region $S := \{\underline{x} \in R^2 \mid |x| > 2 \vee y \neq 0\}$. According to the theorem there exists for every control u and every $\underline{a} \in S$ a solution of (2) on either $[0,\infty)$ or $[0,T_0)$ for some $T_0 > 0$, where in the last case we have $\underline{x}(t) \to \underline{x}_0$ $(t \uparrow T_0)$ for some $\underline{x}_0 \in \partial S = \{\underline{x} \in R^2 \mid |x| \leq 2 \wedge y = 0\}$. Whether the solution can be extended outside of S depends on u. For simplicity we will restrict ourselves to trajectories which are in S.

It follows from (2) that for a trajectory $\underline{x}$ the function $t \mapsto x(t)$ is increasing in the upper half-plane and decreasing in the half-plane $y > 0$. Furthermore, it follows from the symmetry of the problem that $\underline{x}_u(t,\underline{a}) = \underline{x}_{-u}(t,-\underline{a})$ where $t \mapsto \underline{x}_u(t,\underline{a})$ is the trajectory of (2) corresponding to the control u and with initial condition $\underline{x}(0) = \underline{a}$.

For a given natural number N and a real number $a > 2$ we will call a trajectory and the corresponding control a-*start*, N-*step* if the trajectory has initial value $(-a,0)$ and intersects the x-axis N times for $t = t_i$ $(i = 1,\dots,N)$, where $t_0 := 0 < t_1 < \dots < t_N$. The points $(x(t_k),0)$ $(k = 0,\dots,N)$ are called *alteration points* and the numbers $a_k := (-1)^{k+1}x(t_k)$ are called *alteration moduli*. We will require always that $a_k > 2$ $(k = 1,\dots,N)$ holds. The numbers t_k are called *alteration times*.

The problem we are going to consider is the following one:

$Q(a,b,N)$: "For fiven numbers $a,b > 2$ and natural number N find an a-start, N-step control which satisfies $a_N = b$ and such that t_N (called the *final time*) is minimal".

In this section we restrict ourselves to the case $N = 1$. In this case the trajectory has to satisfy $\underline{x}(0) = (-a,0)$, $\underline{x}(T) = (b,0)$ for some $T > 0$, and $y(t) > 0$ $(0 < t < T)$. We will call a control *admissible* if there exists a trajectory satisfying these conditions. It is easily seen that there exists no admissible control if $b \notin [a-4,a]$. In fact, it follows from

$$\tfrac{1}{2}\frac{d}{dt}(x^2 + y^2) = -|y| + uy \le 0 \tag{3}$$

that $b \le a$, and it follows from

$$\tfrac{1}{2}\frac{d}{dt}[(x+2)^2 + y^2] = (1 + u)y \ge 0$$

that $b \ge a-4$ for every trajectory. Also, it is clear from this argument that for $b = a$ the control $u = 1$ is the only admissible control, and for $b = a-4$ the only admissible control is $u = -1$. We can restate problem $Q(a,b,1)$ now as follows:

$\tilde{Q}(a,b)$: "Given numbers a, b satisfying $b \ge 2$ and $a-4 \le b \le a$, find a control u, such that for the corresponding trajectory of the system

$$\dot{x} = y \ ,$$
$$\dot{y} = - x - 1 + u \ , \tag{4}$$

satisfying $\underline{x}(0) = (-a,0)$ we have $\underline{x}(T) = (b,0)$ with minimal T".

We have ignored the condition $y > 0 \quad (0 < t < T)$ for the moment, since it will turn out to be satisfied automatically. We write (4) in vector notation $\underline{\dot{x}} = A\underline{x} + \underline{f}(u)$, with

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} , \quad \underline{f}(u) = \begin{pmatrix} 0 \\ u-1 \end{pmatrix} ,$$

and we apply Theorem (I.5.1). Let $\bar{u}$ be an optimal control. Then $\bar{u}$ satisfies the equation

$$\underline{\psi}(t)\underline{f}(\bar{u}(t)) = \max_{|v|\le 1} \ \underline{\psi}(t)\underline{f}(v)$$

a.e. for some non-trivial solution $\underline{\psi}$ of the adjoint equation $\underline{\dot{\psi}} = - \underline{\psi}A$ (or $\dot{\psi}_1 = \psi_2$, $\dot{\psi}_2 = - \psi_1$). It follows that ψ_2 is of the form $\psi_2(t) = r \cos(t + \theta)$ for some r,θ with $r > 0$. Hence, $\bar{u}(t) = \operatorname{sgn} \psi_2(t)$ a.e. and we may assume that $\bar{u}(t) = \operatorname{sgn} \psi_2(t)$ holds for all t, since the values of $\bar{u}$ on a set of measure zero do not influence the trajectory. As a consequence, $\bar{u}$ assumes only the values ± 1 and there is a time interval π between two consecutive switching times (if there is more than one switch). We show that there is at most one switch in $(0,T)$ by proving the existence of an admissible control with final time $T \le \pi$.

If $b = a$ or $b = a-4$, it follows from the foregoing that $\bar{u}$ is constant (± 1) and this yields $T = \pi$ in both cases. We assume henceforth that $a-4 < b < a$.

The trajectories corresponding to the control $u = 1$ are

$$x = - r \cos(t + \phi)$$
$$y = \ \ r \sin(t + \phi) \tag{6}$$

with constant r, ϕ satisfying $r > 0$. In the phase plane these trajectories are determined by the equation

$$x^2 + y^2 = c^2 \ , \quad y > 0 \ . \tag{7}$$

Let us denote the curve given in (7) by $\gamma_+(c)$.

The trajectories corresponding to the control $u = -1$ are given by

$$\begin{aligned} x &= -2 - r\cos(t+\phi) \ , \\ y &= r\sin(t+\phi) \ . \end{aligned} \tag{8}$$

In the phase plane we have

$$(x+2)^2 + y^2 = (c+2)^2 \ , \quad y > 0 \tag{9}$$

The curve given by (9) will be denoted by $\gamma_-(c)$. Since $x^2 + y^2$ is strictly decreasing on a trajectory given by (8) (compare (3)) and since $a-4 < b < a$ holds, there exists exactly one intersection point Q of $\gamma_-(a)$ and $\gamma_+(b)$, and by similar reasoning there is exactly one intersection point P of $\gamma_+(a)$ and $\gamma_-(b)$ (see fig. 1).

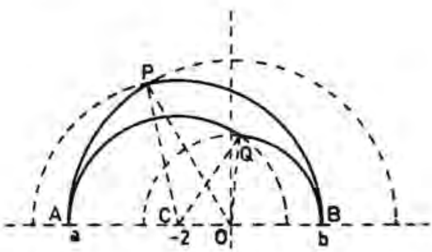


fig 1.

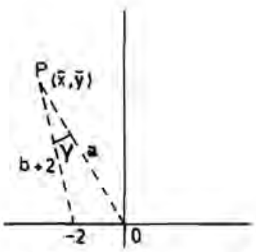


fig 2.

Let us introduce the following notation:

$$A := (-a,0), \ B := (b,0), \ C := (-2,0), \ O := (0,0) \ .$$

Furthermore, we denote by t_P, t_Q the intersection times in P, Q. Then it follows from formula (6) that $t_Q = \angle ACQ$ and it follows from (7) that $T - t_Q = \angle BOQ$. Hence, $T > \pi$ if we use the control $u = -1$ $(0 < t < t_Q)$, $u = 1$ $(t_Q < t < T)$. On the other hand, if we use the control $u = 1$ $(0 < t < t_P)$, $u = -1$ $(t_P < t < T)$, then $t_P = \angle AOP$ and $T - t_P = \angle PCB$. Hence $T = \pi - \gamma$, where $\gamma := \angle OPC > 0$. This control is therefore the optimal one. If we denote the coordinates of P by $(\bar{x},\bar{y})$, it follows by the cosine rule (see fig. 2) that

$$\cos \gamma = \frac{a^2 + (b+2)^2 - 4}{2a(b+2)} \quad .$$

Furthermore, we have $\cos T = -\cos \gamma \quad (0 < T < \pi)$ and hence

$$T = \arccos \frac{4 - a^2 - (b+2)^2}{2a(b+2)} \quad . \tag{10}$$

This result, being derived for $a-4 < b < a$, is also valid if $b = a$ and $b = a-4$, since in these cases $T = \pi$.

We sum up the result obtained for $Q(a,b,1)$

i) There exists an admissible control if and only if $a-4 \le b \le a$.
ii) For the same values of (a,b) there is a unique optimal control.
iii) The minimal time is given by (10).
iv) The optimal control is of the bang-bang type, that is, it only assumes the values ± 1.
v) If $a-4 < b < a$, the optimal control has exactly one switch on some $t_p \in (0,T)$ and we have $u(t) = 1 \quad (0 < t < t_p)$, $u(t) = -1 \quad (t_p < t < T)$.
vi) If $b = a$ or $b = a-4$, there is no switching point. If $b = a$, then we have $u(t) = 1 \quad (0 < t < T)$, and if $b = a-4$, then $u = -1 \quad (0 < t < T)$.

Furthermore, it is easily seen that similar results can be obtained for the case that the trajectory starts at the point $\underline{x}(0) = (a,0)$ and ends at $\underline{x}(T) = (-b,0)$, where $a,b > 2$, whereas $y(t) < 0$ $(0 < t < T)$ holds. Also, it is obvious by the autonomity of the problem that a similar result is obtained if we start at an arbitrary time t_0 instead of $t = 0$. Of course, if t_1 is the final time, then we have $t_1 - t_0 = T$ for the optimal control where T is given by (10).

In all these cases we will call the optimal control "$Q(a,b,1)$-optimal".

III.4. *The case of more alteration points*

We consider now problem $Q(a,b,N)$ for $N > 1$. We will call a control (N,a) *piecewise optimal* if it is an a-start, N-step control with alteration moduli $a_0 = a, a_1, \ldots, a_N$ and such that it is $Q(a_k, a_{k+1}, 1)$-optimal on the step between the alteration points $((-1)^{k-1}a_k, 0)$ and $((-1)^k a_{k+1}, 0)$ for $k = 0, 1, \ldots, N-1$. Then as in the previous chapter we observe that we can restrict ourselves to piecewise optimal controls. Therefore, with a change of notation which is closer to the notation used in the theory of section 1, we can restate $Q(a,b,N)$ as follows:

$Q'(a,b,N)$: "Given numbers $a,b > 2$ and a natural number N, find the sequence $x(0),\ldots,x(N)$ with $x(0) = a$, $x(N) = b$ and $x(k) - 4 \leq x(k+1) \leq x(k) \quad (k \in \hat{N})$ and such that $\sum_{k=0}^{N-1} T(x(k), x(k+1))$ is minimal."

Here the function T is defined by

$$T(p,q) = \arccos \frac{4 - p^2 - (q+2)^2}{2p(q+2)} \tag{1}$$

(compare formula (3.10)). We introduce the following notation:

$$v(k) := x(k+1) - x(k) + 2 \ , \quad T(p,p+r-2) =: S(p,r) \ .$$

Then we can state the problem as follows:

$Q''(a,b,N)$: "Given real numbers $a,b > 2$ and a natural number N, find a strategy $v \in (\hat{N} \to [-2,2])$ such that for the function $x \in (\hat{N} \to R^1)$, defined by

$$\begin{aligned} &x(0) = a \ , \\ &x(k+1) = x(k) + v(k) - 2 \qquad (k \in \hat{N}) \ , \end{aligned} \tag{2}$$

we have $x(N) = b$ and such that $\sum_{k=0}^{N-1} S(x(k), v(k))$ is minimal, where $S \in ((2,\infty) \times [-2,2] \to R^1)$ is defined by

$$S(x,v) := f(x,x+v) , \tag{3}$$

$$f(x,z) := \arccos(-g(x,z)) , \tag{4}$$

$$g(x,z) := (x^2 + z^2 - 4)/(2xz) . \tag{5}$$

It is obvious that there exists an optimal strategy if and only if $a - 4N \le b \le a$ holds. In the cases $b = a$ or $b = a - 4N$ the only admissible (and hence optimal) strategies are $v(k) = 2 \quad (k \in \hat{N})$ and $v(k) = -2 \quad (k \in \hat{N})$ respectively. We will assume henceforth that $a - 4N < b < a$ holds.

In order to solve Q'' we apply Theorem (1.2). First, we must show that the set $\{(x+v-2, S(x,v)+w)' \mid |v| \le 2 \wedge w \ge 0\}$ is convex for every $x > 2$, that is, we have to prove that $v \mapsto S(x,v)$ is convex. To achieve that aim, and also for further calculations, we give formulas for the derivatives of S, f and g:

$$\partial_x g(x,z) = \frac{4 - z^2 + x^2}{2x^2 z} , \tag{6}$$

$$\partial_z g(x,z) = \frac{4 - x^2 + z^2}{2xz^2} , \tag{7}$$

$$\partial_z^2 g(x,z) = \frac{x^2 - 4}{xz^3} , \tag{8}$$

$$\partial_x f(x,z) = (1 - g^2(x,z))^{-\frac{1}{2}} \partial_x g(x,z) , \tag{9}$$

$$\partial_z f(x,z) = (1 - g^2(x,z))^{-\frac{1}{2}} \partial_z g(x,z) , \tag{10}$$

$$\partial_z^2 f(x,z) = (1 - g^2(x,z))^{-3/2} \cdot$$

$$\cdot \left[(1 - g^2(x,z))\partial_z^2 g(x,z) + g(x,z)(\partial_z g(x,z))^2\right] . \tag{11}$$

REMARK 1. Here ∂_z^2 is defined by $\partial_z^2 F(z) := \partial_z(\partial_z F(z))$.

The following inequalities hold for $x > 2$, $|x-z| \le 2$:

$$0 < g(x,z) \le 1 \tag{12}$$

with equality only if $|x-z| = 2$; furthermore, $\partial_z^2 g(x,z) > 0$, and hence $\partial_z^2 f(x,z) > 0$ $(|x-z| < 2)$. Since $\partial_v^2 S(x,v) = \partial_z^2 f(x,x+v)$, we have shown that $v \mapsto S(x,v)$ is strictly convex for all $x > 2$.

Now we give some more properties which will be used in the sequel: We have $\partial_z g(x,x+2) = 2/x > 0$ and $\partial_z g(x,x-2) = -2/x < 0$. This implies that $\partial_z f(x,z) \to \infty$ $(z \uparrow x+2)$ and $\partial_z f(x,z) \to -\infty$ $(z \downarrow x-2)$.

Now we are ready to apply Theorem (1.2). Let $\bar{v}$ be an optimal strategy and let $\bar{x}$ denote the corresponding sequence (according to (2)). Then there exists $\lambda \ge 0$ and $\psi \in (\widehat{N+1} \to R^1)$ such that

i) $\quad \psi(k) = \psi(k+1) + \lambda\partial_x S(\bar{x}(k),\bar{v}(k)) \qquad (k \in \hat{N})$,

ii) $\quad \psi(N) = \mu$ for some real μ,

iii) $\quad \psi(k+1)\{\bar{x}(k) + \bar{v}(k) - 2\} + \lambda S(\bar{x}(k),\bar{v}(k)) =$

$$= \min_{-2\le v\le 2} [\psi(k+1)\{\bar{x}(k) + v - 2\} + \lambda S(\bar{x}(k),v)] .$$

First, we remark that it follows from (2) that $x(k) \ge x(k+1)$ $(k \in \hat{N})$. Hence, $\bar{x}(k) + \bar{v}(k) - 2 > 0$ $(k \in \hat{N})$. Now it is easily seen that we must have $\lambda > 0$. In fact, if $\lambda = 0$ holds, then $\psi(k) = \mu$ $(k \in \widehat{N+1})$ and hence $\bar{v}(k) = 2 \operatorname{sgn} \mu$ $(k \in \hat{N})$. In that case we have either $\bar{x}(N) = a$ or $\bar{x}(N) = a - 4N$, but this contradicts the assumption $a - 4N < b < a$. Because of homogeneity we may set $\lambda = 1$. Then we get, instead of i) and iii):

i)' $\quad \psi(k) = \psi(k+1) + \partial_x S(\bar{x}(k),\bar{v}(k)) \qquad (k \in \hat{N})$,

iii)' $\quad \psi(k+1)\bar{v}(k) + S(\bar{x}(k),\bar{v}(k)) = \min_{-2\le v\le 2} [\psi(k+1)v + S(\bar{x}(k),v)]$.

Since $v \mapsto S(x,v)$ is strictly convex for $x > 2$ it follows that the function $v \mapsto [\psi v + S(x,v)]$ assumes its minimum for every real ψ and $x > 2$ in exactly one point. Let us denote this value of v by $M(\psi,x)$.

Then we can replace iii)' by

iii)" $\bar{v}(k) = M(\psi(k+1),\bar{x}(k)) \quad (k \in N)$.

Because of the asymptotic behaviour of $v \mapsto \partial_v S(x,v)$ for $v \to \pm 2$ and because of the differentiability of S, iii)" is equivalent to

$$\psi(k+1) = - \partial_v S(\bar{x}(k),\bar{v}(k)) \ . \tag{13}$$

The optimal strategy can therefore be obtained by solving the following boundary value problem:

"Find functions $\psi \in (\widehat{N+1} \to R^1)$, $\bar{x} \in (\widehat{N+1} \to R^1)$, $\bar{v} \in (N \to [-2,2])$ satisfying (2), i)', iii)' (or (13)) and $\bar{x}(N) = b$."
If there is more than one solution of this problem, we have to select the one which minimizes $\sum_{k \in \hat{N}} S(\bar{x}(k),\bar{v}(k))$.

In order to find the function M, one has to solve (13) which can be reduced to a cubic equation. By means of a simple artifice we can modify the boundary value problem in such a way that the cubic equation can be avoided and we only have to solve a quadratic equation at each step. For that purpose we introduce the function $z \in (\hat{N} \to R^1)$ by

$$z(k) = \bar{x}(k) + \bar{v}(k) \qquad (k \in \hat{N}) \ , \tag{14}$$

and replace (1) and (13) by

$$\bar{x}(k+1) = z(k) - 2 \ , \tag{15}$$

$$\psi(k+1) = - \partial_z f(\bar{x}(k),z(k)) \ , \tag{16}$$

and i)' by

$$\psi(k) = \psi(k+1) + \partial_x f(\bar{x}(k),z(k)) + \partial_z f(\bar{x}(k),z(k)) \ ,$$

which (using (16)) reduces to

$$\psi(k) = \partial_x f(\bar{x}(k),z(k)) \ . \tag{17}$$

We can solve for z(k) in (17). In fact, the equation

$$\partial_x f(x,z) = \psi \ , \quad z > 0 \tag{18}$$

in z, is equivalent to

$$(\partial_x g(x,z))^2 = \psi^2(1 - g^2(x,z)) \ , \ \operatorname{sgn} \psi = \operatorname{sgn} \partial_x g(x,z) \ , \ z > 0 \tag{19}$$

or

$$(4 - z^2 + x^2)^2 = x^2\psi^2(4x^2z^2 - (x^2 + z^2 - 4)^2) \ . \tag{20}$$

With the substitution $z^2 - 4 =: \zeta$, equation (20) is equivalent to

$$(x^2 - \zeta)^2 = x^2\psi^2(16x^2 - (x^2 - \zeta)^2)$$

which is a linear equation for $(x^2 - \zeta)^2$. Hence,

$$(x^2 - \zeta)^2 = 16x^4\psi^2(1 + x^2\psi^2)^{-1} \ .$$

This yields

$$z^2 = x^2 + 4 \pm 4x^2\psi(1 + x^2\psi^2)^{-\frac{1}{2}} \ .$$

Since $\operatorname{sgn} \partial_x g(x,z) = \operatorname{sgn}(4 - z^2 + x^2)$ for positive z, we see by (19) that (18) is equivalent to:

$$z = [x^2 + 4 - 4x^2\psi(1 + x^2\psi)^{-\frac{1}{2}}]^{\frac{1}{2}} \ . \tag{21}$$

Therefore we can restate the boundary value problem as follows:

"Given numbers a,b, satisfying $b > 2$, $a - 4N < b < a$, find a real number θ such that, if the functions $\psi \in (\widehat{N+1} \to R^1)$, $\bar{x} \in (\widehat{N+1} \to R^1)$, $z \in (\hat{N} \to R^1)$ are defined by $\psi(0) = \theta$, $\bar{x}(0) = a$ and

$$z(k) = [\bar{x}^2(k) + 4 - 4\bar{x}^2(k)\psi(k)(1+\bar{x}^2(k)\psi(k))^{-\frac{1}{2}}]^{\frac{1}{2}} \ ,$$

$$\bar{x}(k+1) = z(k) - 2 \ ,$$

$$\psi(k+1) = (z(k))^{-1}(\bar{x}^2(k)-z(k)-4)[16z^2(k) - (\bar{x}^2(k)-z^2(k)-4)^2]^{-\frac{1}{2}}$$

for $k \in \hat{N}$, then we have $\bar{x}(N) = b$."

REFERENCES

[1] *E.B. Lee* and *L. Markus*, "Foundations of optimal control theory", Wiley, New York (1967).

[2] *L.S. Pontryagin, V.G. Boltyanskii, R.V. Gamkrelidze* and *E.F. Mishchenko*, "The mathematical theory of optimal processes", Wiley, New York (1962).

[3] *H. Halkin*, "A maximum principle of the Pontryagin type for systems, described by non-linear difference equations", J. Siam Control 4, 90-111 (1966).

[4] *R. Bellman* and *R. Kalaba*, "Dynamic programming and modern control theory", Academic Press, New York (1965).

[5] *V.V. Velichenko*, "Optimal control problems for equations with discontinuous right-hand side", Avtomatika i Telemekhanika 27, 20-30 (1965).

[6] *C. Caratheodory*, "Vorlesungen über reelle Funktionen", Leipzig, 1927, reprinted: New York (1948).

[7] *E.J. McShane*, "Integration", Princeton University Press (1944).

[8] *L.W. Neustadt*, "The existence of optimal controls in the absence of convexity conditions", J. Math. Anal. Appl. 7, 110-117 (1963).

[9] *K. Magnus*, "Schwingungen", B.G. Teubner Verlagsgesellschaft (1961).

[10] *G. Owen*, "Game theory", W.B. Saunders company (1968).

[11] *H. Hermes* and *J.P. La Salle*, "Functional analysis and time optimal control", Academic Press, New York (1969).

[12] *N.G. de Bruijn*, "A problem of optimal control", J. Math. Anal. Appl. 14, 185-190 (1966).

[13] *N.G. de Bruijn*, "Asymptotic methods in analysis", North-Holland Publishing Company (1961).

[14] *M. Canon, C. Cullum* and *E. Polak*, "Constrained minimization problems in finite dimensional spaces", J. Siam Control 4, 528-547 (1966).

[15] *W.A. Coppel*, "Stability and asymptotic behaviour of differential equations", Heathand Company (1965).

*

S A M E N V A T T I N G

In dit proefschrift wordt een methode behandeld voor de bepaling van optimale besturingen (optimal controls) voor systemen die beschreven worden door differentiaalvergelijkingen van de vorm $\dot{x} = f(x,u)$, waarbij de functie f discontinuïteiten heeft voor sommige waarden van x. Hierbij is u de besturingsfunctie en t de onafhankelijke variabele. Als er geen discontinuïteiten aanwezig zijn kunnen zulke optimale besturingen worden gevonden met behulp van het maximumprincipe van Pontryagin (zie [2]). Als f discontinuïteiten heeft is het maximumprincipe niet meer geldig. De methode gevolgd in dit proefschrift kan als volgt worden omschreven:

We verdelen de banen (dit zijn de oplossingskrommen van de differentiaalvergelijking) in stukken waarop f wel continu is. Op deze stukken passen we het maximumprincipe toe. Een baan die optimaal is op elk van die stukken noemen we *stuksgewijs optimaal*. Om nu een besturing te vinden die optimaal is voor de hele kromme moeten we de "stukken" aan elkaar passen. Dit resulteert in een discreet optimaliseringsprobleem. Een probleem van dit type kan worden opgelost door middel van dynamische programmering of met behulp van het discrete maximumprincipe, een discrete versie van het maximumprincipe van Pontryagin.

In dit proefschrift wordt geen algemene theorie behandeld. De methode wordt toegepast op enkele voorbeelden, die in detail worden uitgewerkt. Onder optimale besturing wordt steeds tijdoptimale besturing verstaan, hoewel de methode ook kan worden gebruikt bij algemenere optimaliteitscriteria.

Omdat van differentiaalvergelijkingen met discontinu rechterlid de existentie van oplossingen niet uit klassieke stellingen hierover kan worden afgeleid, wordt in hoofdstuk I een criterium voor de existentie van oplossingen van zulke differentiaalvergelijkingen gegeven.

In hoofdstuk II wordt een eenvoudige vergelijking besproken (de zogenaamde jojo-vergelijking) en er worden een aantal optimale besturingsproblemen opgelost voor deze vergelijking. De "discrete gedeelten" van deze problemen worden opgelost door middel van dynamische programmering.

In hoofdstuk III wordt het discrete maximumprincipe gebruikt om de discrete gedeelten van de problemen van hoofdstuk II op te lossen. Verder wordt in hoofdstuk III de optimale besturing van een mechanisch systeem met coulomb-wrijving besproken.

CURRICULUM VITAE

De schrijver van dit proefschrift werd op 27 april 1940 geboren in Helden. Van 1953 tot 1959 doorliep hij het gymnasium op het Bisschoppelijk College te Roermond. Daarna studeerde hij tot 1966 voor wiskundig ingenieur aan de Technische Hogeschool te Eindhoven. Van 1963 tot 1966 was hij werkzaam als student-assistent bij Prof.dr. N.G. de Bruijn. Sinds zijn afstuderen is hij als wetenschappelijk medewerker werkzaam bij de onderafdeling der wiskunde van de Technische Hogeschool te Eindhoven.

S T E L L I N G E N

I

Een rij reële getallen $\{a_n\}_0^\infty$ heeft de eigenschap dat $\Sigma\, a_n x_n^2 > 0$ geldt voor elke niet-triviale eindige rij $\{x_n\}_0^N$ met $\Sigma\, x_n = 0$, dan en slechts dan als voldaan is aan de volgende voorwaarden:

i) Hoogstens één a_n is niet-positief.

ii) Als er een $a_n < 0$ is, dan is $\Sigma\, a_n^{-1}$ convergent en niet-positief.

II

De kubus $\{x \in R^n \mid |x_i| \leq 1\ (i = 1,\ldots,n)\}$ en het hypervlak $\{x \in R^n \mid (a,x) = c\}$ met $a \in R^n$, $a \neq 0$, hebben een niet-lege doorsnede D als $\Sigma\, |a_i| \geq |c|$. Het punt $y \in D$ met de kleinste (euclidische) afstand tot de oorsprong wordt beschreven door $y_i = \mathrm{sat}(\lambda a_i)$ $(i = 1,\ldots,n)$, waarbij sat de oneven functie is met $\mathrm{sat}\,\alpha = \min(1,\alpha)$ $(\alpha \geq 0)$. De waarde van λ kan op een eenvoudige grafische manier worden bepaald uit de vergelijking $(a,y) = c$.

III

Een cel in het platte vlak is een 1×1-vierkant met hoekpunten in roosterpunten (d.w.z. punten (i,j) met i,j geheel). Een kleuring is een afbeelding van de collectie der cellen in een verzameling S (de verzameling der kleuren). Een kleuring heet $n \times m$-invariant als voor elke kleur $s \in S$ het volgende geldt:

Het aantal cellen met de kleur s in een $n \times m$-rechthoek van cellen is onafhankelijk van de positie van de rechthoek.

Er geldt de volgende eigenschap: Als n en m onderling ondeelbaar zijn dan is een $n \times m$-invariante kleuring een $n \times 1$- of een $m \times 1$-invariante kleuring.

<u>Literatuur</u>: M.L.J. Hautus and D.A. Klarner, "Invariant colorings for boxes" (in voorbereiding).

IV

Zij X een genormeerde lineaire ruimte en F een begrensde afbeelding van X in X (d.w.z. F(M) is begrensd als $M \subset X$ begrensd is). Dan bestaat er een rij $\{\lambda_1,\lambda_2,\ldots\}$ van positieve getallen zodanig dat elke rij $\{x_n\}$ in X die voldoet aan $x_{n+1} = F(\lambda_n x_n)$ $(n = 1,2,\ldots)$, begrensd is.

V

De oplossing van de differentiaalvergelijking $y'(x) = -2y^{\frac{1}{2}}(x)$ $(x > 0)$, met beginwaarde $y(0) = 1$ wordt benaderd door het Euler-Cauchy polygoon waarvan de hoekpunten gegeven worden door $x_n = nh$ $(n = 0,1,\ldots)$, $y_0 = 1$, $y_{n+1} = y_n - 2h\, y_n^{\frac{1}{2}}$ $(n = 0,1,\ldots)$. Hierbij is h een positief getal. Laat ξ de waarde van x zijn, waarvoor het polygoon de x-as snijdt. Dan geldt:

$$\xi = 1 + \tfrac{1}{2}\, h \log h + O(h) \qquad (h \to 0)\ .$$

VI

Laten $\phi(x,y)$ en $f(x,y)$ analytische functies zijn in een omgeving van de oorsprong (0,0) in $\mathbb{C}^1 \times \mathbb{C}^n$ met waarden in $\mathbb{C}^1$ resp. $\mathbb{C}^n$. Laat de formele reeks $u(x) = \sum_{k=1}^{\infty} a_k x^k$ formeel voldoen aan de differentiaalvergelijking

$$\phi(x,y)y'(x) = f(x,y)\ . \tag{1}$$

Als $\phi(x,u(x)) =: \sum_{0}^{\infty} b_k x^k$, dan wordt ω gedefinieerd door $\omega := \min\{k \mid b_k \neq 0\}$. (Als $b_k = 0$ voor alle k geldt, dan is $\omega = \infty$.) Er gelden dan de volgende eigenschappen:

i) Als $\omega = 0$ of $\omega = 1$, dan heeft u een positieve convergentiestraal en is derhalve een oplossing van (1) in een omgeving van 0.

ii) Als $\omega < \infty$, dan bestaat er een $\theta > 0$, zodat voor elke sector S in het complexe x-vlak met openingshoek kleiner dan θ een oplossing van (1) bestaat, waarvoor u(x) de asymptotische reeks is voor $x \to 0$, $x \in S$.

Literatuur: W.A. Harris, "Holomorphic solutions of non-linear differential equations at singular points". Adv. in Diff. and Int. Eq. SIAM, Philadelphia, 1969.

J. Malmquist, "Sur l'étude analytique des solutions d'un système d'équations différentielles dans le voisinage d'un point singulier d'indétermination II". Acta Math. 74 (1941), 1-64.

VII

Het bewijs van "Theorem 19.1" in het boek van W. Wasow, "Asymptotic expansions for ordinary differential equations", (Wiley, New York, 1965, p. 111) is niet juist. Het is mogelijk door middel van een modificatie dit bewijs te corrigeren.

VIII

Zij A een $n \times n$-matrix en B een $n \times m$-matrix. Dan heet het paar (A,B) bestuurbaar als de $n \times nm$-matrix $[B, AB, \ldots, A^{n-1}B]$ de rang n heeft. Er geldt de volgende eigenschap: (A,B) is bestuurbaar dan en slechts dan als voor elke eigenwaarde λ van A geldt:
$\mathrm{rang}[A - \lambda I, B] = n$.

Literatuur: M.L.J. Hautus, "Controllability and observability conditions of linear autonomous systems". Nederl. Akad. Wetensch., Proc., Ser. A 72, (1969), 443-448.

IX

Zij S een niet-lege verzameling van complexe getallen, A een $n \times n$-matrix and B een $n \times m$-matrix. Dan bestaat er een matrix D zodanig dat de eigenwaarden van A + BD in S liggen, dan en slechts dan als $\mathrm{rang}[A-\lambda I, B] = n$ voor elke eigenwaarde λ van A buiten S geldt.

Literatuur: M.L.J. Hautus, "Controllability, observability and stabilizability of linear autonomous systems" (in voorbereiding).

X

De voorwaarde, die door R.E. Kalman noodzakelijk en voldoende wordt genoemd voor het bestaan van een "asymptotic state estimator" voor een lineair constant systeem met één output-variable, is wel voldoende maar niet noodzakelijk.

Literatuur: R.E. Kalman, P.L. Falb, M.A. Arbib, "Topics in mathematical system theory". McGraw-Hill, New York, 1969, p. 56.

XI

Zij Ω_n de collectie der niet-lege compacte deelverzamelingen van R^n. Op Ω_n wordt een afstand gedefinieerd door

$$\rho(S,T) := \max\{\max_{x\in S} d(x,T), \max_{y\in T} d(y,S)\} .$$

Hierbij is

$$d(x,A) := \min_{x\in A} |x - a| , \quad (x \in R^n, A \in \Omega_n) .$$

Laat V een absoluut continue afbeelding van $[0,1]$ in Ω_n zijn (d.w.z. voor alle $\varepsilon > 0$ bestaat er een $\delta > 0$, zodat voor alle rijen $s_1,t_1,s_2,t_2,\ldots,s_n,t_n$ met $0 \le s_1 < t_1 \le \ldots \le s_n < t_n \le 1$ met $\Sigma(t_i - s_i) < \delta$ geldt $\Sigma\, d(V(t_i),V(s_i)) < \varepsilon$). Dan bestaat er een absoluut continue afbeelding v van $[0,1]$ in R^n met $v(t) \in V(t)$ $(0 \le t \le 1)$.

Literatuur: H. Hermes, "Existence and properties of solutions of $\dot{x} \in R(t,x)$", Adv. in Diff. and Int. Eq. SIAM, Philadelphia, 1969.

Eindhoven, 29 juni 1970

M.L.J. Hautus