ON THE "BANG-BANG" CONTROL PROBLEM*

BY

R. BELLMAN, I. GLICKSBERG AND O. GROSS

The RAND Corporation, Santa Monica, Calif.

Summary. Let S be a physical system whose state at any time is described by an n-dimensional vector $x(t)$, where $x(t)$ is determined by a linear differential equation $dz/dt = Az$, with A a constant matrix. Application of external influences will yield an inhomogeneous equation, $dz/dt = Az + f$, where f, the "forcing term", represents the control. A problem of some importance in the theory of control circuits is that of choosing f so as to reduce z to 0 in minimum time. If f is restricted to belong to the class of vectors whose ith components can assume only the values $\pm b_i$, the control is said to be of the "bang-bang" type.

Various aspects of the above problem have been treated by McDonald, Bushaw, LaSalle and Rose. We shall consider here the case where all the solutions of $dz/dt = Az$ approach zero as $t \to \infty$. In this case we prove that the problem of determining f so as to minimize the time required to transform the system into the rest position subject to the requirement that f_1, the ith component, satisfies the constraint $| f_i | \leq b_i$ may be reduced to the case where $f_i = \pm b_i$. Furthermore, we show that if all the characteristic roots of A are real and negative, f_i need change value only a finite number of times at most, dependent upon the dimension of the system.

Finally, an example is given for $n = 2$, illustrating the procedure that can be followed and the results that can be obtained.

1. Introduction. Let z be an n-dimensional vector function of t satisfying the linear differential equation

$$\frac{dz}{dt} = Az + f, \qquad z(0) = c, \tag{1.1}$$

where we assume that:

a. A is a real, constant matrix of order n, whose characteristic roots all have negative real parts;

b. f is restricted to be real, measurable, and to have components satisfying the constraints, $| f_i | \leq 1$.

The first condition is the necessary and sufficient condition that all the solutions of (1.1) approach zero as $t \to \infty$.

The problem we wish to consider is that of determining the vectors f which, subject to the constraint (b), reduce z to zero in minimum time. This is a problem of Bolza of rather unconventional type, and the techniques we shall employ are quite different from the classical ones.

We shall establish two results:

THEOREM 1. *Under the above conditions, an f which reduces z to zero in minimum time exists, and has components f_i for which $| f_i | = 1$.*

*Received December 10, 1954; revised manuscript received March 14, 1955.

THEOREM 2. *If the characteristic roots of A are real, distinct, and negative, a minimizing f exists with components f_i for which $| f_i | = 1$, and each f_i changes sign at most $(n - 1)$ times.*

The statement in Theorem 1 has been assumed in the past on an intuitive basis, see McDonald, [3], and has been established in various cases by Bushaw [1], LaSalle [2], and Rose [4]. The only paper we have had access to is that by Rose, and his methods are distinct from ours. In addition, he is primarily interested in the case where the condition in (a) is not satisfied.

Problems of this type arise in connection with many different types of control processes. A discussion of the connection with servomechanisms is sketched in [2].

2. Proof of Theorem 2. We shall consider in detail only the case of Theorem 2, where the characteristic roots of A are real and negative. It will be clear from the treatment of this case how the proof of Theorem 1 goes.

Let X be a square matrix whose columns are the n linearly independent eigenvectors x_i of A, and let λ_j $(j = 1, \cdots, n)$ be the corresponding n distinct, negative eigenvalues of A; clearly, X is non-singular and all its elements are real. Finally, denote by Λ the diagonal matrix whose jth diagonal element is λ_j. We have

$$Ax_j = \lambda_j x_j , \tag{2.1}$$

whence we see that $AX = X\Lambda$; hence

$$X^{-1}AX = \Lambda. \tag{2.2}$$

If now in (2.1) we make the transformation $z = Xy$, we obtain using (2.2),

$$\begin{aligned} y'(0) &= X^{-1}c, \\ y'(t) &= \Lambda y(t) + X^{-1}f(t), \end{aligned} \tag{2.3}$$

or, componentwise,

$$y_i'(t) = \lambda_i y_i(t) + \sum_{j=1}^{n} \alpha_{ij} f_j(t), \tag{2.4}$$

where the α's are the elements of X^{-1}. Solving for $y_i(t)$, we obtain

$$y_i(t) = y_i(0) \exp(\lambda_i t) + \exp(\lambda_i t) \int_0^t \exp(-\lambda_i s) \sum_{j=1}^{n} \alpha_{ij} f_j(s)\, ds. \tag{2.5}$$

Since $z(t) = 0$ is equivalent to $y(t) = 0$, we wish to find the least t for which, for some f, $y_i(t) = 0$, $i = 1, \cdots, n$, i.e., for which

$$-y_i(0) = \int_0^t \exp(-\lambda_i s) \sum_{j=1}^{n} \alpha_{ij} f_j(s)\, ds, \qquad i = 1, \cdots, n \tag{2.6}$$

for some f.

Our first observation is that, given any starting value $y(0) \neq 0$ there exists a $t > 0$ and an f, such that (2.6) is satisfied. In fact, there is a constant vector $f(s) = k$ which does the trick for some t sufficiently large. For substituting $f_j(s) = k_j$ in (2.6), we obtain

$$\sum_{j=1}^{n} \alpha_{ij} k_j = \frac{y_i(0)}{-\int_0^t \exp(-\lambda_i s)\, ds} = \frac{\lambda_i y_i(0)}{\exp(-\lambda_i t) - 1},$$

whence, by virtue of the definition of the α's,

$$k_j = \sum_i x_{ji} \frac{\lambda_i y_i(0)}{\exp(-\lambda_i t) - 1} . \tag{2.7}$$

Since $-\lambda_i > 0$ the right member of (2.7) can be made as small in magnitude as we please for sufficiently large t, and hence we can insure that $|k_j| \leq 1$.

For each $t \geq 0$ we have a linear mapping ρ_t taking f into the n-dimensional vector with ith component

$$\int_0^t \exp(-\lambda_i s) \sum_j \alpha_{ij} f_j(s)\, ds, \tag{2.8}$$

and this mapping clearly takes our basic convex set of f's onto a convex subset $C(t)$ of euclidean n-space. For any f in our basic set there is another, $\bar{f}$, in the set which agrees with f for $s \leq t$ and vanishes for $s > t$, so that, for $t' > t$, $\rho_t \cdot f = \rho_t \bar{f} = \rho_t f$, by (2.8), and $\rho_t f$ is in $C(t')$. Thus $C(t)$ increases with t.

Now our desired least time is, by (2.6), the least $t \geq 0$ for which $C(t)$ contains the vector $-y(0)$. Since $C(t)$ increases, we have an interval (t_0, ∞) for which $C(t)$ contains this vector, while for $t < t_0$ this is not the case. We can see that $C(t_0)$ also contains this vector as follows.

Denoting for any vector $x = (x_1, x_2, \cdots, x_n)$ the euclidean norm $(\sum_i x_i^2)^{1/2}$ by $\| x \|$, we obtain, using (2.8), a constant $k = k(t_0)$ with the property that for all f and t, t' in a finite interval $[0, t_0]$ we have $\| \rho_t f - \rho_{t'} f \| \leq k | t - t' |$; thus, for $| t - t' |$ small every point of $C(t')$ is close to a point of $C(t)$. Since $-y(0)$ is in $C(t)$ for all $t > t_0$, $-y(0)$ must be at zero distance from $C(t_0)$ so that if we show this set is closed $-y(0)$ must actually be in it. But each $C(t)$ is closed, since by a well known fact about Banach spaces [5], our basic set of f's may be topologized so as to be compact and render each ρ_t continuous. Thus $C(t)$, as the continuous image of a compact set, is compact, hence closed.

Let us return to the fact that $-y(0)$ is not in $C(t)$ for $t < t_0$. From the theory of convex sets [6] this implies that we have a vector θ^t of unit norm, for which, in the usual inner product notation, $(\theta^t, \rho_t f) \leq [\theta^t, -y(0)]$ for every f. Since the vectors of unit norm are compact in the euclidean topology, we may select a sequence t_n increasing to t_0 for which θ^{t_n} converges to some vector θ of unit norm. But since $\rho_{t_n} f$ converges to $\rho_{t_0} f$, $(\theta, \rho_{t_0} f) = \lim (\theta^{t_n}, \rho_{t_n} f) \leq \lim [\theta^{t_n}, -y(0)] = [\theta, -y(0)]$. Thus if f^* denotes an f for which $\rho_{t_0} f^* = -y(0)$ we have $(\theta, \rho_{t_0} f) \leq (\theta, \rho_{t_0} f^*)$ for all f, hence constants $\theta_1, \cdots, \theta_n$, not all zero for which f^* maximizes the expression

$$\sum_i \theta_i \int_0^{t^*} \exp(-\lambda_i s) \sum_j \alpha_{ij} f_j(s)\, ds = \sum_j \int_0^{t^*} \Big(\sum_i \theta_i \alpha_{ij} \exp[-\lambda_i s]\Big) f_j(s)\, ds. \tag{2.9}$$

But this expression clearly has as its maximum

$$\sum_j \int_0^{t^*} \Big| \sum_i \theta_i \alpha_{ij} \exp(-\lambda_i s) \Big|\, ds \tag{2.10}$$

achieved by setting $f_j(s) = \operatorname{sgn}(\sum_i \theta_i \alpha_{ij} \exp[-\lambda_i s])$. Thus it is clear that $f_j^*(s) = \operatorname{sgn}(\sum_i \theta_i \alpha_{ij} \exp[-\lambda_i s])$ almost everywhere on the set where $\sum_i \theta_i \alpha_{ij} \exp(-\lambda_i s) \not\equiv 0$.

Our principal result now follows, namely that we can achieve minimal time by restricting f to assume componentwise ± 1 on a finite number of intervals; in fact, in

the case considered, each component need change sign *at most* $n - 1$ times. This latter statement is a simple consequence of the fact that unless the continuous function ϕ_i given by $\phi_i(s) = \sum_{i=1}^{n} \theta_i \alpha_{ij} \exp(-\lambda_i s)$ is identically zero (in which case it makes no difference as to our choice of f_i^*), it can have at most $n - 1$ real zeros. This is well known and there is a simple inductive proof.

3. A special case of $n = 2$. Consider the problem as before, with

$$A = \begin{pmatrix} -3 & -2 \\ 1 & 0 \end{pmatrix};$$

thus,

$$z_1' = -3z_1 - 2z_2 + f_1, \qquad z_2' = z_1 + f_2. \tag{3.1}$$

The transformation

$$z_1 = 2y_1 - y_2, \qquad z_2 = -y_1 + y_2 \tag{3.2}$$

reduces the above system to

$$y_1' = -2y_1 + f_1 + f_2, \qquad y_2' = -y_2 + f_1 + 2f_2, \tag{3.3}$$

and we obtain, as before, for the set of admissible starting values, for a given t and f_1, f_2,

$$\begin{aligned} -y_1(0) &= \int_0^t e^{2s}[f_1(s) + f_2(s)]\, ds, \\ -y_2(0) &= \int_0^t e^{s}[f_1(s) + 2f_2(s)]\, ds. \end{aligned} \tag{3.4}$$

From the preceding section, we know that if t^* is minimal, then the optimal f^* is given by

$$\begin{aligned} f_1(s) &= \operatorname{sgn}(\theta_1 e^{2s} + \theta_2 e^{s}), \\ f_2(s) &= \operatorname{sgn}(\theta_1 e^{2s} + 2\theta_2 e^{s}). \end{aligned} \tag{3.5}$$

If we now ask the question "For what set of starting values y is it optimal to choose $f_1 = 1$, $f_2 = 1$ on an *initial interval*?" with a similar question for the other combinations ± 1, it is readily seen that the answers will determine an *optimal policy*. This is clear, since any continuation of an optimal policy must be again optimal with respect to the new starting values. We thus have

$$\begin{aligned} -y_1(0) &= \int_0^{t^*} e^{2s}\{\operatorname{sgn}(\theta_1 e^{2s} + \theta_2 e^{s}) + \operatorname{sgn}(\theta_1 e^{2s} + 2\theta_2 e^{s})\}\, ds, \\ -y_2(0) &= \int_0^{t^*} e^{s}\{\operatorname{sgn}(\theta_1 e^{2s} + \theta_2 e^{s}) + 2\operatorname{sgn}(\theta_1 e^{2s} + 2\theta_2 e^{s})\}\, ds. \end{aligned} \tag{3.6}$$

To answer the first question, for what values of y is it optimal to set $f_1 = f_2 = 1$ on an initial interval, we note that this is equivalent to the conditions

$$\begin{aligned} t^* &> 0, \\ \theta_1 + \theta_2 &> 0, \\ \theta_1 + 2\theta_2 &> 0. \end{aligned} \tag{3.7}$$

Now, since the functions $\theta_1 e^{2t} + \theta_2 e^t$, $\theta_1 e^{2t} + 2\theta_2 e^t$ can each vanish at most once, we see that the above case breaks down into four sub-cases, namely:

$$\text{(a)}\quad \theta_1 \exp(2t^*) + \theta_2 \exp(t^*) > 0, \qquad \text{(b)}\quad >, <, \qquad \text{(c)}\quad <, >, \qquad \text{(d)}\quad <, < \tag{3.8}$$
$$\theta_1 \exp(2t^*) + 2\theta_2 \exp(t^*) > 0.$$

Case (a) is trivial and consists of the arc α' illustrated in Fig. 1. α' is defined parametrically by

$$\begin{aligned} y_1(0) &= 1 - \exp(2t^*) \\ & \qquad\qquad t^* > 0 \\ y_2(0) &= 3(1 - \exp[t^*]), \end{aligned} \tag{3.9}$$

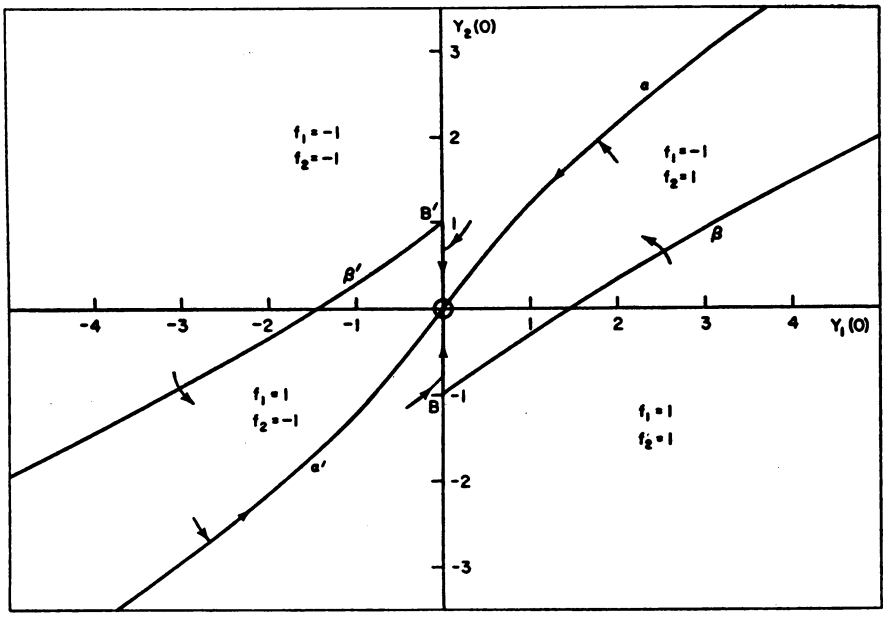

FIG. 1.

as one can readily verify by working out the integrals. Moreover, the curve defines an optimal path, since the solution of the differential equation is, with $f_1 = f_2 = 1$ identically, and $y_1(0)$, $y_2(0)$ defined as above, precisely a sub-arc of α' beginning at $y(0)$ and terminating at the origin.

Case (b) is vacuous, for if we have

$$\begin{aligned} &\theta_1 \exp(2t^*) + \theta_2 \exp(t^*) > 0 \qquad \text{and} \\ &\theta_1 \exp(2t^*) + 2\theta_2 \exp(t^*) < 0, \end{aligned} \tag{3.10}$$

we obtain, by subtraction, and the condition $t^* > 0$, that $\theta_2 < 0$. But, $\theta_1 + \theta_2 > 0$, whence $\theta_1 > 0$. We thus have

$$\begin{aligned} \theta_1 \exp(2t^*) + 2\theta_2 \exp(t^*) &> \theta_1 \exp(t^*) + 2\theta_2 \exp(t^*) \\ &= \exp(t^*)(\theta_1 + 2\theta_2) > 0 \end{aligned} \tag{3.11}$$

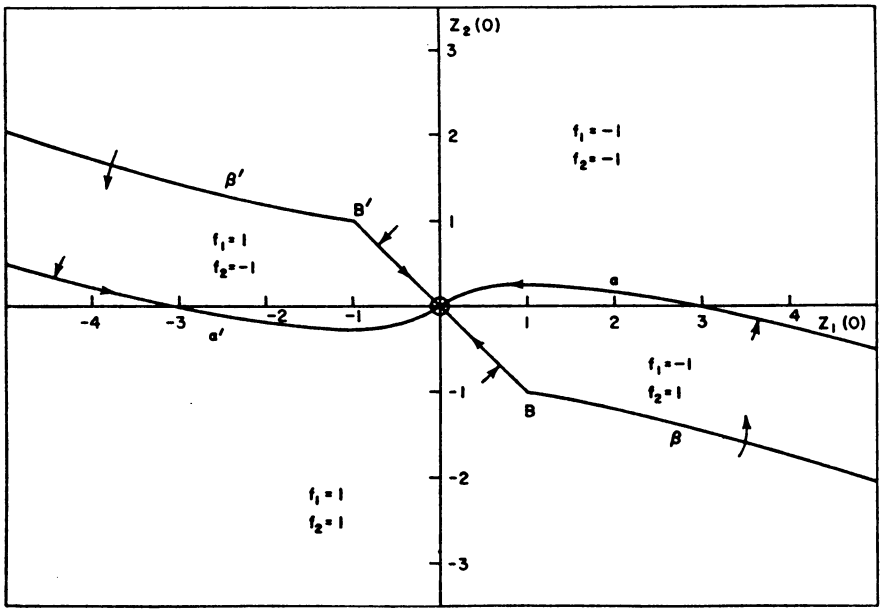

FIG. 2.

which contradicts

$$\theta_1 \exp(2t^*) + 2\theta_2 \exp(t^*) < 0. \tag{3.12}$$

We shall treat case (3.8c) in detail. Case (3.8d) can be treated similarly, but is a trifle more involved, albeit elementary, and will be omitted on those grounds.

We have, upon substituting in (3.6) for case (c):

$$\begin{aligned} -y_1(0) &= \int_0^{\ln(-\theta_2/\theta_1)} \exp(2s)\,ds - \int_{\ln(-\theta_2/\theta_1)}^{t^*} \exp(2s)\,ds + \int_0^{t^*} \exp(2s)\,ds, \\ -y_2(0) &= \int_0^{\ln(-\theta_2/\theta_1)} \exp(s)\,ds - \int_{\ln(-\theta_2/\theta_1)}^{t^*} \exp(s)\,ds + 2\int_0^{t^*} \exp(s)\,ds. \end{aligned} \tag{3.13}$$

Simplifying, we obtain

$$\begin{aligned} -y_1(0) &= (\theta_2/\theta_1)^2 - 1, \\ -y_2(0) &= -2(\theta_2/\theta_1) + \exp(t^*) - 3. \end{aligned} \tag{3.14}$$

If now we set $x^* = \exp(t^*)$, our conditions become

$$\begin{aligned} x^* &> 1, \\ \theta_1 + \theta_2 &> 0, \\ \theta_1 + 2\theta_2 &> 0, \\ \theta_1 x^* + \theta_2 &< 0, \\ \theta_1 x^* + 2\theta_2 &> 0, \\ y_1(0) &= 1 - (\theta_2/\theta_1)^2, \\ y_2(0) &= 3 + 2(\theta_2/\theta_1) - x^*. \end{aligned} \tag{3.15}$$

We easily obtain from the above that $\theta_1 < 0$. By homogeneity, we can set $\theta_1 = -1$, $\theta_2 = \lambda$ and we obtain the equivalent conditions

$$2\lambda > x^* > \lambda > 1 \qquad \text{(A)}$$

$$y_1(0) = 1 - \lambda^2$$

$$\text{(B)}$$

$$y_2(0) = 3 - 2\lambda - x^*$$

i.e., we wish to find the image of all pairs (x^*, λ) satisfying (A) under the mapping defined by (B). Pictorially this is represented by Fig. 3.

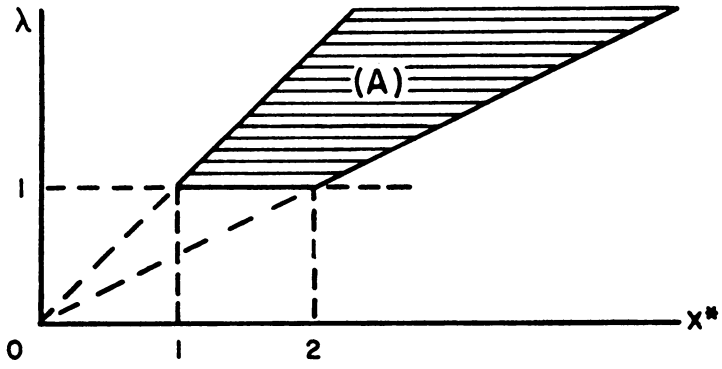

FIG. 3.

On the other hand, the Jacobian of the transformation (B) is given by

$$J_B = \begin{vmatrix} -2\lambda & 0 \\ -2 & -1 \end{vmatrix} = 2\lambda \neq 0 \tag{3.16}$$

throughout (A); hence the transformation is non-singular and the boundary of the image is the image of the boundary. Making use of this fact we obtain the region for case (3.8c):

$$y_1(0) < 0 \tag{3.17}$$

and

$$3 - 4[1 - y_1(0)]^{1/2} < y_2(0) < 3 - 3[1 - y_1(0)]^{1/2}. \tag{3.18}$$

In a similar manner we obtain a region for case (3.8d). The union of cases (3.8a) through (3.8d) is the set of all starting values for which $f_1 = f_2 = 1$ is optimal on an initial interval. In a similar manner we obtain the region $f_1 = 1$, $f_2 = -1$. (Notice that we need not compute the other regions since they can be obtained by skew-symmetry.)

The final result of our calculations is illustrated in Figs. 1 and 2. Figure 2 is the image of Fig. 1 under our initial transformation and gives the optimal policy in terms of our initial starting vector $c = [z_1(0), z_2(0)]$.

In terms of optimal paths (see Fig. 2) we can state the following: A path initiating in the (1,1) region continues with $f_1 = 1$, $f_2 = 1$ until it strikes either the straight segment OB or the parabolic arc β. In the former case f_1 switches from 1 to -1 and the

path continues along OB to the origin. In the latter case f_1 switches to -1 at β and the path continues in the $(-1, 1)$ region until it intercepts the parabolic arc α at which f_2 changes from 1 to -1 and α is followed to the origin with $f_1 = f_2 = -1$. Similar remarks hold for the skew-symmetric regions.

Bibliography

1. D. W. Bushaw, Ph.D. Thesis, Department of Mathematics, Princeton University, 1952
2. J. P. LaSalle, Abstract 247t, Bull. Amer. Math. Soc. **60**, 154 (1954)
3. D. McDonald, *Nonlinear techniques for improving servo performance*, Cook Research Laboratories, Bulletin S-2, Chicago, 1950
4. N. J. Rose, *Theoretical aspects of limit control*, Report No. 459, Experimental Towing Tank, Stevens Institute of Technology, November 1953
5. L. Alaoglu, *Weak topologies of normed linear spaces*, Ann. of Math. **41**, 252-267 (1940)
6. T. Bonnesen and W. Fenchel, *Theorie der konvexen Körper*, Ergebnisse der Mathematik **3**, 1, Berlin 1934, New York 1948