A MINIMUM-MAXIMUM PROBLEM FOR DIFFERENTIAL EXPRESSIONS

D. S. CARTER

1. Introduction. In the study of approximate methods for solving ordinary differential equations, an interesting question arises. To state it roughly for a single first order expression, let $y_0(t)$ be the solution of the equation

$$f(t, y, \dot{y}) = 0 \tag{1.1}$$

which satisfies the initial condition $y(a) = \eta_a$. Let η_b be an *approximation* to the value of y_0 at a later time, $t = b$. Unless this approximation is exact, there is no continuous function which satisfies (1.1), together with the two boundary conditions

$$y(a) = \eta_a, \; y(b) = \eta_b. \tag{1.2}$$

The question is whether there exists a continuous function satisfying (1.2), for which the maximum absolute value of $f(t, y, \dot{y})$ on the interval $[a, b]$ is minimized; and if so, how to find it.

Stated for higher order and multiple-component systems, this problem should find application in engineering and some branches of "operations analysis." For example, in a dynamical system being driven between two preassigned points in phase-space, it may be of interest to minimize the peak value of a stress which is expressible, through accelerations or friction, as the absolute value of a differential form.

A moment's reflection provides some insight into the nature of the solution. If there is a solution $y^*(t)$, then $f(t, y^*, \dot{y}^*)$ must be constant in absolute value, so that $|f|$ is everywhere equal to its supremum. Otherwise it would be possible to decrease $|f|$ near its maxima at the expense of increases elsewhere. This does not mean that f *itself* need be constant. In fact, it turns out that the value of f at the solution will generally have a number of discontinuities of sign, especially for higher order expressions.

For the first order system (1.1), (1.2), this situation is illustrated by the following rough variational argument. Let $y(t)$ satisfy (1.2), and let the partial derivative $f_{\dot{y}}(t, y, \dot{y})$ be constant in sign throughout $[a, b]$. The first variation of f corresponding to a variation δy is

$$\delta f = f_y \, \delta y + f_{\dot{y}} \, \delta \dot{y}.$$

Received January 6, 1956. Work performed under the auspices of the U.S. Atomic Energy Commission.

The author wishes to express his indebtedness to J. Lehner and S. Ulam of the Los Alamos Scientific Laboratory for helpful discussions of this problem.

Solving for δy, and using the fact that $\delta y = 0$ at $t = a, b$, we have

$$\int_a^b ds[\delta f(s)/f_{\dot{y}}(s)] \exp\left\{ \int_b^s dr\, f_y(r)/f_{\dot{y}}(r) \right\} = 0,$$

which is a necessary and sufficient condition for the admissibility of δf. Since neither $f_{\dot{y}}$ nor the exponential factor changes sign, this requires little more than that δf shall change sign. Unless f is constant throughout $[a, b]$ it is easy to construct an admissible δf differing in sign everywhere from f, so $\sup|f + \delta f| < \sup|f|$. Conversely, if f is everywhere constant, then δf must agree in sign with f on part of the interval; and $\sup|f + \delta f| > \sup|f|$ for every non-vanishing δy. That is to say, $\sup|f|$ has at least a *local* minimum at $y = y^*$ if and only if y^* satisfies (1.2) and

$$f(t, y^*, \dot{y}^*) = c$$

for some constant c. Allowing for variation of c, this equation has a two-fold infinity of solutions; and in the simplest cases this is just enough to yield a unique solution satisfying (1.2).

The assumption that $f_{\dot{y}}$ be constant in sign may be relaxed simply by taking

$$f(t, y^*, \dot{y}^*) = \pm c,$$

where the sign agrees with that of $f_{\dot{y}}$.

The object of this paper is to present a complete theory for linear systems alone, of which the chief results are contained in Theorem 1, §5, and Theorem 2, §6. Detailed consideration of non-linear cases is left for later publication. Meanwhile, rough extensions of the present theory by variational arguments (i.e., linearization near the solution) will undoubtedly yield correct results for most applications.

2. Formulation of the first order problem. All functions encountered are real, single-valued, and defined on a closed, bounded interval $I = [a, b]$.

By a "vector" we mean an ordered n-tuple of functions, $f = (f^i)$, for some fixed n. Similarly, a "matrix" is an nxn square matrix of functions. A vector or matrix is said to be continuous, or summable, etc., if and only if each of its components has that property. For matrix operations, vectors are regarded as rows or columns according to context.

Two spaces are fundamental to the discussion. The trial solutions $y(t)$ are taken from the space Y of absolutely continuous vectors; the values of the differential expressions lie in the space M of measurable vectors. Elements of M are regarded as equivalent, written $f \sim g$, when they are equal almost everywhere.

For all vectors we define the "vector absolute value"

$$|f| = |(f^i)| = (|f^i|), \tag{2.1}$$

and for elements of M we have the "maximum-norm"

https://doi.org/10.4153/CJM-1957-017-6 Published online by Cambridge University Press

$$||f|| = \sup\{\text{ess. sup.}|f^i| \mid i = 1, \ldots, n\}, \tag{2.2}$$

where the essential supremum of each component is taken over I.

Given an ordered n-tuple $J = (J^i)$ of subsets of I, the vector whose components are the characteristic functions of the corresponding sets J^i is called the characteristic vector of J.

To define the problem we are given

(a) a first order differential operator

$$\mathfrak{L}y = A[\dot{y} + By + c] \tag{2.3}$$

which serves to map Y into M. A is an almost everywhere finite and non-singular matrix, whose inverse A^{-1}, together with the matrix B and vector c, are Lebesgue summable.

(b) an ordered n-tuple J of measurable subsets of I, at least one of whose components has positive measure. The components of $\mathfrak{L}y$ will be free to vary on the corresponding component sets J^i, but must vanish almost everywhere on the complements $I - J^i$.

(c) a pair of initial conditions for vectors $y \in Y$, one for each endpoint of I:

$$y(a) = \eta_a, \quad y(b) = \eta_b. \tag{2.4}$$

Let $F \subset M$ be the space of equivalence classes of essentially bounded vectors f, whose components f^i vanish almost everywhere on the corresponding sets $I - J^i$. Let $X \subset Y$ be the set of all absolutely continuous vectors x which satisfy the initial conditions (2.4), and such that $\mathfrak{L}x \in F$. Then a vector x_0 is said to be a solution of the minimum-maximum problem if and only if $x_0 \in X$ and

$$||x_0|| = \inf\{||x|| \mid x \in X\}. \tag{2.5}$$

It is important to notice that a given problem may be inconsistent, in the sense that the set X is empty. (For example, $n > 1$, $\mathfrak{L}y \equiv \dot{y}$. J^1 is nul, and $\eta^1{}_a \neq \eta^1{}_b$. Here $\dot{x}^1 \sim 0$ for $x \in X$, which contradicts the requirement that $x^1(a) \neq x^1(b)$.) However, we will find that every consistent problem has a solution.

3. The set $G = \mathfrak{L}X$. Although the solutions are defined in terms of X, it is more convenient to work with the image G of X under $\mathfrak{L}$. This is possible because of

LEMMA 1. *The sets X and G are in 1:1 correspondence, for $\mathfrak{L}$ has a unique inverse on G.*

The proof consists of the observation that for every $f \in F$ the differential equation $\mathfrak{L}y \sim f$ has a unique solution satisfying *either* of the two initial conditions (2.4).[1]

[1]The properties of solutions of all the differential equations considered in this section are essentially the same as if the coefficients were continuous. They may be derived by a slight extension of the standard methods (**1**; **2**, chap. IX).

In view of this correspondence, the solutions may be defined in terms of G: g_0 is a solution if and only if $g_0 \in G$ and

$$||g_0|| = \inf\{||g|| \mid g \in G\}.$$

The remainder of this section is devoted to a derivation of the necessary and sufficient conditions stated in Lemma 2 for an element of F to belong to G. First we will find expressions for the two inverses of $\mathfrak{L}$ corresponding to the two initial conditions at a and b. The required conditions follow from the fact that these inverses are equal on G.

Let W be the space of solutions of the homogeneous equation

$$\dot{y} + By \sim 0,$$

and let Z be the space of solutions of the adjoint equation

$$\dot{y} - yB \sim 0.$$

W and Z are both n-dimensional subspaces of Y. For every pair $w \in W$ and $z \in Z$ the "scalar product"

$$zw = \sum_{i=1}^{n} z^i(t)w^i(t)$$

is constant, independent of t.

If n vectors $\{e_j\}$ form a basis for W, they may be combined into a matrix E which is everywhere non-singular. The matrix

$$K(t, s) = E(t)\, E^{-1}(s)$$

is independent of the choice of basis, and plays the role of "translation operator" in W and Z:

$$w(t) = K(t, s)\, w(s), \quad z(s) = z(t)\, K(t, s). \tag{3.1}$$

For every $f \in F$, the pair y_a, y_b of solutions of

$$\mathfrak{L}y \sim f$$

which satisfy the initial conditions (2.4) at a and b, respectively, are given by

$$y_\alpha(t) = w_\alpha(t) + \int_\alpha^t K(t, s)[A^{-1}(s)\, f(s) - c(s)]\, ds, \qquad \alpha = a, b, \tag{3.2a}$$

where $w_\alpha(t)$ is the element of W which is equal to η_α at $t = \alpha$:

$$w_\alpha(t) = K(t, \alpha)\, \eta_\alpha. \tag{3.2b}$$

Now $y_a = y_b$ if and only if $f \in G$, so we have, with the help of (3.1):

LEMMA 2. *$g \in G$ if and only if $g \in F$ and*

$$u(t) = \int K(t, s)\, A^{-1}(s)\, g(s)\, ds \tag{3.3}$$[2]

[2]Integrals written without limits mean integrals from a to b.

https://doi.org/10.4153/CJM-1957-017-6 Published online by Cambridge University Press

where u is an element of W which is determined by the vector c and the initial values η_a and η_b:

$$u(t) \equiv w_b(t) - w_a(t) + \int K(t, s)\, c(s)\, ds. \tag{3.4}$$

Moreover, (3.3) *is equivalent to the condition*

$$zu = \int \zeta(s)\, g(s)\, ds \qquad \text{for every } z \in Z \tag{3.5}$$

where, for each $z \in Z$, the vector ζ is defined by

$$\zeta = zA^{-1}. \tag{3.6}$$

4. The function $\mu(z)$. Our plan is to reduce the problem so that its solution amounts to maximizing a function on the n-dimensional space Z, rather than minimizing a function on the infinite-dimensional space F. This reduction depends on the following inequality, which is a direct consequence of (3.5) and the definition of F.

Let j be the characteristic vector of J. Then for every pair $z \in Z$ and $g \in G$,

$$|zu| \leqslant ||g|| \int |\zeta|\, j\, ds. \tag{4.1}$$

Now consider the set

$$V = \{z | z \in Z \text{ and } \int |\zeta|\, j\, ds = 0\}, \tag{4.2}$$

which is clearly a linear subspace of Z. When its dimension is less than n, the function

$$\mu(z) = zu / \int |\zeta|\, j\, ds \tag{4.3}$$

is defined on the complement CV of V in Z, and satisfies the inequality

$$\sup\{|\mu(z)| \mid z \in CV\} \leqslant \inf\{||g|| \mid g \in G\}. \tag{4.4}$$

We will see that equality actually holds, as long as μ is bounded.

Some useful properties of $\mu(z)$ are contained in the proof of

LEMMA 3. *When $V \neq Z$, $\mu(z)$ is bounded on CV if and only if*

$$zu = 0, \text{ for every } z \in V. \tag{4.5}$$

And if μ is bounded, then $|\mu|$ attains its supremum on CV.

Proof. Choose components with respect to an arbitrary basis as coordinates in Z, so that Z becomes homeomorphic to Euclidean n-space. Then both numerator and denominator in the expression (4.3) are clearly continuous on Z. If $\mu(z)$ is bounded on CV, then since V forms the boundary of CV, the numerator must vanish with the denominator as $z \to V$, and (4.5) must be true.

Conversely, let (4.5) be satisfied. Let U be any linear subspace of Z complementary to V, so that every $z \in CV$ is expressible in the form $z = z_1 + z_2$, where $z_1 \in U$, $z_2 \in V$, and $z_1 \neq 0$. Substituting into (4.3), we find that $\mu(z_1 + z_2) = \mu(z_1)$. Moreover, for every real α, $\mu(\alpha z) = \pm \mu(z)$, where the

https://doi.org/10.4153/CJM-1957-017-6 Published online by Cambridge University Press

sign agrees with that of α. Hence μ assumes all its functional values for z on the unit sphere in U. The proof is completed by recalling that a continuous function on a closed and bounded set in Euclidean space is bounded, and attains its supremum on that set.

5. Solution of the first order problem. The way is now prepared for

THEOREM 1. *The problem is consistent if and only if the condition* (4.5) *of Lemma* 3 *is satisfied. Moreover, every consistent problem has a solution as follows:*

(a) *in the trivial case* $u = 0$, *there is a unique solution* $g_0 \backsim 0$.

(b) *in the "degenerate case"* $V = Z$, *the problem is consistent only if* $u = 0$.

(c) *in all other cases the function* $|\mu(z)|$ *has a positive supremum which is attained for some* $z \in CV$. *Let* z_0 *be any such point, and let*

$$\mu_0 = \mu(z_0), \quad \zeta_0 = z_0 A^{-1}. \tag{5.1}$$

Then every solution has the property

$$||g_0|| = |\mu_0| \tag{5.2}$$

and its components are uniquely determined on the sets

$$J_0^i = \{t | t \in J^i, \quad \zeta_0^i(t) \neq 0\} \tag{5.3}$$

by the equivalences

$$g_0^i \sim \mu_0 \zeta_0^i / |\zeta_0^i|. \tag{5.4}$$

Thus each $g_0{}^i$ *is equal in absolute value to the constant* $|\mu_0|$ *almost everywhere on* $J_0{}^i$; *and the solution is unique if every* $J_0{}^i$ *has the same measure as the corresponding* J^i.

Proof. The problem is inconsistent if (4.5) is violated, for otherwise the inequality (4.1) provides an obvious contradiction. Now the result for case (b) follows, since (4.5) implies $u = 0$ when $V = Z$. And in view of Lemma 2, the result for case (a) is trivial.

The rest of the proof deals with case (c), in which $u \neq 0$ and $V \neq Z$. Assuming that (4.5) is satisfied, we will construct a $g_0 \in F$ that satisfies (3.3), (5.2), and (5.4). Then in view of Lemma 2, $g_0 \in G$, so (4.5) implies consistency. And by virtue of (4.4), g_0 is also a solution. To show that the components of every solution satisfy (5.4), let g_1 be any other solution. Since

$$|g_1{}^i| \leqslant |g_0{}^i| = |\mu_0|$$

almost everywhere on $J_0{}^i$, there exists a vector e such that $g_0{}^i - g_1{}^i \sim e_0{}^i g_0{}^i$ on each $J_0{}^i$, where $0 \leqslant e^i \leqslant 2$. Since g_0 and g_1 both satisfy (3.5), it follows that

$$\int \zeta_0 (g_0 - g_1)\, ds = \mu_0 \sum_i \int_{J_0^i} |\zeta_0^i|\, e^i ds = 0.$$

Hence each $e^i \sim 0$, and $g_1{}^i \sim g_0{}^i$ on $J_0{}^i$.

To construct g_0, consider the vectors $h \in F$ and $v \in W$ defined by

$$h^i = \begin{cases} \mu_0 \zeta_0^i / |\zeta_0^i| \text{ on } J_0^i, \\ 0 \quad \text{elsewhere,} \end{cases}$$

$$v = \int K(t, s)\, A^{-1}(s)\, h(s)\, ds$$

If $v = u$, take $g_0 \sim h$. If $v \neq u$, define a new problem with the same spaces W and Z but with u and J replaced by

$$u' = u - v,$$
$$J'^i = J^i - J_0^i.$$

It is shown below that the inequality

$$|zu'| \leqslant |\mu_0| \int |\zeta|\, j' ds \tag{5.5}$$

holds for every $z \in Z$, where j' is the characteristic vector of J'. Then it is easy to check that the new problem falls under case (c), and that

$$\sup |\mu'(z)| \leqslant |\mu_0|.$$

Moreover, the dimension of the new space V' exceeds that of V, for $V \subset V'$ and $z_0 \in V'$, while $z_0 \notin V$.

The procedure just outlined may be repeated to give a sequence of problems, which terminates as soon as the condition $v = u$ is satisfied. Termination occurs after at most $n - d$ steps, where d is the dimension of the original V. For if the sequence continues until V is of dimension $n - 1$, then $V' = Z$, and it follows from (5.5) that $u' = u - v = 0$.

If h_j denotes the vector h for the jth step, it is easy to check that

$$g_0 \sim \sum_j h_j \tag{5.6}$$

has all the required properties.

To prove (5.5), note first that

$$(z_0 + \alpha z)\, u/\mu_0 \leqslant \int |\zeta_0 + \alpha \zeta|\, j ds \tag{5.7}$$

for every z and every real α. Keeping z fixed, consider the vector k and the sets L_α defined by

$$k^i = \begin{cases} \zeta^i\, \zeta_0^i / |\zeta_0^i| \text{ on } J_0^i, \\ 0 \quad \text{elsewhere,} \end{cases}$$
$$L_\alpha^i = \{t | t \in J_0^i \text{ and } |\xi_0^i| + \alpha k^i < 0\}.$$

Then on $J_0{}^i$

$$|\zeta_0^i + \alpha \zeta^i| = |\zeta_0^i / |\zeta_0^i| (|\zeta_0^i| + \alpha \zeta^i\, \zeta_0^i / |\zeta_0^i|)| = ||\zeta_0^i| + \alpha k^i|$$

and on J'^i,

$$|\zeta_0^i + \alpha \zeta^i| = |\alpha| |\zeta^i|.$$

Hence, if j_0 and l_α are the characteristic vectors of J_0 and L_α respectively,

$$\int |\zeta_0 + \alpha \zeta|\, j ds = \int [|\zeta_0| + \alpha k]\, j_0 ds + |\alpha| \int |\zeta|\, j' ds - 2 \int [|\zeta_0| + \alpha k]\, l_\alpha\, ds. \tag{5.8}$$

https://doi.org/10.4153/CJM-1957-017-6 Published online by Cambridge University Press

Now in view of (3.1) and (3.6),

$$zv = \int \zeta(s)\, h(s)\, ds$$

and, since $\zeta h \equiv \mu_0 k j_0$,

$$\int k j_0 ds = zv/\mu_0$$

Moreover,

$$\int |\zeta_0|\, j_0 ds = \int |\zeta_0|\, j ds = z_0 u/\mu_0.$$

Combining the last two equations with (5.8) and (5.7) and using the fact that

$$0 < -[|\zeta_0^i| + \alpha k^i] < |\alpha k^i| \qquad \text{on } L_\alpha^i,$$

it follows that

$$\pm zu'/\mu_0 \leqslant \int |\zeta|\, j' ds + 2\int |k|\, l_\alpha\, ds,$$

where the sign agrees with that of α.

For every monotone vanishing sequence $\{\alpha_j\}$, the corresponding sequences $\{L^i_{\alpha j}\}$ are non-decreasing, and the infinite products $\prod_j L^i_{\alpha j}$ contain only points at which the k^i are infinite. Hence the measure of each L^i_α vanishes with α. The proof is completed by taking the two limits $\alpha \to \pm 0$ in the last inequality.

6. Higher order systems. The theory is easily extended to cases in which the operator $\mathfrak{L}$ is of higher order, by rewriting the given system in an equivalent first order form. The case of a single higher order expression provides an interesting illustration, in view of the special results obtained below.

Taking $n = 1$, let $\mathfrak{L}$ be replaced by the mth order operator

$$\mathfrak{L}_m y = a\left(\frac{d^m y}{dt^m} + \sum_{j=m-1}^{0} b_j \frac{d^j y}{dt^j} + c\right) \tag{6.1}$$

where a is almost everywhere finite, and $1/a$, the b_j, and c are summable functions. $||\mathfrak{L}_m x||$ is to be minimized over all functions x such that

(a) x and its first $m-1$ derivatives are absolutely continuous, and assume given values at the endpoints of I.

(b) $\mathfrak{L}_m x$ is essentially bounded, and vanishes almost everywhere on $I - J_m$, where J_m is a given subset of positive measure.

On taking

$$y^i = \frac{d^{i-1} y}{dt^{i-1}}, \qquad i = 1, \ldots, m$$

this reduces to a first order problem in which $n = m$, all the J^i are empty except $J^m = J_m$, and the matrix A is diagonal, with unit diagonal elements except $a^{mm} = a$. Accordingly, the function μ has the form

$$\mu(z) = zu/\int_{J_m} |z^m/a|\, ds$$

where the integral extends over J_m.

https://doi.org/10.4153/CJM-1957-017-6 Published online by Cambridge University Press

It is shown below that for every $z \neq 0$ the component z^m has no more than a finite number of zeros on I. Hence the set V consists of the single element $z = 0$, and the condition (4.5) is always satisfied. And if $|\mu|$ attains its supremum at z_0, the set

$$J_0^m = \{t | t \in J_m, \quad z^m/a \neq 0\}$$

differs from J_m by a set of measure zero. Thus we have

THEOREM 2. *For a single differential expression of the form* (6.1) *the problem is always consistent and has the unique solution*

$$g_{m0} \sim \begin{cases} z_0^m a/|z_0^m a| & \text{on } J_m, \\ 0 & \text{on } I - J_m. \end{cases}$$

And except in the trivial case $u = 0$, *the solution has a finite or infinite number of sign changes according as the factor* $a(t)$ *changes sign a finite or infinite number of times on* J_m.

Proof. Let z be any element of Z for which z^m has an infinity of zeros on I. These zeros must have a limit point $\tau \in I$. We will show that if, for any j in the range $m \leqslant j < 1$, τ is a common limit point of the zeros of each component $z^m, z^{m-1}, \ldots, z^j$, then τ is also a limit point of the zeros of z^{j-1}. Hence the zeros of *every* component have τ as a common limit point. Since z is continuous, it follows that every component vanishes at $t = \tau$, and hence that z vanishes identically.

Now the adjoint equations

$$z^i - b_{i-1} z^m + z^{i-1} \sim 0, \qquad j \leqslant i \leqslant m,$$

may be regarded as equations for $(z^j, \ldots, z^m)$ with the inhomogeneous term $(z^{j-1}, 0, \ldots, 0)$. If $z^j, \ldots, z^m$ all vanish at $t = \tau$, it follows from the analog of (3.2) that

$$z^j(t) = -\int_\tau^t k_j(t, s)\, z^{j-1}(s)\, ds$$

where k_j is continuous in both its arguments, and $k_j(\tau, \tau) = 1$. Hence, for sufficiently small $|t - \tau|$, the continuous function z^{j-1} must change sign between τ and each zero of z^j. This completes the proof.

REFERENCES

1. G. D. Birkhoff and R. E. Langer, *The boundary problems and developments associated with a system of ordinary differential equations of the first order*, Proc. Amer. Acad. Arts Sci. *58* (1923), 345–424.
2. E. J. McShane, *Integration* (Princeton, 1944).

Los Alamos Scientific Laboratory