57

ON THE CONVERGENCE OF A NUMERICAL METHOD FOR OPTIMAL CONTROL PROBLEMS

by

James W. Daniel

Technical Report #44

September 1968

ON THE CONVERGENCE OF A NUMERICAL METHOD FOR OPTIMAL CONTROL PROBLEMS*

by

James W. Daniel

1. INTRODUCTION

We seek to compute numerically an approximate solution to an optimal control problem of the following general type: Minimize $f(x, u) = \int_0^1 c(t, x(t), u(t))dt$ over the collection of functions (x, u) satisfying $\dot{x} \equiv \frac{dx}{dt} = s(t, x, u)$, $x(t) \in X(t)$, $u(t) \in U(t)$, $0 \le t \le 1$, where $X(t)$ and $U(t)$ are subsets of E^ℓ and E^k respectively; for the moment we remain vague as to the continuity properties of c, s, x, and u. The following numerical method has been proposed [6]: for positive integers n, set $h = \frac{1}{n}$, $t_i = ih$, $0 \le i \le n$, find vectors $x_n = (x_{n,0}, \cdots x_{n,n})$, $u_n = (u_{n,0}, \cdots u_{n,n})$ minimizing $h \sum_{i=1}^{n} c(t_i, x_{n,i}, u_{n,i})$ over the collection of vectors satisfying $\frac{x_{n,i+1} - x_{n,i}}{h} = s(t_i, x_{n,i}, u_{n,i})$ for $i = 0, \cdots n-1$, $x_{n,i} \in X(t_i)$, $u_{n,i} \in U(t_i)$, for $i = 0, \cdots n$. This method has proved useful in practice; under certain assumptions [6], the nonlinear programming problem defined by the numerical approximation can be computed rapidly by a variant of Newton's method. In this paper we are concerned not with methods for computing x_n, u_n, but with whether or not the sequence x_n,

---

*Prepared under Contract Number N00014-67-A-0128-0004 at the University of Wisconsin.

u_n, (or, more precisely, approximations to x_n, u_n) converges in some sense to a solution x, u to the original control problem.

We shall in Section 2 describe in general what criteria must be met for there to be any hope of proving convergence. Then in Section 3, using the point of view described in [1], we shall state the control problem as a variational problem in Hilbert space, consider the programming problem as a discretized variational problem in a finite dimensional space, and show how the precise statements of the general criteria given in Section 2 allow us to prove convergence in a weak sense and, in some conditions, in a stronger sense. In Section 4 we discuss generally how one can determine whether the criteria of Section 3 are in fact met; Section 5 briefly applies the ideas to linear problems, i.e., linear $s(t, x, u)$.

2. FEASIBILITY OF NUMERICAL METHOD; DESCRIPTION OF RESULTS

In order for the numerical method to have a chance of success, a number of obvious criteria must be met; these criteria will essentially be sufficient to guarantee convergence. Since the precise statements become rather complex in Section 3, we first pause to see intuitively what is needed.

First, the nature of the sets $X(t)$, $U(t)$ must be revealed fully by their nature at the discrete points t_i; for example, if $X(t) = \{t\}$ for irrational t but $X(t) = (-\infty, \infty)$ for rational t, the numerical method using $X(\frac{i}{n})$ would never detect restrictions. Thus we need to assume that $X(t)$, $U(t)$ vary

nicely in the sense that given vectors x_n, u_n with $x_{n,i} \in X(t_i)$, $u_{n,i} \in U(t_i)$, there exist functions x, u with $x(t) \in X(t)$, $u(t) \in U(t)$ and x and u near x_n and u_n in some sense. In the proof of convergence, it is necessary to know that to x_n, u_n as described and satisfying $x_{n,i+1} - x_{n,i} \leq hs(t_i, x_{n,i}, u_{n,i})$ there corresponds x, u satisfying $\dot{x} \leq s(t, x, u)$, with the distances of $x(t)$, $u(t)$ from $X(t)$, $U(t)$ being very small, and with x, u near x_n, u_n in some sense; the point (x, u) will be called $p_n(x_n, u_n)$. This says that to feasible points for the programming problem there are nearby almost feasible points for the control problem. This assumption will guarantee in essence that $X(t)$, $U(t)$ are well behaved.

Secondly we cannot hope to get an accurate solution unless there are feasible points for the programming problem arbitrarily near the solution to the control problem; this is not always the case as the following example indicates. Solve $\dot{x} = u$, $t^2 \leq x \leq t^2 + t$, $2t \leq u \leq 3t$, $t \in [0, 1]$, minimizing $\int_0^1 x^2 + u^2 \, dt$. The solution is $x = t^2$, $u = 2t$, but there are no feasible points at all for the programming problem with constraints $x_{i+1} = x_i + hu_i$, $(ih)^2 \leq x_i \leq (ih)^2 + ih$, $2ih \leq u_i \leq 3ih$ because $x_0 = u_0 = 0$ implies $x_1 = 0$. In this example however it is clear that there exist points satisfying the equality constraints and which are very near to being in $X(t_i)$, $U(t_i)$; if the constraints on the program are relaxed so as to allow such nearby points, we can prove convergence. Thus we will assume that there is a mapping r_n of the solution x, u of the control problem onto points x_n, u_n which satisfy the discrete

equality constraint, lie very near the discrete sets $X(t_i)$, $U(t_i)$, and are sufficiently near x, u.

In [6], the author actually treats inequality constraints of the form $x_{n,i+1} \le x_{n,i} + hs(t_i, x_{n,i}, u_{n,i})$ and indicates that one can solve the problem under equality constraints by a penalty function approach; we continue with that approach in this paper.

For a sequence of positive numbers a_n converging to zero we will minimize $h \sum_{i=1}^{n} [s(t_{i-1}, x_{n,i-1}, u_{n,i-1}) - \frac{x_{n,i} - x_{n,i-1}}{h}] + a_n h \sum_{i=1}^{n} c(t_i, x_{n,i}, u_{n,i})$ under inequality constraints and with the sets $X(t_i)$, $U(t_i)$ slightly expanded; the minimizing points will be shown to converge to a solution of the control problem under equality constraints. We will indicate in general how one might have mappings p_n and r_n and, in the particular case of a linear differential equation, observe what is sufficient to guarantee the existence of such mappings.

For notational convenience, we restrict ourselves henceforth to scalar problems; that is, we assume that x and u are in E^1. The situation for $x \in E^\ell$, $u \in E^k$ is exactly the same except that some statements, such as those regarding convexity of functions $s(t, x, u)$, must be taken to mean componentwise.

3. EXPLICIT ASSUMPTIONS; CONVERGENCE

Define the Hilbert space $\mathcal{H} = \{(x, u);\ u \in L_2(0,1),\ \dot{x} \in L_2(0,1),\ x$ is absolutely continuous$\}$; for $V = (x, u) \in \mathcal{H}$, let $\|V\|^2 = \int_0^1 \dot{x}^2\, dt + \int_0^1 u^2\, dt + x(0)^2$. Define the discretized space $\mathcal{H}_n = \{(x_n, u_n);\ x_n = (x_{n,0}, \cdots x_{n,n}),\ u_n = (u_{n,0}, \cdots u_{n,n})\}$; for $V_n = (x_n, u_n) \in \mathcal{H}_n$, let $\|V_n\|_n^2 = h \sum_{i=1}^{n} \left[\frac{x_{n,i} - x_{n,i-1}}{h}\right]^2 + h \sum_{i=1}^{n} u_{n,i}^2 + x_{n,0}^2$. We remark that weak convergence in $\mathcal{H}$ is equivalent to weak convergence of the compontnts $\dot{x}$ and u in $L_2(0,1)$ and convergence of $x(0)$ in E^1; this implies, by Ascoli's theorem [7], uniform convergence of x, i.e., convergence in $C[0,1]$.

Next we define functionals $f(V)$, $f_n(V_n)$, $g_n(V_n)$ as follows:

$f(V) = \int_0^1 c(t, x, u)dt$; $g(V) = \int_0^1 [s(t, x, u) - \dot{x}]\, dt$; $f_n(V_n) = h \sum_{i=1}^{n} c(t_i, x_{n,i}, u_{n,i})$; $g_n(V_n) = h \sum_{i=1}^{n} \left[s(t_{i-1}, x_{n,i-1}, u_{n,i-1}) - \frac{x_{n,i} - x_{n,i-1}}{h}\right]$; it is assumed that $c(t, x, u) \geq 0$. We seek to minimize $f(V)$ over the set $Q' \equiv \{(x, u);\ x(t) \in X(t),\ u(t) \in U(t)$ a.e., $\dot{x} \leq s(t, x, u)$ a.e.$\}$ with the additional constraint that $g(V) = 0$, which implies $\dot{x} = s(t, x, u)$ a.e. We assume that we need only consider V in some bounded subset (hence weakly compact) $S_B = \{V; \|V\| \leq B\}$ of $\mathcal{H}$; this would be true for example if the sets $X(t)$, $U(t)$ are bounded above and below by functions in $L_2(0,1)$ or if $f(V)$ satisfies $\lim_{\|V\| \to \infty} f(V) = \infty$. We assume that $Q \equiv S_B \cap Q'$ is weakly compact. Suppose f and g are

weakly lower semicontinuous functionals and that for our control problem there exists a solution $V' = (x', u') \in Q$ with $g(V') = 0$, i.e., $f(x', u') \leq f(x, u)$ if $(x, u) \in Q'$ and $g(x, u) = 0$.

As indicated by the second remark of Section 2, we assume the existence of a mapping r_n from $\mathcal{H}$ into $\mathcal{H}_n$ such that $(x'_n, u'_n) = V'_n \equiv r_n V'$ (recall that V' solves the control problem) satisfies $\lim_{n \to \infty} f_n(V'_n) = f(V')$, $\frac{x'_{n,i+1} - x'_{n,i}}{h} = s(t_i, x'_{n,i}, u'_{n,i})$ and $d_n \to 0$, where $d_n \geq d(V'_n) \equiv$ $\max_{0 \leq i \leq n}$ {distance from $x'_{n,i}$ to $X(t_i)$, distance from $u'_{n,i}$ to $U(t_i)$}. We then define the slightly enlarged feasible set for the programming problem as $Q_n = \{V_n = (x_n, u_n);\ \|V_n\|_n \leq B + d_n,\ d(V_n) \leq d_n,\ \frac{x_{n,i+1} - x_{n,i}}{h} \leq s(t_i, x_{n,i}, u_{n,i})\}$.

As indicated by the first remark of Section 2, we must assume the existence of a mapping p_n from $\mathcal{H}_n$ into $\mathcal{H}$ such that: if $V_n \in Q_n$ and $g_n(V_n) \to 0$, then $\lim_{n \to \infty} |f_n(V_n) - f(p_n V_n)| = \lim_{n \to \infty} |g_n(v_n) - g(p_n V_n)| = 0$, $(y_n, w_n) \equiv p_n V_n \in \mathcal{H}$ satisfies $\dot{y}_n \leq s(t, y_n, w_n)$ a.e., and $e_n \to 0$ where $e_n \geq e(p_n V_n) \equiv \max_{0 \leq t \leq 1}$ {distance of $(p_n x_n)(t)$ from $X(t)$, distance of $(p_n u_n)(t)$ from $U(t)$}, and $\|p_n V_n\| \leq B + e_n$. Finally we define the slightly enlarged feasible set for our control problem as $Q^n = \{V = (x, u);\ \|V\| \leq B + e_n,$ $e(V) \leq e_n,\ \dot{x} \leq s(t, x, u)$ a.e.$\}$

Theorem 3.1 Let f, g, f_n, g_n, Q, Q_n, Q^n, r_n, p_n be as described above. Let $a_n > 0$ converge to 0. For each n let W_n nearly minimize $g_n(V_n) + a_n f_n(V_n)$ over Q_n in the sense that $g_n(W_n) + a_n f_n(W_n) \leq \varepsilon_n a_n + g_n(V_n) + a_n f_n(V_n)$ for all $V_n \in Q_n$ with $\varepsilon_n \to 0$. Then all weak limit points W of $p_n W_n$, at least one of which exists, solve the optimal control problem, i.e., if $W = (x, u)$, then $\dot{x} = s(t, x, u)$ a.e., $W \in Q$, and $f(W) \leq f(V)$ for all $V \in Q$ with $g(V) = 0$.

Proof: First consider the auxiliary sequence $Y_n \in Q^n \subset \mathcal{H}$ satisfying $g(Y_n) + a_n f(Y_n) \leq g(Y) + \varepsilon_n$ for all Y in Q^n; this sequence exists. Recalling that $g(V') = 0$, we have $0 \leq g(Y_n) \leq g(Y_n) + a_n f(Y_n) \leq g(V') + a_n f(V') + \varepsilon_n$ which implies that $g(Y_n) + a_n f(Y_n)$ converges to zero. Now $g(Y_n) + a_n f(Y_n) \leq g(p_n W_n) + a_n f(p_n W_n) + \varepsilon_n \equiv g_n(W_n) + a_n f_n(W_n) + \varepsilon_n + \delta_n \leq g_n(r_n V') + a_n f_n(r_n V') + \varepsilon_n a_n + \varepsilon_n + \delta_n \equiv g(V') + a_n f(V') + \eta_n + \varepsilon_n a_n + \delta_n$ where η_n and δ_n, as defined by the inequalities, converge to zero because of the assumptions on r_n and p_n. Therefore $g(p_n W_n) + a_n f(p_n W_n)$ converges to zero.

By the boundedness of Q^n, weak limit points W of $p_n W_n$ exist and clearly must lie in Q since Q is weakly closed; thus $0 \leq g(W) \leq \lim_{n \to \infty} \inf g(p_n W_n) = 0$, so $g(W) = 0$, that is, if $W = (x, u)$, $\dot{x} = s(t, x, u)$ a.e. Thus W is a candidate for solving the control problem; we only need show that $f(W) \leq f(V')$.

We have $0 = g_n(r_n V') \le g_n(W_n) \le g_n(W_n) + a_n f_n(W_n) \le g_n(r_n V') + a_n f_n(r_n V') + a_n \varepsilon_n \le g_n(W_n) + a_n f_n(r_n V') + a_n \varepsilon_n$ which implies, from the fourth and sixth terms in the inequalities, that $f_n(W_n) \le f_n(r_n V') + \varepsilon_n$. Then $f(W) \le \lim_{n \to \infty} \inf f(p_n W_n) = \lim_{n \to \infty} \inf f_n(W_n) \le \lim_{n \to \infty} \inf [f_n(r_n V') + \varepsilon_n] = f(V')$. Since $f(W) \le f(V')$ and V' solves the control problem so must W. Q.E.D.

Simply stated, the above theorem shows that if one nearly solves a penalty function form of the programming problem over a slightly enlarged feasible set, then the minimizing points yield an arbitrarily accurate solution to the original control problem with the differential equality constraint. We have only shown of course that weak limit points solve the problem; if the solution V' is unique then clearly $p_n W_n$ weakly converges to it. In some cases, for instance if the functional f is uniformly convex, this implies convergence in norm in $\mathcal{H}$ [3, 4]; nonetheless, weak convergence here is such that $p_n x_n$ converges in $C[0,1]$, a strong condition.

4. THE EXISTENCE OF r_n, p_n

The theorem above involves many assumptions, e.g., weak compactness of Q, weak lower semicontinuity of f and g, and, most importantly, the existence of mappings r_n and p_n as described; the most troublesome assumptions are those on r_n and p_n.

First we examine the other assumptions. In [6], in order to prove that the numerical method for minimizing $g_n + a_n f_n$ works, it was assumed that

$s(t, x, u)$ and $c(t, x, u)$ were convex jointly in x and u; it is a simple matter to show, under that assumption and the additional one that $s_x(t, x, u)$, $s_u(t, x, u)$, $c_x(t, x, u)$, $c_u(t, x, u)$ exist and are in $L_2(0, 1)$ for $(x, u) \in \mathcal{H}$, that f, g, and Q satisfy their needed assumptions.

Let us see how one might define an operator r_n. Suppose $V' = (x, u)$ satisfies $\dot{x} = s(t, x, u)$ a.e.; suppose one knows, as is sometimes the case [4, 5], that u is continuous. Define x_n, u_n as follows: $u_{n,i} = u(t_i)$, $x_{n,i+1} = x_{n,i} + hs(t_i, x_{n,i}, u_{n,i})$, $x_{n,0} = x(0)$; clearly x_n is an approximation to $x(t)$ by Euler's method and it can be shown [2] that $|x_{n,i} - x(t_i)| = O(h + w(h))$, where $w(h) = \max |s(t, x, u(t)) - s(t^*, x, u(t^*))|$ for $|t - t^*| \le h$, $0 \le t, t^* \le 1$, and x in a certain bounded set. Moreover, if $x \in C^2[0, 1]$, then $|x_{n,i} - x(t_i)| = O(h)$ and we could take $d_n = h^{1/2}$ for constructing r_n. The condition that $f_n(r_n V') \to f(V')$ would follow, for example, from continuity of $c(t, x, u)$. Unfortunately it is not always the case that $u(t)$ and hence $s(t, x, u(t))$ will be continuous. Often, however, $u(t)$ is piecewise continuous with a finite number of discontinuities all of which are finite jumps. By a modification of the argument in [2], it is possible to show that $|x_{n,i} - x(t_i)| = O(h + w(h))$ where $w(h)$ has the same definition as before except that t and t^* are required to lie in the same interval of continuity for $u(t)$. Without an actual estimate for $w(h)$, however, the number d_n and hence the set Q_n is unknown; if $u(t)$ is known to be piecewise constant, then $w(h)$ could be estimated. We do not know how to guarantee the existence of r_n or estimate d_n in other situations.

The mapping p_n is somewhat more difficult. Given x_n, u_n with $x_{n,i+1} = x_{n,i} + hs(t_i, x_{n,i}, u_{n,i}) - h\, b_{n,i}$, $b_{n,i} \geq 0$, with $h \sum_{i=1}^{n} b_{n,i} \to 0$, one could define $p_n u_n$ as the piecewise linear interpolant of u_n, $b_n(t)$ as the piecewise linear interpolant of $b_{n,i}$, and $p_n x_n$ as the solution of $\dot{x} = s(t, x, p_n u_n) - b_n(t)$. Again x_n is the numerical solution by Euler's method for this equation. One needs to know that $x(t)$ and x_n are very close uniformly in u_n and d_n; even then, one needs to know that $X(t)$ changes smoothly enough that $x(t_i)$ near $X(t_i)$ for all i will imply $x(t)$ near $X(t)$ for all t, and similarly for u, U. This will be true for example if $X(t)$ (and also $U(t)$) has the form $X(t) = \{x;\ m(t) \leq x \leq M(t)\}$, where m and M are piecewise continuous having a finite number of discontinuities, all of jump type; values of $\pm \infty$ are allowed for m and M. It is understood that the grid $(t_0, \cdots t_n)$ should be modified so as to include for all n the finite number of points of discontinuity of the functions m, M defining $X(t)$ (and $U(t)$). In addition one needs that $f(p_n V_n)$ and $f_n(V_n)$, $g(p_n V_n)$ and $g_n(V_n)$ are near each other; in the next Section we indicate a condition under which this occurs.

5. LINEAR PROBLEMS

Suppose that $s(t, x, u) = A(t)\, x + B(t)\, u + C(t)$ with A, B, and C continuous. As remarked above, if the u corresponding to the solution V' is continuous and if $x \in C^2[0,1]$, then r_n exists and we can expand our constraint sets

$X(t_i)$ and $U(t_i)$ by $h^{1/2}$ and be sure of containing $r_n V^*$. If $X(t)$ and $U(t)$ are as described in the preceding section, then by a straightforward argument, one can show that the mapping p_n also described in the preceding Section satisfies the needed assumptions if c satisfies the conditions $|c(t_1, x_1, u) - c(t_2, x_2, u)| \leq (a + b|u|^2)|c(t_1, x_1, 0) - c(t_2, x_2, 0)|$ with $c(t, x, 0)$ continuous in (t, x). This requirement is used to show that $|f_n(V_n) - f(p_n V_n)| \to 0$; the fact that $p_n V_n$ and V_n are uniformly close follows from the natures of $s(t, x, u)$ and of $X(t)$, $U(t)$.

REFERENCES

1. Daniel, J. W., "Approximate minimization of functionals," Univ. of Wisconsin, Computer Sciences Technical Report #42 (1968).

2. Henrici, P., Discrete variable methods in ordinary differential equations, Wiley, New York (1962), 16-29.

3. Levitin, E. S., and Poljak, B. T., "Convergence of minimizing sequences in conditional extremum problems," Soviet Math. Dokl., Vol. 7 (1966), 764-767.

4. Poljak, B. T., "Existence theorems and convergence of minimizing sequences in extremum problems with restrictions," Soviet Math. Dokl., Vol 7 (1966), 72-75.

5. Pontryagin, L. S., Boltyanskii, V. G., Gamkrelidze, R. V., and Miscenko, E. F., The mathematical theory of optimal processes, Wiley, New York (1962).

6. Rosen, J. B., "Iterative solution of nonlinear optimal control problems," J. SIAM Control, Vol. 4 (1966), 223-244.

7. Taylor, A. E., Introduction to functional analysis, Wiley, New York (1961), 276.