Suboptimal Control of Nonlinear Systems:

I. Unconstrained

A. P. J. WEBER and L. LAPIDUS

Princeton University, Princeton, New Jersey

An algorithm for suboptimal control of nonlinear unconstrained systems is developed. The result is closed loop, uses the familiar linear-quadratic optimal approach, and minimizes computational time and effort. Numerous examples are presented.

A vast flow of recent publications deals with the theoretical details of the optimal control of dynamical systems. However, the application of these theories has lagged behind the theory because of the difficulty of implementing the algorithms and the uncertainties in the actual specification of the system models.

In this first paper we shall develop the computational features of the suboptimal control of nonlinear systems with essentially quadratic indices. Such a development has the advantages of minimizing computational time and effort, of using the well-developed linear-quadratic (L-Q) control problem, of being able to include constraints in almost any form, and of generating a feedback or closed-loop type of control. The disadvantage of this approach is that optimal control is rarely achieved. As shall be pointed out in many numerical examples, however, this disadvantage is not a serious one.

We shall also outline algorithms for achieving suboptimal control of nonlinear dynamical systems with quadratic indices when constraints are not present. To allow for the inclusion of constraints, Part II, which follows, will specify a means of solving the L-Q optimal control problem with control and state constraints. These new results can then be used to generate suboptimal control of nonlinear constrained systems with quadratic indices.

The number of publications in the suboptimal control area is small but growing rapidly. McClamroch (*11*) obtained suboptimal control policies as a modification of those for related processes with known optimal policies. To achieve suboptimal control, Friedland (*7*), Rubin (*17*), Kokotovic and Sannuti (*9*) and Meditch (*12*) involved a reduction of the system dimension or a partitioning of the system. Paradis and Perlmutter (*13*), Brosilow and Handley (*3*), and Chant and Luus (*5*) used an instantaneous minimization of the control criteria.

Suboptimal control was also obtained by Eller and Aggarwal (*6*) who used a special form of linearization and by Pearson (*14*), Westcott et al. (*21*), and Baldwin and Sims-Williams (*2*) who used another form of linearization. Thiriet and Deledieq (*19*) analyzed a special ordering of suboptimal control policies. Burghart (*4*) considered the suboptimal linear regulator with parameter variations while Seinfeld and Kumar (*18*) considered the suboptimal control of distributed systems. We do not consider here these optimal iterative techniques which approach but do not achieve optimality because of premature termination of the iterations.

THE DISCRETE L-Q OPTIMAL CONTROL PROBLEM

Because of their importance to this work, we include the necessary final results of the discrete L-Q optimal control problem; the explicit manipulations can be found in the literature (*10*).

Considering a fixed final time and a lumped-parameter deterministic dynamical system, we define the discrete L-Q optimal control problem as that sequence of discrete optimal control vectors $\mathbf{u}_k^0$, $k = 0, 1, 2, \ldots, N-1$, such that the performance index

$$I[\mathbf{x}_0, N] = \tau \sum_{k=1}^{N} [\mathbf{x}_k^T \mathbf{Q}\mathbf{x}_k + \mathbf{u}^T_{k-1} \mathbf{R}\mathbf{u}_{k-1}] \qquad (1)$$

is minimized subject to the system constraints of

$$\mathbf{x}_{k+1} = \varphi\mathbf{x}_k + \mathbf{\Delta}\mathbf{u}_k \qquad (0 \leq k \leq N-1) \qquad (2)$$

$$\mathbf{x}_0 \text{ given}$$

Here $\mathbf{x}_k$ is an n-dimensional vector (the state), $\mathbf{u}_k$ is an r-dimensional vector (the control), φ is an $n \times n$ matrix, $\mathbf{\Delta}$ is an $n \times r$ matrix and $\mathbf{Q}$ and $\mathbf{R}$ are $n \times n$ and $r \times r$ weighting matrices. Also, I is a scalar function (performance index), τ is a sampling period, and k is an index locating $\mathbf{x}_k$ and $\mathbf{u}_k$ in the overall interval $k = 0$ to $k = N - 1$. No constraints are specified on $\mathbf{x}_k$ or $\mathbf{u}_k$.

The interested reader may consult reference *10* for a detailed description of the implications and importance of the weighting matrices $\mathbf{Q}$ and $\mathbf{R}$; that is, they are required to be symmetric and positive semidefinite. There it is shown also that the solution to this problem is given by the recursive set of matrix equations

$$\mathbf{K}_{k-1} = [\mathbf{\Delta}^T (\mathbf{Q} + \mathbf{P}_k)\mathbf{\Delta} + \mathbf{R}]^{-1} \mathbf{\Delta}^T (\mathbf{Q} + \mathbf{P}_k)\varphi \qquad (k = N, N-1, \ldots 1) \qquad (3)$$

$$\mathbf{P}_{k-1} = (\varphi - \mathbf{\Delta}\mathbf{K}_{k-1})^T (\mathbf{Q} + \mathbf{P}_k)\varphi \qquad (k = N, N-1, \ldots 1) \qquad (4)$$

with a starting condition of

$$\mathbf{P}_N = \mathbf{0} \qquad (5)$$

and with an optimal feedback control of

$$\mathbf{u}_k^0 = -\mathbf{K}_k\mathbf{x}_k \qquad (k = 0, 1, \ldots, N-1) \qquad (6)$$

Briefly, Equations (3) to (4) are used by starting with the given condition (5); with $\mathbf{P}_N = \mathbf{0}$ known, Equation (3) can be used to generate $\mathbf{K}_{N-1}$ and then (4) to generate $\mathbf{P}_{N-1}$. These, in turn, generate in a backwards direction $\mathbf{K}_{N-2}$, $\mathbf{P}_{N-2}$, $\mathbf{K}_{N-3}$, $\mathbf{P}_{N-3}$, . . . until after N steps $\mathbf{K}_0$ is achieved. The values of $\mathbf{K}_0, \mathbf{K}_1, \ldots, \mathbf{K}_{N-1}$ are then used in Equation (6) to generate the optimal controls in a feedback sense, as $\mathbf{x}_k$ rather than just $\mathbf{x}_0$ is used. This implies, of course, that when $\mathbf{u}_k^0$ is available it may be used with $\mathbf{x}_k$ in Equation (2) to generate $\mathbf{x}_{k+1}$.

Thus, the solution to the discrete L-Q control problem, when φ, $\mathbf{\Delta}$, $\mathbf{Q}$ and $\mathbf{R}$ are constant matrices, involves a backwards pass using Equations (3) to (5) to generate $\mathbf{K}_k$ and $\mathbf{P}_k$ and a forward pass using (2) and (6) to generate $\mathbf{x}_k$ and $\mathbf{u}_k^0$.

One special case is also of interest. When N becomes large and $\mathbf{Q}$ and $\mathbf{R}$ have the properties mentioned above, it is possible to show that $\mathbf{K}_k \rightarrow \mathbf{K}$ = constant matrix. In this case the optimal feedback equation becomes merely

$$\mathbf{u}_k{}^0 = -\mathbf{K}\mathbf{x}_k \tag{7}$$

and it is not necessary to store $\mathbf{K}_{N-1}, \mathbf{K}_{N-2}, \ldots$.

With these equations in hand, we should point out that the discrete system Equation (2) can be obtained directly from a linear continuous system form, namely

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) \tag{8}$$

By treating Equation (8) as a piecewise constant coefficient equation, that is by approximating $\mathbf{A}(t)$ and $\mathbf{B}(t)$ with $\mathbf{A}_k = \mathbf{A}(k\tau)$ and $\mathbf{B}_k = \mathbf{B}(k\tau)$ for $k = 0, 1, \ldots, N-1$, it may be solved directly to give

$$\mathbf{x}_{k+1} = \varphi_k \mathbf{x}_k + \mathbf{\Delta}_k \mathbf{u}_k \tag{9}$$

φ_k and $\mathbf{\Delta}_k$ are defined as

$$\varphi_k = \exp\left[\mathbf{A}_k\tau\right] = \sum_{i=0}^{\infty} \frac{(\mathbf{A}_k\tau)^i}{i!} \tag{10}$$

and

$$\mathbf{\Delta}_k = \left[\int_0^{\tau} \exp\left(\mathbf{A}_k \lambda\right) d\lambda\right] \mathbf{B}_k \tag{11}$$

When $\mathbf{A}$ and $\mathbf{B}$ and thus φ and $\mathbf{\Delta}$ are constant, Equation (2) follows directly from Equation (9). In these equations $N = t_f/\tau$ where t_f is the final time of control; φ is termed the transition matrix, and it can be shown that φ is always nonsingular.

All of this optimal control material is well known (*8, 10*). However, these results are based on the simplifying assumptions of constant φ, $\mathbf{\Delta}$, $\mathbf{Q}$, and $\mathbf{R}$ and are insufficient for our purpose of suboptimal control. In fact, we will need to consider all four matrices as functions of k. When the weighting matrices $\mathbf{Q}$ and $\mathbf{R}$ are time varying, the performance index corresponding to Equation (1) is

$$I[\mathbf{x}_0, N] = \tau \sum_{k=1}^{N} \left[\mathbf{x}_k{}^T \mathbf{Q}_k \mathbf{x}_k + \mathbf{u}^T{}_{k-1} \mathbf{R}_{k-1} \mathbf{u}_{k-1}\right] \tag{12}$$

By an argument of mathematical induction (*20*) we have shown that, even when φ, $\mathbf{\Delta}$, $\mathbf{Q}$ and $\mathbf{R}$ are time varying, the same recurrence relations as Equations (3) to (5) hold except that the subscript k must be added to all matrices. In other words, the sequence

$$\mathbf{K}_{k-1} = \left[\mathbf{\Delta}^T{}_{k-1} (\mathbf{Q}_k + \mathbf{P}_k) \mathbf{\Delta}_{k-1} + \mathbf{R}_{k-1}\right]^{-1} \mathbf{\Delta}^T{}_{k-1} (\mathbf{Q}_k + \mathbf{P}_k) \varphi_{k-1} \qquad (k = N, N-1, \ldots, 1) \tag{13}$$

$$\mathbf{P}_{k-1} = (\varphi_{k-1} - \mathbf{\Delta}_{k-1}\mathbf{K}_{k-1})^T (\mathbf{Q}_k + \mathbf{P}_k) \varphi_{k-1} \qquad (k = N, N-1, \ldots, 1) \tag{14}$$

with

$$\mathbf{P}_N = \mathbf{0} \tag{15}$$

yields the unique solution to the discrete L-Q optimal control problem even when the system matrices and the weighting matrices in the performance index are time varying. The arrays $\mathbf{P}_k$ are symmetric and positive semidefinite. The only requirements are that $\mathbf{Q}_k$ and $\mathbf{R}_k$ be symmetric and nonnegative definite and that the array $\mathbf{\Delta}^T{}_{k-1}(\mathbf{Q}_k + \mathbf{P}_k)\mathbf{\Delta}_{k-1} + \mathbf{R}_{k-1}$ be positive definite. This derivation is developed by the use of Bellman's principle of optimality and represents a contribution not previously presented in the literature.

APPARENT LINEARIZATION

As the suboptimal algorithms to be mentioned shortly are to be applied to nonlinear dynamical systems as opposed to the linear ones of Equations (2) or (8), it is necessary to apply some type of linearization. The normal variational approach in which a nonlinear system of the form

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}, \mathbf{u}, t) \tag{16}$$

is converted via Taylor Series to

$$\delta\dot{\mathbf{x}}(t) = \mathbf{f}_{\mathbf{x}}\,\delta\mathbf{x} + \mathbf{f}_{\mathbf{u}}\,\delta\mathbf{u} \tag{17}$$

or

$$\delta\dot{\mathbf{x}}(t) = \mathbf{A}(t)\delta\mathbf{x} + \mathbf{B}(t)\delta\mathbf{u} \tag{18}$$

is not employed for linearization. In Equations (17) and (18), $\delta\mathbf{x}$ and $\delta\mathbf{u}$ represent perturbations around a given $\mathbf{x}(t)$ and $\mathbf{u}(t)$, and $\mathbf{f}_{\mathbf{x}}$ and $\mathbf{f}_{\mathbf{u}}$ are derivatives of $\mathbf{f}$ with respect to $\mathbf{x}$ and $\mathbf{u}$, respectively. Instead of using this variational approach for linearization, we prefer to use a method proposed by Pearson (*14*) in which the original system equations are used although the selection of the coefficient matrices is not unique. Two examples briefly illustrate this method. In the first, Equation (16) is written as

$$\dot{x}_i(t) = g_i(\mathbf{x}, u, t) + h_i(\mathbf{x}, u, t) + p_i(\mathbf{x}, u, t) \quad (i = 1, 2) \tag{19}$$

with at least 1 term nonzero. We now rewrite (19) as

$$x_i(t) = \left[\frac{g_i}{x_1}\right] x_1 + \left[\frac{h_i}{x_2}\right] x_2 + \left[\frac{p_i}{u}\right] u \tag{20}$$

where

$$\lim_{x_1 \to 0} \left|\frac{g_i}{x_1}\right| < \infty, \quad \lim_{x_2 \to 0} \left|\frac{h_i}{x_2}\right| < \infty, \quad \text{and}$$

$$\lim_{u \to 0} \left|\frac{p_i}{u}\right| < \infty \quad (i = 1, 2) \tag{21}$$

The coefficient matrices of the linearized system

$$\dot{\mathbf{x}} = \mathbf{A}(\mathbf{x}, u, t)\mathbf{x} + \mathbf{B}(\mathbf{x}, u, t)u \tag{22}$$

would be

$$\mathbf{A} = \begin{bmatrix} g_1/x_1 & h_1/x_2 \\ g_2/x_1 & h_2/x_2 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} p_1/u \\ p_2/u \end{bmatrix} \tag{23}$$

By providing state and control trajectories, the dependence of $\mathbf{A}$ and $\mathbf{B}$ on $\mathbf{x}$ and u can be eliminated to yield $\mathbf{A}(t)$ and $\mathbf{B}(t)$.

In the second example, consider the following special case of Equation (16)

$$\dot{\mathbf{x}}(t) = \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} x_1{}^2 x_2 + t(x_2 + u_2) + x_3 u_1 \\ x_1 x_3 + \exp(x_2) - 1 + x_2 u_2{}^{1/4} \\ x_2 + \tanh x_3 + x_3 u_2{}^2 \end{bmatrix} \tag{24}$$

or

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{u}(t) \tag{25}$$

where one of several possible choices of $\mathbf{A}(t)$ and $\mathbf{B}(t)$ is

$$\mathbf{A}(t) = \begin{bmatrix} x_1 x_2 & t & 0 \\ x_3 & [\exp(x_2) - 1]/x_2 + u_2{}^{1/4} & 0 \\ 0 & 1 & (\tanh x_3)/x_3 \end{bmatrix} \tag{26}$$

$$\mathbf{B}(t) = \begin{bmatrix} x_3 & t \\ 0 & 0 \\ 0 & x_3 u_2 \end{bmatrix}$$

Note that $\lim_{x_2 \to 0} \{[\exp(x_2) - 1]/x_2\} = 1$

and $\lim_{x_3 \to 0} [(\tanh x_3)/x_3] = 1$

so the elements of $\mathbf{A}(t)$ and $\mathbf{B}(t)$ are well behaved.

Note that this form of system rearrangement merely transforms the equations so that if $\mathbf{x}$ and $\mathbf{u}$ are specified the equations appear to be linear. Thus we refer to it as an apparent linearization.

METHODS OF SUBOPTIMAL CONTROL

Before detailing methods of suboptimal control, we should present some figure of merit or degree of suboptimality for comparing suboptimal to optimal control. Here we use the figure of merit (which may not be as sensitive as desired) given by

$$\text{degree of suboptimality in } \% = \left(\frac{I - I^0}{I^0}\right) 100 \quad (28)$$

where I^0 and I represent the optimal and suboptimal indices, respectively. The suboptimal control is perfect when the degree of suboptimality is 0%.

We have developed at least two noniterative suboptimal algorithms including the fixed final time (FFT) method and the infinite final time method (IFT). One iterative method (ITER) will also be discussed because it has application in the constrained control case to be presented in Part II. An understanding of all these methods can be obtained by examining how the FFT method develops $\mathbf{u}_0$. Given a nonlinear system equation of the form of (16), the apparent linearization is carried out to yield Equation (22) or (25). For $t = 0$, the matrices in $\mathbf{A}$ and $\mathbf{B}$ are evaluated with $\mathbf{u}_0 = \mathbf{0}$, i.e.,

$$\mathbf{A} = \mathbf{A}(\mathbf{x}_0, \mathbf{0}, 0)$$
$$\mathbf{B} = \mathbf{B}(\mathbf{x}_0, \mathbf{0}, 0) \quad (29)$$

This may be looked upon as using the assumed control and state trajectories

$$\mathbf{u}(t) = \mathbf{0}$$
$$\mathbf{x}(t) = \mathbf{x}_0 \quad (30)$$

With these values of $\mathbf{A}$ and $\mathbf{B}$, the dynamical equation can be put into the discrete-time form (2) by the method of Equations (9) to (11). Now we have the performance index of (12) and the system Equation (2) forming a linear-quadratic control problem. This problem may be solved by using the equation sequence (13) to (15) to generate $\mathbf{K}_0$. In turn, $\mathbf{u}_0$ is calculated from

$$\mathbf{u}_0 = -\mathbf{K}_0\mathbf{x}_0 \quad (31)$$

and then by integrating the nonlinear equation

$$\dot{x}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}_0, t) \quad (32)$$

over $t = 0$ to $t = \tau$, the value of $\mathbf{x}_1 = \mathbf{x}(\tau)$ is generated. Note that this proecdure uses the apparent linearization to freeze the coefficient matrices in the dynamical equations and then the standard L-Q problem to generate the first $\mathbf{u}_0$.

The remaining $N - 1$ control vectors $\mathbf{u}_k$ for $k = 1, 2, \ldots, N - 1$ are produced in the same manner as $\mathbf{u}_0$. Suppose $\mathbf{x}_q = \mathbf{x}(q\tau)$ has just been calculated; $\mathbf{u}_q$ may then be evaluated by the sequence of steps of first calculating $\mathbf{A}$ and $\mathbf{B}$ at $t = q\tau$ with the $\mathbf{u}_{q-1}$ just used to evaluate $\mathbf{x}_q$. In other words, this is equivalent to supplying the assumed control and state trajectories

$$\mathbf{u}(t) = \mathbf{u}_{q-1}$$
$$\mathbf{x}(t) = \mathbf{x}_q \quad (33)$$

to form

$$\mathbf{A} = \mathbf{A}(\mathbf{x}(q\tau), \mathbf{u}_{q-1}, q\tau)$$
$$\mathbf{B} = \mathbf{B}(\mathbf{x}(q\tau), \mathbf{u}_{q-1}, q\tau) \quad (34)$$

Secondly, one solves the new L-Q problem with the new $\mathbf{A}$ and $\mathbf{B}$ to generate $\mathbf{K}_q$ and in turn

$$\mathbf{u}_q = -\mathbf{K}_q\mathbf{x}_q \quad (35)$$

Finally, the system equation

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}_q, t) \quad (36)$$

is integrated from $t = q\tau$ to $t = (q + 1)\tau$ to generate $\mathbf{x}_{q+1}$.

According to this method, by stepping from one state to another until $\mathbf{x}_N$ is calculated, it is possible to generate the entire state trajectory in one complete integration of the nonlinear system equation. No state and control trajectories have to be made available from another control method in order to initiate the calculations.

This control is suboptimal for two reasons. First, each time a control vector is calculated, the nonlinear system is treated as a linear one. Because the nonlinearity is not taken into account, the control cannot be optimal. Furthermore, the linearization makes it impossible to determine, at a given time step, a control policy from that time onward which is optimal with respect to the state at that time step. When applied to a true linear system with time-varying coefficients, the equivalent behavior can be achieved by pretending that the values of $\mathbf{A}$ and $\mathbf{B}$ are known only at the time corresponding to the state just calculated.

In contrast to the method just described, the IFT method generates the feedback matrices $\mathbf{K}_q$ as if the final time were infinite. In other words, the sequence of Equations (13) to (15) is stepped backwards until $\mathbf{K}$ becomes constant. This constant matrix is then used in the preceding calculations. Note, however, that $\mathbf{Q}_k$ and $\mathbf{R}_k$ must be constant matrices to assure that a constant $\mathbf{K}$ is achieved. Also, every time the system equation is integrated forward one time step, $\mathbf{A}$ and $\mathbf{B}$ will in general change making it necessary to reevaluate $\mathbf{K}$.

Finally, we mention the ITER method which starts with an assumed control $\hat{\mathbf{u}}(t)$ and subsequently calculated $\hat{\mathbf{x}}(t)$. The matrices $\mathbf{A}$ and $\mathbf{B}$ may now be treated as functions of time according to

$$\mathbf{A}(t) = \mathbf{A}(\hat{\mathbf{x}}, \hat{\mathbf{u}}, t); \quad \mathbf{B}(t) = \mathbf{B}(\hat{\mathbf{x}}, \hat{u}, t) \quad (37)$$

and the system equations of (9) developed directly. Now we have a complete L-Q problem again and the feedback matrices $\mathbf{K}_{N-1}, \mathbf{K}_{N-2}, \ldots, \mathbf{K}_0$ may be generated. These in turn yield directly $\mathbf{u}_0, \mathbf{u}_1 \ldots, \mathbf{u}_{N-1}$ and $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$. Thus, in the ITER method the control must be completely assumed, and a new suboptimal control is generated with only one solution of the L-Q problem. Obviously, the initial control sequence could be generated from one of the previous noniterative suboptimal control methods and the iteration can be repeated any number of times desired.

Test Problems

In order to test the feasibility of the different suboptimal control methods we have chosen a number of dynamical systems whose unconstrained control can be analyzed. These systems span the spectrum from the very simple to the complex, and only the final equations and the corresponding $\mathbf{A}$ and $\mathbf{B}$ matrices will be presented. Further

TABLE 1. THREE SYSTEMS FOR UNCONSTRAINED CONTROL

System 1	System 2	System 3
In the dynamic equation of the form $\dot{\mathbf{x}} = f(\mathbf{x},t)\mathbf{x} + \mathbf{u}$		
$f(\mathbf{x},t) = -\frac{1}{2}t$	$2.5(1\text{-}0.08t)\sin\{4.0[(1.0\text{-}0.1t)t + 0.1]\} + 0.25$	$\begin{bmatrix} -x_1^2 & 0 \\ 0 & -x_2^2 \end{bmatrix}$
Linear L-Q problem with time varying coefficient Analytical solution available	Linear L-Q problem with damped oscillatory coefficient	Nonlinear problem Analytical solution available
$\mathbf{A}(t) = \begin{bmatrix} -\frac{1}{2}t & 0 \\ 0 & -\frac{1}{2}t \end{bmatrix}$	$\begin{bmatrix} f(t) & 0 \\ 0 & f(t) \end{bmatrix}$	$\begin{bmatrix} -x_1^2 & 0 \\ 0 & -x_2^2 \end{bmatrix}$
$\mathbf{B}(t) = \mathbf{I}$	$\mathbf{I}$	$\mathbf{I}$

details can be found in Weber (*20*).

In Table 1, each of the first three systems is shown to have two states, two controls, initial conditions of $\mathbf{x}(0) = [1\ -1]^T$, and performance indices of the continuous quadratic form. While these three systems might be looked upon as special, the two final systems are of more direct chemical engineering interest. The first of these is the CSTR (continuous stirred-tank reactor) with proportional control of the reactor accomplished by regulating the flow of coolant through cooling coils. Details of the dynamical system equations have been presented by Aris and Amundson (*1*) and Lapidus and Luus (*10*). In normalized dimensionless terms, the system equations are given by

$$\begin{aligned} \dot{x}_1 &= -(1 + \exp E)x_1 - 0.5(\exp E - 1) \\ \dot{x}_2 &= (\exp E)x_1 - \omega x_2^2 - (2 + 0.25\omega)x_2 \\ &\quad + 0.5(\exp E - 1) + u \end{aligned} \tag{38}$$

where $E = 25x_2/(x_2 + 2)$. In addition

$$\begin{aligned} &-0.5 \leq x_1 \leq 0.5 \\ &-0.25 \leq x_2 \quad \ldots \text{ for } \omega = 0 \\ &-1/\omega \leq x_2 \leq 1/\omega \quad \ldots \text{ for } \omega > 4 \\ &-1.75 < u \end{aligned} \tag{39}$$

Here x_1, x_2, and u are the equivalent output concentration, output temperature, and inlet temperature variables. ω is a parameter such that when $\omega = 0$, the steady state solution of (38) yields one unstable and two stable equilibrium points; when $\omega = 8.9$, a single stable limit cycle exists; and when $\omega = 20$, a single stable equilibrium point exists.

Equation (38) is linearized by the method of Pearson to yield the **A** and **B** matrices

$$\mathbf{A}(\mathbf{x}) = \begin{bmatrix} -(1 + \exp E) & -0.5(\exp E - 1)/x_2 \\ \exp E & -\omega x_2 - (2 + 0.25\omega) + 0.5(\exp E - 1)/x_2 \end{bmatrix}; \quad \mathbf{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{40}$$

The initial condition $\mathbf{x}(0)$ on the state is a function of the parameter ω. In particular, the following cases have been used

$$\mathbf{x}(0) = \begin{bmatrix} 0.09 \\ -0.04 \end{bmatrix}; \begin{bmatrix} -0.1111889 \\ 0.0323358 \end{bmatrix}; \begin{bmatrix} -0.15 \\ 0.05 \end{bmatrix} \tag{41}$$

$\omega = 0$ $\qquad$ $\omega = 8.9$ and 20 $\qquad$ $\omega = 8.9$

along with $t_f = 5$ and 10 and $\tau = 0.1$. The performance indices used are the previous ones, but with $\mathbf{Q} = \mathbf{I}$ and $\mathbf{R}$ as the scalar 1 (only one control variable). Optimal control of this CSTR system is achieved by minimizing a continuous-time performance index using the method of quasilinearization detailed by Rothenberger and Lapidus (*16*).

The final system investigated was a nonlinear absorber with six state and two control variables as detailed by Rothenberger (*15*). The normalized state equations are

$$\begin{aligned} \dot{x}_1 &= \{-[40.8 + 66.7(M_1 + 0.08x_1)]x_1 \\ &\quad + 66.7(M_2 + 0.08x_2)x_2\}/Y_1 + 40.8u_1/Y_1 \\ \dot{x}_i &= \{40.8x_{i-1} - [40.8 + 66.7(M_i + 0.08x_i)]x_i \\ &\quad + 66.7(M_{i+1} + 0.08x_{i+1})x_{i+1}\}/Y_i \\ &\quad (i = 2, 3, 4, 5) \\ \dot{x}_6 &= \{40.8x_5 - [40.8 + 66.7(M_6 + 0.08x_6)]x_6\}/Y_6 \\ &\quad + 66.7(M_7 + 0.08u_2)u_2/Y_6 \end{aligned} \tag{42}$$

with

$$Y_i = (M_i + 0.16x_i) + 75 \quad (i = 1, \ldots, 6)$$

The vector **M** is

$$\mathbf{M} = \begin{bmatrix} M_1 \\ M_2 \\ M_3 \\ M_4 \\ M_5 \\ M_6 \\ M_7 \end{bmatrix} = \begin{bmatrix} 0.7358 \\ 0.7488 \\ 0.7593 \\ 0.7677 \\ 0.7744 \\ 0.7797 \\ 0.7838 \end{bmatrix} \tag{43}$$

Apparent linearization of these equations yields the **A** and **B** matrices

$$\mathbf{A}(\mathbf{x}) = \begin{bmatrix} A_{11}(x_1) & A_{12}(x_1, x_2) & 0 & 0 & 0 & 0 \\ A_{21}(x_2) & A_{22}(x_2) & A_{23}(x_2, x_3) & 0 & 0 & 0 \\ 0 & A_{32}(x_3) & A_{33}(x_3) & A_{34}(x_3, x_4) & 0 & 0 \\ 0 & 0 & A_{43}(x_4) & A_{44}(x_4) & A_{45}(x_4, x_5) & 0 \\ 0 & 0 & 0 & A_{54}(x_5) & A_{55}(x_5) & A_{56}(x_5, x_6) \\ 0 & 0 & 0 & 0 & A_{65}(x_6) & A_{66}(x_6) \end{bmatrix}$$

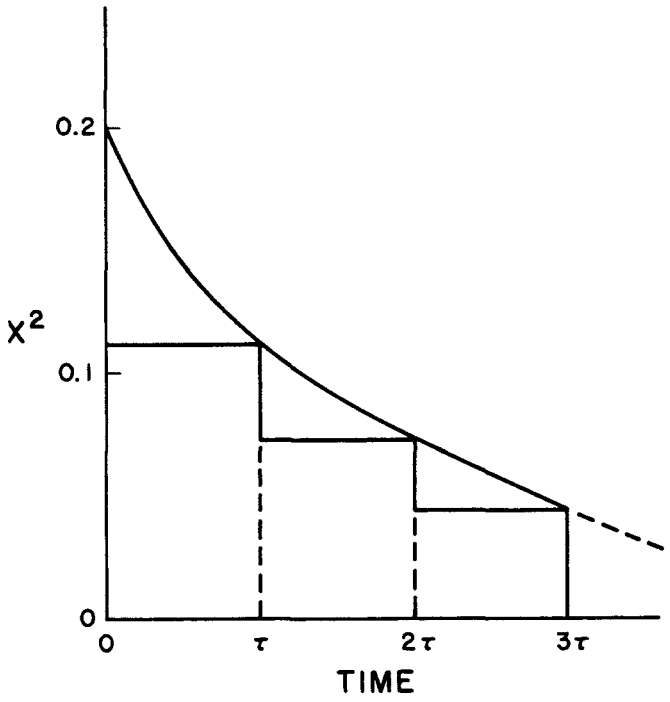

Fig. 1. x^2 vs. time, t, with areas under the curve delineated.

and

$$\mathbf{B}(\mathbf{x}, \mathbf{u}) = \begin{bmatrix} 40.8/Y_1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 66.7(M_7 + 0.08u_2)/Y_6 \end{bmatrix}$$

with

$$A_{ii}(x_i) = -[40.8 + 66.7\,(M_i + 0.08x_i)]/Y_i \qquad (i = 1, \ldots, 6)$$

$$A_{i,i+1}(x_i, x_{i+1}) = 66.7\,(M_{i+1} + 0.08x_{i+1})/Y_i \qquad (i = 1, \ldots, 5)$$

$$A_{i,i-1}(x_i) = 40.8/Y_i \qquad (i = 2, \ldots, 6)$$

The initial conditions associated with this system are given by

$$\mathbf{x}(0) = \begin{bmatrix} -0.0342 \\ -0.0619 \\ -0.0837 \\ -0.1004 \\ -0.1131 \\ -0.1224 \end{bmatrix} \qquad (44)$$

Once again, the control objective is to minimize a quadratic index (with $\mathbf{Q} = \mathbf{I}$ and $\mathbf{R} = \mathbf{I}$) with optimal control achieved via quasilinearization. Values of $t_f = 10$ and 20 were used along with a τ of 0.2. Whenever the absorber system is encountered, it will be understood that all times given are in minutes.

NUMERICAL RESULTS

Before presenting the explicit comparisons between the optimal and suboptimal control of the five foregoing test systems, a few general words are in order. In all cases, the weighting matrices in the quadratic performance index are given by $\mathbf{Q}_k = \mathbf{Q} = \mathbf{I}$ and $\mathbf{R}_k = \mathbf{R} = \mathbf{I}$. These specific values are not required, but their use minimizes the detail of results to be presented. Also, even though the problems are treated as fixed final time control systems, the values of $\mathbf{x}(t_f)$ actually achieved are all close to zero, corresponding theoretically to a large value of t_f.

In terms of the figures to be presented, the discrete-time control vector plots require special attention. These control laws come from the L-Q method of optimal control and from all of the suboptimal control methods used in the numerical examples. At the beginning of each sampling interval, the control variables are determined and held constant throughout the interval. However, for the plotting of the control variables vs. time, smooth curves are drawn through the controls at their values at the beginning of every time step.

Computing times mentioned refer to an IBM 7094 computer and are of the order of magnitude of the times of an IBM 1800 on-line computer. Thus, one may extrapolate the results here to those needed for an on-line control system.

Numerical data have been kept to a minimum to conserve space, but the interested reader can refer to Weber (*20*) for explicit values of the states and controls as a function of time.

Finally, we wish to make a point regarding the explicit numerical calculation of the performance index. This is motivated by the fact that some of the optimal control results are based upon a continuous-time index whereas the suboptimal control results are based upon a discrete-time index. For simplicity, consider $\mathbf{x}(t)$ and $\mathbf{u}(t)$ as scalar functions and the indices

$$I_C = \int_0^{t_f} x^2(\lambda)\, d\lambda \qquad (45)$$

and

$$I_D = \tau \sum_{k=1}^{N} x_k^2 = \tau \sum_{k=1}^{N} x^2(k\tau) \qquad (46)$$

Figure 1 shows a hypothetical trajectory of x^2 vs. t for three time steps of equal length τ. Here I_C is the area under the continuous curve, and I_D is the area under the set of steps shown in the figure. Then $I_C - I_D$ is the area between the curve and the set of steps. It is possible to show that, in general, $I_D \leq I_C$ but we shall not pursue this point. Of importance is that once the control is specified, either I_C or I_D can be calculated.

To discuss the suboptimal control computational results we turn to the first linear system. Using $t_f = 5.0$ and $\tau =$

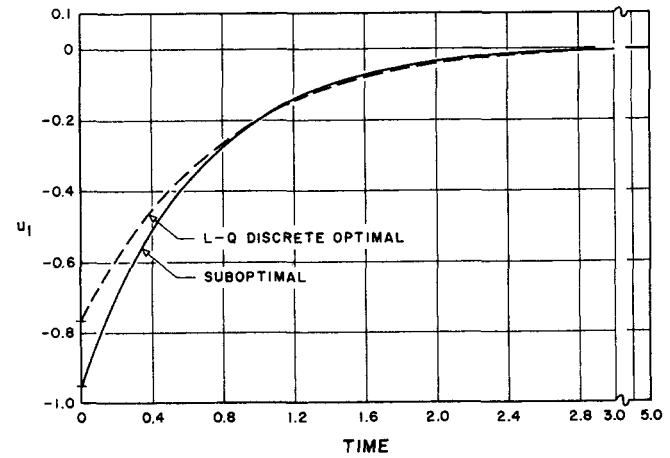

Fig. 2. The L-Q discrete optimal control vs. the FFT suboptimal control for the first system: $u_1(t)$ vs. t.

0.1, we can generate an analytical optimal control result, an optimal result using the standard L-Q discrete algorithm and a suboptimal result using the FFT algorithm. In addition, we have generated a sample and hold answer from the analytical solution. We should emphasize that suboptimal control is not needed here as the system is linear and the optimal control is known; however, it does serve as a convenient test on the suboptimal control. For this calculation, it is assumed that $-1/2\ t$ is known only at the time corresponding to the state just currently calculated.

Table 2 and Figure 2 show some of the results obtained on this test system. As anticipated, the optimal control is better than the suboptimal result; however, the degree of suboptimality is 1.9% on the continuous or the discrete basis. Thus, the suboptimal control behaves quite well although the computing time is slightly higher than for the optimal methods.

From Figure 2 it can be seen that the main place where the suboptimal control and the L-Q discrete optimal control differ is at the beginning of the control period. In the present case we can actually analyze why this difference occurs. For the equation solved each sampling period is

$$\dot{x}_1(t) = -\frac{1}{2} k_\tau x_1(t) + u_1(k_\tau) \qquad (47)$$

However, the coefficient $-1/2\ k_\tau$ is an eigenvalue of this linear equation and, as k goes from 0 to $N-1$, this eigenvalue ranges from 0 to -2.45 (with $\tau = 0.1$ and $N = 50$). Thus the free system corresponding to $u_1 = 0$ becomes more stable with each successive step. Because the suboptimal control cannot anticipate this increase in stability, an overestimate of the amount of control initially required is incurred by the suboptimal procedure. As time increases, the suboptimal control policy approaches the optimal policy.

Next we turn to the second test system with $t_f = 5$ and $\tau = 0.1$. This is much more complex than the first system because the coefficient is a damped sinusoidal function. We have two sets of computational results, these being with the L-Q discrete optimal and the FFT suboptimal algorithms. Figure 3 shows a plot of the corresponding controls for the two modes of operation.

As expected, the two controls differ markedly because of the time-varying coefficient. In fact, the discrete performance indices are 7.075 and 9.187 for the two modes yielding a degree of suboptimality of 29.9%. These differences can be explained via a stability analysis analogous to that for the first system, but we shall not do so here. Computer times are comparable to those for the first system.

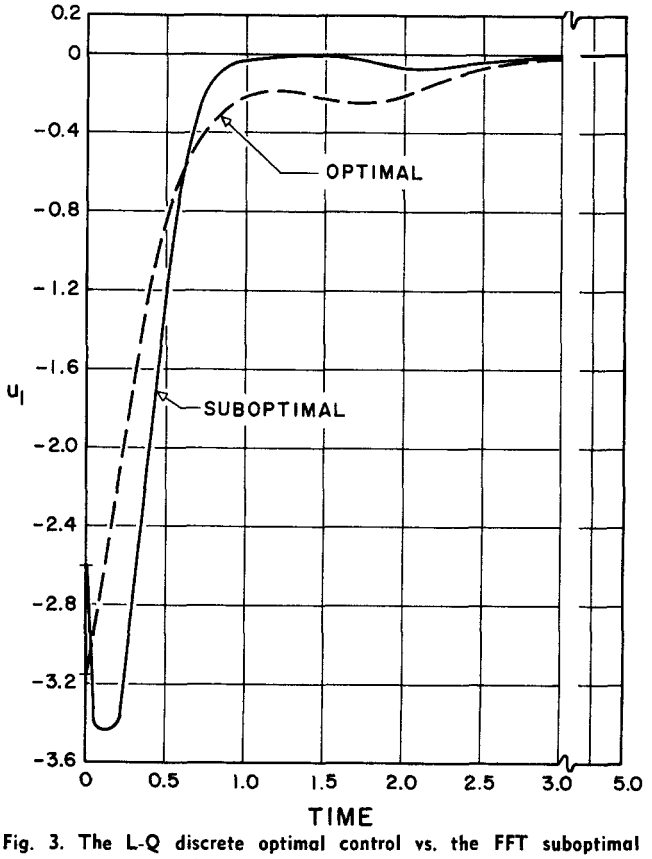

Fig. 3. The L-Q discrete optimal control vs. the FFT suboptimal control for the second system: $u_1(t)$ vs. t.

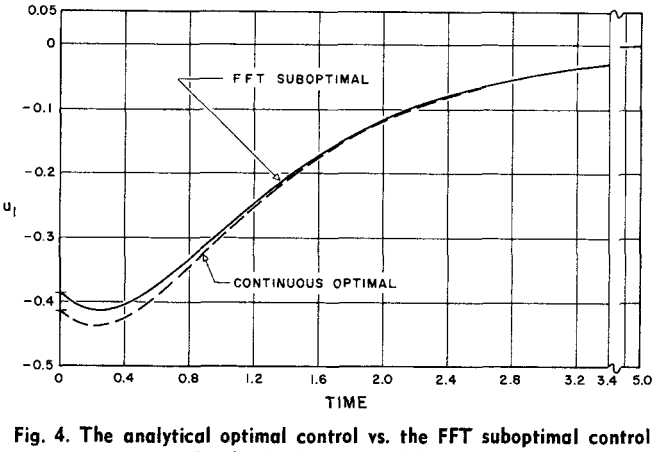

Fig. 4. The analytical optimal control vs. the FFT suboptimal control for the third system: $u_1(t)$ vs. t.

TABLE 2. VALUE OF THE PERFORMANCE INDEX FOR THE FIRST SYSTEM

Method	I_C	I_D	Computing time (sec.)
No control	2.507		
Continuous analytical	1.596	1.565	
Sample and hold analytical	1.599	1.500	75
L-Q optimal	1.597	1.489	
Suboptimal (FFT)	1.626	1.527	70

TABLE 3. VALUE OF THE PERFORMANCE INDEX FOR THE THIRD SYSTEM

Method	I_C	I_D	Computing time (sec.)
No control	2.398		
Continuous analytical	1.295	1.217	
Sample and hold analytical	1.296	1.200	
Suboptimal (FFT)	1.296	1.201	20
Suboptimal (ITER)	1.311	1.215	

The third test system is a nonlinear system, but one for which optimal control by an analytical solution is available (*20*). Using $t_f = 5.0$ and $\tau = 0.1$, we thus have an analytical continuous optimal solution, a sample and hold solution, and two suboptimal modes, namely, the FFT and ITER algorithms. In the ITER mode the iterations were terminated after eight total iterations with the $\mathbf{x}_k$ and $\mathbf{u}_k$ agreeing with the previous values to within 10^{-7}.

Table 3 and Figure 4 show some results of these calculations. The degree of suboptimality is essentially zero

using the FFT algorithm. The ITER algorithm is not quite as good, but the suboptimal results are still excellent.

Having examined the suboptimal control of the three test systems, we analyze the CSTR as described by Equations (38) to (41), using parameter values of $\omega = 0$, 8.9, and 20, t_f values of 5 and 10, and a τ of 0.1. Optimal results were generated using a quasilinearization algorithm whereas suboptimal results were obtained with the FFT and IFT modes. For quasilinearization four and five iterations were needed to achieve convergence when t_f was 5 and 10, respectively.

Figure 5 shows FFT suboptimal results for $\omega = 8.9$ and 20; Figure 6 compares the FFT suboptimal and the quasilinearization optimal results for $\omega = 20$. Next to the points in both figures are shown the corresponding values of the time of control. Table 4 presents performance index values for many of the runs made. As can be seen from all these data, the IFT and FFT algorithms yield identical results. In the case of $\omega = 20$, the suboptimal results agree almost exactly with the optimal results. Thus one concludes that the suboptimal algorithms yield excellent approximations to optimal control for a fairly complicated system. However, the suboptimal is closed loop whereas the optimal is open loop. Incidentally, both types of algorithms require about the same computation time, 45 to 60 sec.

Finally, we examine computational results on the non-

TABLE 4. PERFORMANCE INDICES FOR CSTR AS A FUNCTION OF ω

ω	x(0) Initial condition*	Method of control	t_f	I_C or I_D	Index value
0		IFT	10	I_C	0.009704
0		IFT	5	I_C	0.009704
0		FFT	5	I_C	0.009704
0		FFT	5	I_C	0.009211
8.9	on L.C.	IFT	10	I_C	0.005423
8.9	on L.C.	IFT	5	I_C	0.005423
8.9	on L.C.	FFT	5	I_C	0.005423
8.9	on L.C.	FFT	5	I_D	0.004763
8.9	outside L.C.	IFT	10	I_C	0.010598
8.9	outside L.C.	IFT	5	I_C	0.010598
20		IFT	10	I_C	0.003833
20		IFT	5	I_C	0.003833
20		FFT	5	I_C	0.003833
20		FFT	5	I_D	0.003179
20		Quasilinearization	10	I_C	0.003846
20		Quasilinearization	5	I_C	0.003846

* When $\omega = 8.9$, the initial condition x(0) is either on or outside of the limit cycle (L.C.).

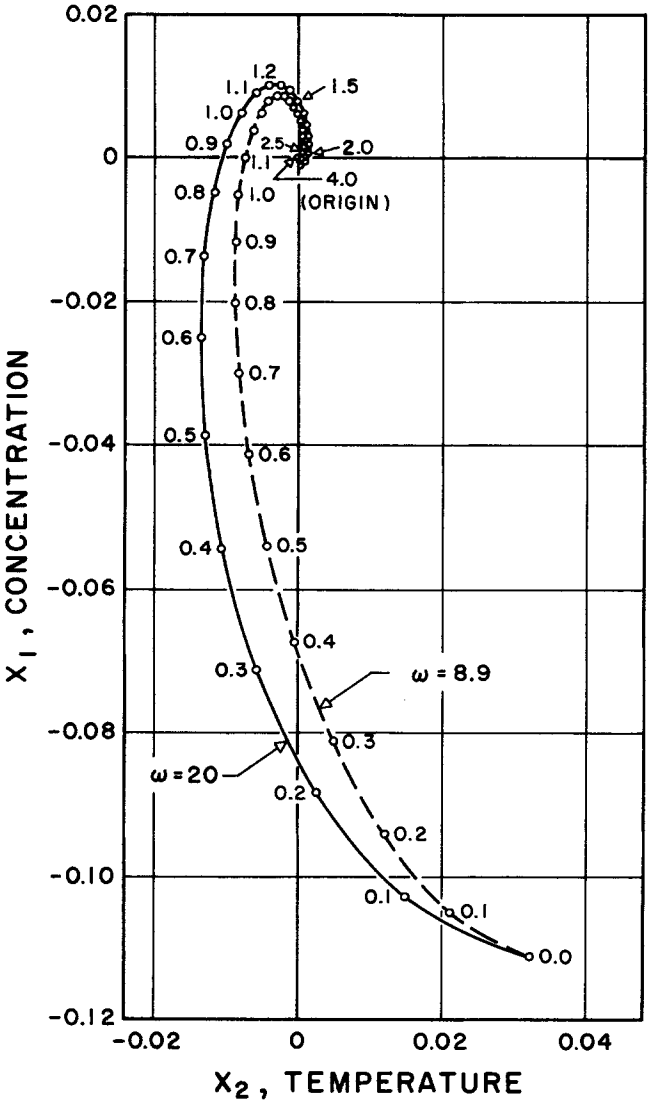

Fig. 5. FFT suboptimal control for CSTR for $\omega = 8.9$ and 20.

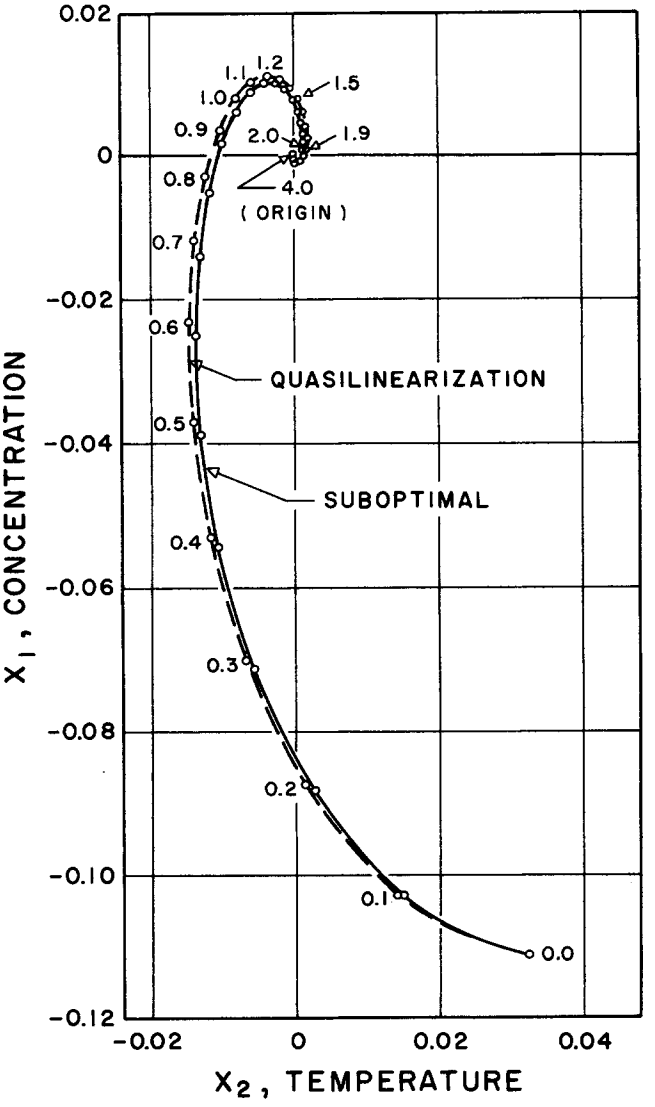

Fig. 6. Optimal and FFT suboptimal control for CSTR for $\omega = 20$.

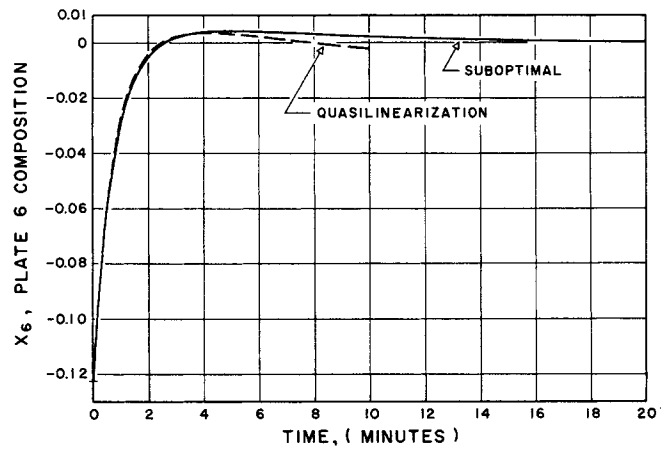

Fig. 7. Optimal and IFT suboptimal control for nonlinear absorber. State 6 vs. time.

linear absorber of Equations (42) to (44). Using $t_f = 10$ and 20 and $\tau = 0.2$, optimal control was achieved with quasilinearization while suboptimal control was achieved with the IFT algorithm. Five iterations were used to achieve the quasilinearization optimal control. The type of results in Figure 7 shows the behavior of x_6, the 6th state variable, as a function of time for the two modes of control generation.

Of primary interest is that for $t_f = 10$, an I_C of 0.1281 was obtained for the quasilinearization mode whereas the corresponding value for the IFT suboptimal control was 0.1284; these, in turn, yield a degree of suboptimality of 0.2%. Thus, as in the CSTR calculations, the suboptimal control yields a feedback type of closed-loop control which provides an excellent approximation to the open-loop optimal control. In a real on-line control system, with partially known parameters in the system model, the suboptimal control is highly preferred.

Interestingly, the quasilinearization results in Figure 7 are given only to $t_f = 10$ whereas the suboptimal control reaches either $t_f = 10$ or $t_f = 20$. This is due to the computational problems associated with quasilinearization when one or more eigenvalues of the linearized equations have large absolute values (*15*). Instability thus results at large values of time. By contrast, the suboptimal control has no such difficulties and can be used with any value of t_f.

CONCLUSIONS

These results show that, except for the special case with a highly time-varying coefficient, the suboptimal control algorithms yield excellent feedback or closed-loop control policies. Such results indicate that the approach is suitable for a real on-line control system as long as constraints on the state or control are not present. In Part II, which follows, we illustrate how these constraints may be included and thus extend the suboptimal algorithms to almost all control situations.

ACKNOWLEDGMENT

The authors wish to acknowledge support of this work from National Science Foundation Grant NSF-GP-2858. Furthermore, this work made use of Princeton University computer facilities supported in part by National Science Foundation Grant NSF-GP-579.

NOTATION

$\mathbf{A}$ = matrix
$\mathbf{B}$ = matrix
$\mathbf{f}$ = vector
g, h, p = scalars
H = Hamiltonian
$\mathbf{I}$ = identity matrix
I = scalar performance index
I_C = continuous performance index
I_D = discrete performance index
k = integer
$\mathbf{K}$ = feedback matrix
$\mathbf{M}$ = vector
N = integer
$\mathbf{P}$ = matrix of Riccati equation
$\mathbf{Q}$ = weighting matrix
$\mathbf{R}$ = weighting matrix
t = time
t_f = final time
$\mathbf{u}$ = control vector
$\mathbf{x}$ = state vector
$\mathbf{y}$ = vector
ω = proportionality constant in CSTR
$\mathbf{\Delta}$ = control matrix
φ = transition matrix
τ = time step size

LITERATURE CITED

1. Aris, R., and N. R. Amundson, *Chem. Eng. Sci.*, **7**, 121, 132, 148 (1958).
2. Baldwin, J. F., and J. H. Sims-Williams, *J. Math. Anal. Applications*, **22**, 523 (1968).
3. Brosilow, C. B., and K. R. Handley, *AIChE J.*, **14**, 467 (1968).
4. Burghart, J. H., *IEEE Trans. Autom. Control*, **14**, 284 (1969).
5. Chant, V. G., and R. Luus, *Can. J. Chem. Eng.*, **46**, 376 (1968).
6. Eller, D. H., and J. K. Aggarwal, *Intern. J. Control*, **8**, 113 (1968).
7. Friedland, B., "A Technique of Quasi-Optimum Control," *JACC Proc.* (1965).
8. Kalman, R. E., and T. S. Englar, "An Automatic Synthesis Program for Optimal Filters and Control Systems," *RIAS Rept.*, Baltimore, Md. (July, 1963).
9. Kokotovic, P., and P. Sannuti, "Singular Perturbation Method for Reducing the Model Order in Optimal Control Design," *JACC Proc.* (1968).
10. Lapidus, L., and R. Luus, "Optimal Control of Engineering Processes," Blaisdell Publishing Co., Waltham, Mass. (1967).
11. McClamroch, N. H., *IEEE Trans. Autom. Control*, **14**, 282 (1969).
12. Meditch, J. S., "A Class of Suboptimal Linear Controls," *JACC Proc.* (1966).
13. Paradis, W. O., and D. D. Perlmutter, *AIChE J.*, **12**, 876 (1966).
14. Pearson, J. D., *J. Electron. Control*, **13**, 453 (1962).
15. Rothenberger, B. F., "Quasilinearization as a Numerical Method," Ph.D. dissert., Princeton University, Princeton, N. J. (1966).
16. Rothenberger, B. F., and L. Lapidus, *AIChE J.*, **13**, 114 (1967).
17. Rubin, O., *IEEE Trans. Autom. Control*, **14**, 737 (1969).
18. Seinfeld, J. H., and K. S. P. Kumar, *Intern. J. Control*, **7**, 417 (1968).
19. Thiriet, L., and A. Deledicq, *Ind. Eng. Chem.*, **60**, No. 2, 23 (1968).
20. Weber, A. P. J., "Suboptimal Control of Nonlinear and Inequality Constrained Control Systems," Ph.D. dissert., Princeton University, Princeton, N. J. (1969).
21. Westcott, J. H. et al., "Approximation Methods in Optimal and Adaptive Control," *Proc. Second IFAC Cong., Basle, 1963*, **2**, 263, Butterworths, London (1964).

Manuscript received January 1, 1970; revision received April 27, 1970; paper accepted April 29, 1970.