Discrete Dynamic Optimization Applied to On-Line Optimal Control

MARSHALL D. RAFAL and WILLIAM F. STEVENS
Northwestern University, Evanston, Illinois

A general method has been developed for controlling deterministic systems described by linear or linearized dynamics. The discrete problem has been treated in detail. Step-by-step optimal controls for a quadratic performance index have been derived. The method accommodates upper and lower limits on the components of the control vector.

A small binary distillation unit was considered as a typical application of the method. The control vector was made up of feed rate, reflux ratio, and reboiler heat load. Control to a desired state and about a load upset was effected.

Calculations are performed quite rapidly and only grow significantly with an increase in the dimension of the control vector. Extension to much larger distillation units with the same controls thus seems practical.

The advent of high-speed computers has made possible the on-line digital control of many chemical engineering processes. In on-line control a three-step procedure is adhered to:

1. Sense the current state.
2. Calculate a suitable control action.
3. Apply this control for a period of time known as the sampling period.

The present study proposes a method for performing step 2. The technique developed is based on linearized dynamics. The strongly nonlinear binary distillation unit provides a suitable system for this study. While much has been published recently (*2*, *3*, *8*) on modeling distillation, little if anything has appeared on the optimal control of such units.

In recent years, a good deal has been published by Kalman, Lapidus, and others (*4* to *7*) on the control of linear or linearized nonlinear systems by minimizing a quadratic function of the states resulting from a sequence of control actions. Their controls are always unconstrained, although the introduction of a quadratic penalty function limits this effect somewhat. The general constrained problem has been treated numerically (*1*) for a single control variable. It was Wanninger (*10*, *11*) who first chose to look at the problem on a one-step-at-a-time basis rather than considering a sequence of controls. However, he made no attempt to solve completely the resulting quadratic programming problem.

The approach taken in the present work is to set up the problem on a one-step basis. This is quite compatible with the on-line digital control scheme. The problem is then shown to be a special case of the quadratic programming problem and as such has a special solution. The particulars concerning the theory underlying the solution scheme and its implementation on a digital computer have been presented (9). In addition, a derivation of the theorems upon which the computational algorithm is based is presented in the Appendix.

The authors wish to be very careful to point out that *optimal*, as used herein, refers only to a single step of control. Even for truly linear systems, the step-by-step optimal control need not be overall optimal. A recent text by Athans and Falb (*1a*) presents both the virtues and defects of such a one-step method. In the present work, the one-step approach is taken because it is amenable to practical solution of the problem and is well suited to nonlinear situations where updating linearization is useful.

THE PROBLEM

The system under consideration is described by a set of matrix differential equations:

$$\dot{X}(t) = AX(t) + BM(t) + \delta(t) \tag{1}$$

Marshall D. Rafal is with Esso Research and Engineering Company., Florham Park, New Jersey.

where A and B are $n \times n$ and $n \times m$ matrices of constants, X and δ are n-dimensional vectors, M is an m-dimensional vector, and $n \geq m$.

Consider the discrete system with a sampling period τ. Suppose $M(\tau)$ is constant on each sampling interval and $\delta = \delta(t)$ is constant for all time. Following a well-known procedure (6), the discrete solution results:

$$X(K+1) = \Phi X(K) + \Delta M(K) + \gamma \qquad (2)$$

where

$$\Phi = e^{A\tau} = \sum_{i=0}^{\infty} \frac{A^i \tau^i}{i!}$$

$$\Delta = \left[\sum_{i=0}^{\infty} \frac{A^i \tau^{i+1}}{(i+1)!} \right] B$$

$$\gamma = \left[\sum_{i=0}^{\infty} \frac{A^i \tau^{i+1}}{(i+1)!} \right] \delta$$

In the following, $X(0)$ will denote an initial or sampled state and $X(1)$ will denote the state resulting from a control action $M(0)$. By translating coordinates, the problem can be set up so that the desired state is the origin. A reasonable criterion function is the minimization of

$$J[M(0)] = X(1)^T QX(1) = [\Phi X(0) + \Delta M(0) + \gamma]^T \times Q[\Phi X(0) + \Delta M(0) + \gamma] \qquad (3)$$

$M(0)$ is further subject to the *physical* constraints

$$M_{\min} \leq M(0) \leq M_{\max} \qquad (4)$$

Given that Q is a positive diagonal matrix, and Δ is of rank m, it readily follows that J is a positive function of the control vector $M(0)$. This follows from the fact that

$$\frac{\partial^2 J}{\partial M^2} = 2\Delta^T Q \Delta = [(2Q)^{1/2} \Delta]^T [(2Q)^{1/2} \Delta] \qquad (5)$$

which, by linear algebra theory, must be positive.

The problem then satisfies all conditions of the general quadratic programming problem. A special algorithm, utilizing the special constraints of Equation (4), was set up, based on the fact that the solution to the problem must lie on a bounding hyperface of the control space unless the unconstrained optimum

$$M(0) = -(\Delta^T Q \Delta)^{-1} \Delta^T Q[\Phi X(0) + \gamma] \qquad (6)$$

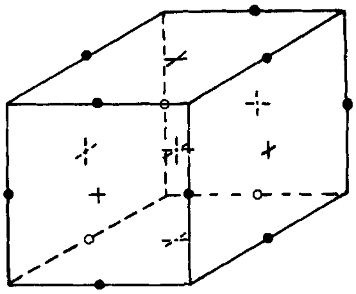
lies within the admissible control space.

Figure 1 depicts a three-dimensional constrained control space. Equation (6) is first used for a guide to which faces and edges of the space must be searched. In fact, subsequent calculations provide information on what further calculations must be done. When a face or edge is searched, a direct calculation is really done. Because of the special set of constraints, any one of the six faces or twelve edges simply requires a calculation of the form of Equation (6), where the minimization is taken with respect to those control parameters free to vary in the face or edge under study. Those controls which are fixed simply become fixed parameters in the analysis.

Details of the algorithm are presented elsewhere (9), and a derivation of its characteristics is given in the Appendix.

SATURATION

The unconstrained optimum of Equation (6) provides some indication of the importance of saturation in these problems.



KEY-

✛ Original linearization point

+ Linearization point with one constrained variable

○ Linearization point with two constrained variables

Fig. 1. Linearization points in a three-dimensional control space.

The form of the control law is

$$M(0) = CX(0) + G \qquad (7)$$

which implies that for two-state variables

$$[C]_{i1} X_1(0) + [C]_{i2} X_2(0) = M_{i\max} \quad i = 1, \ldots, m \qquad (8)$$

$$[C]_{i1} X_1(0) + [C]_{i2} X_2(0) = M_{i\min} \quad i = 1, \ldots, m \qquad (9)$$

where for simplicity $\gamma = 0$.

For a given i, the parallel lines define a strip in the state space. If the initial state lies within this strip, the control component M_i is not saturated in the unconstrained calculation. In Figure 2 two-dimensional control is considered. It should be clear that unless the strips intersect at slight angles, only a small region of the state space will exist where a totally unconstrained control will be optimal. This tends to stress the importance of a scheme for solving the constrained problem.

DISTILLATION MODEL

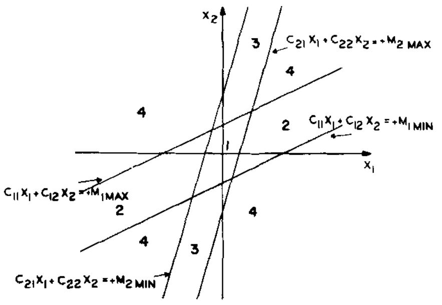
With only a few assumptions, a compact dynamic model may be derived for a binary distillation unit. The work of Huckaba et al. (3) provides a basis for this development.

Figure 3 depicts a three-tray tower with reboiler and condenser. Feed is onto the middle tray and is composed of two species. The state vector for the system is

$$X = (x_1, x_2, x_3, x_4, x_5) \qquad (10)$$

where x_i are light component mass fractions. The control vector is

$$M = (F, R, Q_s) \qquad (11)$$


x2
x1
$C_{21}X_1 + C_{22}X_2 = +M_{2\,MAX}$
$C_{11}X_1 + C_{12}X_2 = +M_{1\,MIN}$
$C_{11}X_1 + C_{12}X_2 = +M_{1\,MAX}$
$C_{21}X_1 + C_{22}X_2 = +M_{2\,MIN}$
1
2
2
3
3
4
4
4
4

1- Both variables unconstrained
2- Variable M_2 constrained
3- Variable M_1 constrained
4- Both variables constrained

Fig. 2. Saturation in two dimensions.

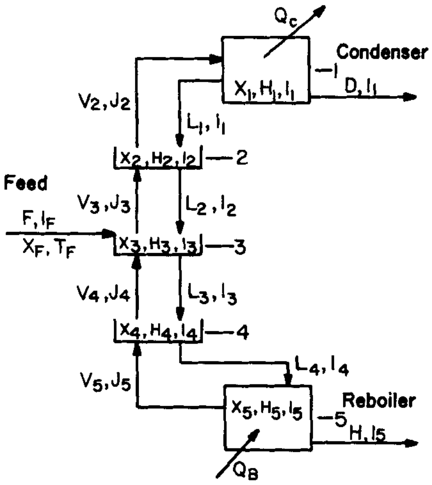


Fig. 3. Tower system.

For simplicity, the assumption is made that the materials are of similar density. The result is that

$$\frac{dH_i}{dt} = 0 \tag{12}$$

The fifteen relevant equations describing the dynamics are then:

Overall material balance:

$$V_{i+1} + L_{i-1} - V_i - L_i + A_i = 0 \tag{13}$$

where

$$L_o = 0,\ L_5 = 0,\ A_1 = -D,\ A_2 = 0,\ A_3 = F,\ A_4 = 0,\ A_5 = -H,\ V_1 = 0,\ V_6 = 0$$

Enthalpy balances:

$$H_i \frac{dI_i}{dt} = V_{i+1} J_{i+1} + L_{i-1} I_{i-1} - V_i J_i - L_i I_i + B_i \tag{14}$$

where

$$B_1 = -DI_1 - Q_c,\ B_2 = 0,\ B_3 = F[I_F - C_p(T_b - T_F)],\ B_4 = 0,\ B_5 = Q_s - HI_5$$

Light component material balances:

$$H_i \frac{dx_i}{dt} = V_{i+1} y_{i+1} + L_{i-1} x_{i-1} - V_i y_i - L_i x_i + C_i \tag{15}$$

where

$$C_1 = -Dx_1,\ C_2 = 0,\ C_3 = Fx_F,\ C_4 = 0,\ C_5 = -Hx_5$$

The equilibrium relation

$$y_i = f(x_i) \tag{16}$$

holds as well.

With the feed a subcooled liquid, there are five degrees of freedom among the twenty variables. The assumption of linear enthalpy relations

$$I_i = ax_i + c \tag{17}$$

$$J_i = by_i + d \tag{18}$$

enables algebraic elimination of all variables but x_i, $i = 1, \ldots, 5$, X_F, T_F, Q_s, F, and $R = L_1/D$.

The resultant dynamics are

$$H_1 \frac{dx_1}{dt} = \frac{E(y_2 - x_1)}{y_2 + \alpha} \tag{19}$$

$$H_2 \frac{dx_2}{dt} = E\left[\frac{y_3 - x_2}{y_3 + \alpha} + \frac{Rx_1 + x_2 - (R+1)y_2}{(R+1)(y_2 + \alpha)}\right] \tag{20}$$

$$H_3 \frac{dx_3}{dt} = E\left[\frac{x_2 - y_3}{y_2 + \alpha} + \frac{x_3 - x_2}{(R+1)(y_2 + \alpha)}\right] + \left[\frac{Q_s}{b - a}\right]\left[\frac{y_4 - x_3}{y_4 + \alpha}\right] + F(x_F - x_3) \tag{21}$$

$$H_4 \frac{dx_4}{dt} = \left[\frac{E}{(R+1)(y_2 + \alpha)} - F\right](x_4 - x_3) + \left[\frac{Q_s}{b - a}\right]\left[\frac{y_5 - x_4}{y_5 + \alpha} + \frac{x_3 - y_4}{y_4 + \alpha}\right] \tag{22}$$

$$H_5 \frac{dx_5}{dt} = \left[\frac{E}{(R+1)(y_2 + \alpha)} - F\right](x_5 - x_4) + \left[\frac{Q_s}{b - a}\right]\left[\frac{x_4 - y_5}{y_5 + \alpha}\right] \tag{23}$$

where

$$\alpha = \frac{d - c}{b - a} \tag{24}$$

$$E = \frac{Q_s - FC_p(T_b - T_F)}{b - a} \tag{25}$$

Piece-by-piece linear equilibrium relations may be developed, giving

$$y_i = k_i x_i + l_i \tag{26}$$

From this model, linear coefficients may be developed so that the relation

$$\frac{dX}{dt} = f(X, M) \tag{27}$$

is cast in the form

$$\frac{dX}{dt} = AX + BM + \gamma \tag{28}$$

where

$$A_{ij} = \left.\frac{\partial f_i}{\partial x_j}\right|_{\substack{X=X_o \\ M=M_o}} \tag{29}$$

TABLE 1. STEADY STATE TOWER PARAMETERS

$F = 125.4$ lb./hr.
$R = 3.79$
$Q_s = 166{,}800$ B.t.u./hr.
$X_F = 0.57$
$T_F = 83$ °F.
$C_p = 1.0$ B.t.u./(lb.) (°F.)
$c = 115.0$ B.t.u./lb.
$d = 349.8$ B.t.u./lb.
$a = -44.0$ B.t.u./lb.
$b = 194.2$ B.t.u./lb.
Upper limits on control: $F - +15.0$ $R - +1.0$ $Q_s - +50{,}000$
Lower limits on control: $F - -15.0$ $R - -1.0$ $Q_s - -50{,}000$

Linearization constants for:	$y_i = k_i x_i + l_i$
$k_1 = 0.4890$	$l_1 = 0.5228$
$k_2 = 0.8197$	$l_2 = 0.2951$
$k_3 = 1.017$	$l_3 = 0.2021$
$k_4 = 1.4393$	$l_4 = 0.0355$
$k_5 = 1.615$	$l_5 = 0.0000$

Tray holdups:
$H_1 = 57.8$ lb.
$H_2 = 5.0$ lb.
$H_3 = 5.0$ lb.
$H_4 = 5.0$ lb.
$H_5 = 94.2$ lb.

TABLE 2. INITIAL OFFSETS—TOWER SYSTEM

$\tau = 0.01$ hr. $Q = I$ Control vector: (F, R, Q_S)

a = No control
b = Boundary control—single linearization
c = Boundary control—updated linearization

Case	Initial offsets	IES*	Steps to origin or offsets at 0.95 hr.
2a		1.023	+0.012, +0.013, +0.011, +0.009, +0.006
2b	+0.05, +0.05, +0.05, +0.05, +0.05	0.0794	73
2c		0.0792	71
3a		0.0414	−0.007, −0.008, −0.007, −0.006, −0.004
3b	−0.01, −0.01, −0.01, −0.01, −0.01	0.0018	48
4a		0.7720	+0.033, +0.036, +0.032, +0.026, +0.018
4b	+0.05, −0.05, +0.05, −0.05, +0.05	0.1479	69
4c		0.0509	67
5a		0.1226	+0.012, +0.013, +0.011, +0.009, +0.006
5b	+0.05, 0.00, 0.00, 0.00, 0.00	0.1248	73
7a		0.0043	+0.001, +0.001, +0.001, +0.001, +0.001
7b	0.00, 0.00, +0.05, 0.00, 0.00	0.0025	3
9a		0.3621	+0.022, +0.025, +0.022, +0.018, +0.013
9b	0.00, 0.00, 0.00, 0.00, +0.05	0.0163	39
11a		13.22	+0.121, +0.140, +0.128, +0.108, +0.078
11b	−0.22, −0.04, +0.14, +0.28, +0.38	6.13	+0.004, 0.000, 0.000, 0.000, 0.000

*IES - sum of discrete error square contributions.

$$B_{ij} = \left.\frac{\partial f_i}{\partial M_j}\right|_{\substack{X=X_o\\ M=M_o}} \tag{30}$$

$$\gamma_i = \left. f_i \right|_{\substack{X=X_o\\ M=M_o}} \tag{31}$$

The forty coefficients described by Equations (29) to (31) will not be presented owing to their ponderous nature.

The general assumptions made above were:

1. There is no holdup variation in the liquid phase. For the reboiler and condenser this implies liquid level control. Material of similar density is also implied.
2. No holdup in the vapor phase occurs.
3. There is a linear enthalpy relationship for both the liquid and vapor phases.
4. A sequence of linear approximations serves to describe the equilibrium.
5. There is adequate condenser surface to condense all vapors from the top tray.
6. There is uniform composition in each phase at each stage.
7. The stages are ideal; Murphree efficiency is equal to unity.
8. Tower operation is adiabatic.
9. Feed is a subcooled liquid and reflux is liquid at the boiling point.
10. The pressure throughout the tower is atmospheric.
11. Suitable heat capacity and boiling point models are available.

LINEARIZATION TECHNIQUES

In controlling a nonlinear system, a model of the form (27) is linearized about $X = X_o$ and $M = M_o$, yielding

$$\frac{dX}{dt} = f(X_o, M_o) + \left.\frac{\partial f}{\partial X}\right|_{\substack{X=X_o\\ M=M_o}} \overline{X} + \left.\frac{\partial f}{\partial M}\right|_{\substack{X=X_o\\ M=M_o}} \overline{M} \tag{32}$$

where

$$\overline{X} = X - X_o$$
$$\overline{M} = M - M_o$$

Two means of choosing X_o and M_o are presented below.

1. It is assumed that an off-line calculation has indicated a particular state X_d at which to operate the system. M_d is the corresponding control vector which, in the absence of external disturbance or initial offset, will hold the system at X_d. Linearization for the unconstrained calculation is done with $X_o = X_d$ and $M_o = M_d$. If a search of several boundaries is necessary, linearization is done about $X_o = X_d$ and M_o broken down as follows. Those variables constrained at extreme values on the boundary contribute the corresponding extreme values to M_o. The variables free to vary between their usual limits contribute the corresponding components of M_d to M_o.

2. Control is effected by updating the linearization at each step. In other words, $X_o = X(0)$ and $M_o = M_d$ for the first step, but thereafter $X_o = X(0)$* and M_o is the control used on the previous step. Linearization in a hyperface proceeds as before except that those control variables not fixed at an extreme in the hyperface contribute components to M_o in the same manner as just described for the unconstrained case.

By previous analysis

$$\overline{M}(0) = C\overline{X}(0) + G \tag{33}$$

where C and G depend on the linearized coefficients. In Method 1 above, a single C and G correspond to each hyperface independent of the current state. Storage of these matrices for future use as they are calculated becomes a significant economic factor. Method 2, which updates X_o, does not permit this economy. Computationally, then, the first method is more attractive.

Hereafter, Method 1 will be called a *single linearization technique.* The second method will be called *updated linearization.* Refer to Figure 1 for the locations of linearization points.

RESULTS

Table 1 presents the array of parameters used to define a steady state operation for the tower. From these values, a Newton-Raphson technique applied to the steady state

*Since a one-step procedure is being used, the resultant state $X(1)$ from a control $M(0)$ becomes the initial condition $X(0)$ for the next step.

Table 3. Step Upsets

$\tau = 0.01$ hr. $Q = I$ Control vector: (F, R, Q_S)

Steady state: $x_F = 0.057$ $T_F = 83$

No initial offsets

a = No control
b = Boundary control—updated linearization
c = Boundary control—single linearization

Case	x_F	T_F	*IES*	Offsets at 0.95 hr.
1a			0.3104	(+0.034, +0.047, +0.048, +0.035, +0.021)
1b	0.62	83	0.0019	(0.000, 0.000, +0.002, −0.003, −0.003)
1c			0.0017	(0.000, −0.001, +0.002, −0.002, −0.003)
2a			0.0003	(−0.001, −0.002, −0.002, −0.001, −0.001)
2b	0.57	93	3×10^{-7}	(0.000, 0.000, 0.000, 0.000, 0.000)
2c			3×10^{-7}	(0.000, 0.000, 0.000, 0.000, 0.000)
5a			0.2900	(+0.033, +0.046, +0.047, +0.033, +0.020)
5b	0.62	93	0.0018	(0.000, 0.000, +0.002, −0.003, −0.003)
5c			0.0017	(0.000, −0.001, +0.002, −0.002, −0.003)
6a			0.2890	(−0.033, −0.046, −0.047, −0.033, −0.020)
6b	0.52	73	0.0018	(+0.001, +0.001, −0.003, +0.002, +0.003)
6c			0.0018	(0.000, 0.000, −0.003, +0.002, +0.003)

form of Equations (19) to (23) yields a state vector:

$$X_d = (0.7931, 0.6076, 0.4336, 0.2929, 0.1869)$$

Studies were then made on the motion of the system:

1. Initially offset from X_d. This is a set point upset.
2. Originally at X_d, but subject to a step change in some input parameter. This is a load upset.

Three means of dealing with the system were compared:

1. No control was effected. The dynamic response was simulated by a fourth-order Kutta-Runge technique.
2. Control effected by the methods previously described with a single linearization approach used.
3. Same as 2, but with an updated linearization approach used.

Tables 2 and 3 show results of studies based on the two kinds of upsets considered. These results are supplemented by Figures 4 to 6, which present a continuous description of some particular cases. The first study deals with initial offsets in a variety of patterns, all yielding a controlled response far more satisfactory than the uncontrolled. Particularly interesting is the subject of Figure 5, which is the simulation of a start-up operation. The compositions on each tray are originally the same as that of the feed to the tower. A satisfactory response under control results in about 20 min., while several hours would be needed for the uncontrolled response to achieve the same.

Response in the reboiler and condenser was sluggish relative to that of the individual trays, due to the larger holdup in these sections. The condenser composition, which is unaffected directly by the reflux ratio, is particularly hard to control relative to the others.

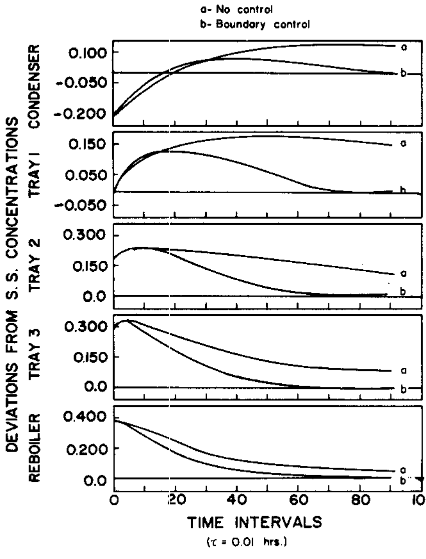


Fig. 5. Tower offset of (—0.223, —0.038, +0.136, +0.277, +0.383). *a* = no control; *b* = boundary control.

Saturation of control was seen to be quite significant. Reboiler or condenser offsets of more than 0.005 units seemed to require saturated control parameters.

Several cases were subjected to updated linearization. The results show little, if any, improvement. A more complete discussion of this comparison is given below.

Table 3 gives results for load upsets. Step upsets in feed composition and feed temperature were considered. Updated control is pursued more thoroughly than in the study of set point upsets.

With either linearization technique, in controlling a load upset the desired state was approached with the result that the system settled at a steady state with the following properties: relatively small offset from the desired state; an optimal control which dictates the repeat of the prior control, holding the steady state; a lack of any potential improvement under the criterion function.

Some further attention should be given to these results. The fact that the desired state could not be achieved is to be expected in a system where the state dimension exceeds that of the control. The steady state system corresponding

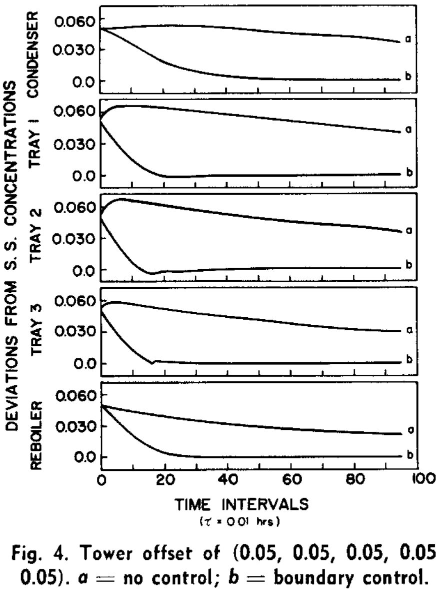


Fig. 4. Tower offset of (0.05, 0.05, 0.05, 0.05 0.05). *a* = no control; *b* = boundary control.

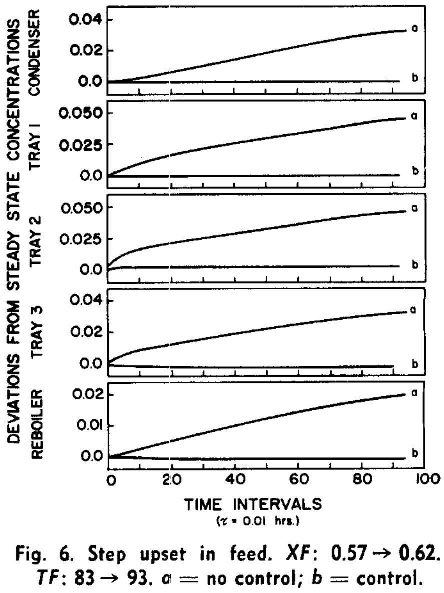


Fig. 6. Step upset in feed. *XF*: 0.57→ 0.62. *TF*: 83→ 93. *a* = no control; *b* = control.

to the desired state is modeled by n algebraic equations in m controls (variables). Hence, no solution need exist for such a system.

The motivation for updated linearization may not be as strong as a superficial analysis might indicate. Any linearization is best within the region close to the point $X(0)$ being linearized about. An accurate representation of the system in the region of $X(1)$ is desired which will be the best of all possible states resulting from the control action $M(0)$. Since this state is not known *a priori*, unless modest motion from state to state occurs, the current point may be no better for linearization than the region of the desired state. For the tower system, the gross response of the trays of the column tends to minimize the value of updating.

In a set point change, the desired state is eventually reached and the last control actions are best taken with a single linearization approach. The last steps under a load upset, where an offset steady state is reached, seems best updated. The results achieved, however, hardly indicate that this effect is important.

A brief study of computer requirements showed that the CDC-3400 computer took an average of 0.013 min. to obtain a control applied for a period of 0.600 min. With the added economy afforded by a single linearization technique, the average time is reduced to 0.003 min. Most of this is attributable to the first few steps of control, where most of the significant calculation is performed. In both cases, on-line computation is clearly feasible.

CONCLUSIONS

A technique was developed which is applicable to on-line digital control of nonlinear processes. A study made on the model of a distillation tower supports the value attributed to the method.

A careful study of two linearization techniques, and possibly others, should be made with respect to the nature of the particular system. However, the more efficient single linearization scheme seems most desirable. The updated mode of control may be needed only under appreciable load upset.

The deterministic tower model is a useful tool and one applicable to quite general control studies. References 9 and *11* should convince the reader that the computer time per calculation is roughly proportional to the cube of the dimension of the state vector while it is combinatorially dependent on the dimension of the control vector. If we accept this and assume that the tower still has only three controls, a fifteen-tray tower can be handled with current hardware. The authors recognize that the above analysis requires the measurement of all tray compositions. A more practical scheme could possibly be worked out based on measurements at several points in a large tower. One approach would be to predict all compositions from these few, or perhaps lump several trays about a measured tray into a single pseudo tray.

Multicomponent models should provide a basis for analysis similar to that performed for the binary.

ACKNOWLEDGMENT

The authors acknowledge with thanks the computer time made available by the Northwestern Computing Center and the NDEA Title IV Fellowship granted to Dr. M. D. Rafal, both of which made this work possible.

NOTATION

A = linearized coefficient matrix for state variables
a = slope in the linear liquid enthalpy relation
A_i = constant used to develop overall material balance
B = linearized coefficient matrix for control variables
B_i = constant used to develop enthalpy balance
b = slope in the linear vapor enthalpy relation
C = composite constant matrix for optimal control
c = intercept in the linear liquid enthalpy relation
C_p = heat capacity
D = flow rate of condenser effluent
d = intercept in the linear vapor enthalpy relation
E = composite function of parameters
F = feed rate
f = function
G = composite constant matrix for optimal control
H = flow rate of reboiler effluent
H_i = holdup at the i^{th} stage
I_i = enthalpy of liquid on i^{th} tray
J = criterion function
J_i = enthalpy of vapor from i^{th} tray
k_i = slope in the linear equilibrium relation
L_i = flow rate of liquid from i^{th} tray
l_i = intercept in the linear equilibrium relation
M = control vector
Q = positive diagonal matrix
Q_c = condenser heat load
Q_s = reboiler heat load
R = reflux ratio
t = time
T_b = boiling point
T_F = feed temperature
V_i = flow rate of vapor from i^{th} tray
X = state vector
X_F = feed composition
x_i = liquid composition on the i^{th} tray
y_i = vapor composition on the i^{th} tray

Greek Letters

α = composite tower constant
γ = constant in solution of matrix differential equation
Δ = coefficient of control in solution of matrix differential equation
ζ = zero-order linearization term
τ = sampling interval
Φ = coefficient of state in solution of matrix differential equation

LITERATURE CITED

1. Deley, G. W., and G. Franklin, *J. Basic Eng.*, 87, 81 (1965).
1*a*. Athans, M., and P. L. Falb, "Optimal Control, An Introduction to the Theory and Its Applications," Chap. 10, p. 824 ff, McGraw-Hill, New York (1966).
2. Duffin, J. H., and J. D. Gamer, paper presented at AIChE Dallas meeting, Tex. (1966).
3. Huckaba, C. E., E. R. Franke, and E. P. May, *Chem. Eng. Progr. Symp. Ser. No. 46*, 59, 38 (1963).
4. Kalman, R. E., *J. Basic Eng.*, 82, 35 (1960).
5. ———, and R. W. Koepcke, "Proceedings of the Western Joint Computer Conference," p. 107, Institute of Radio Engineers, New York (1959).
6. Kalman, R. E., Leon Lapidus, and Eugene Shapiro, *Chem. Eng. Progr.*, 56, 55 (1960).
7. Lapidus, L., Eugene Shapiro, Saul Shapiro, and R. E. Stillman, *AIChE J.*, 7, 288 (1961).
8. Moczek, J. S., R. E. Otto, and T. J. Williams, *Proc. Sec. Congr. IFAC*, 2, 238 (1963).
9. Rafal, M. D., Ph.D. thesis, Northwestern Univ., Evanston, Ill. (1966).
10. Stevens, W. F., and L. A. Wanninger, *Can. J. Chem. Eng.*, 44, 158 (1966).
11. Wanninger, L. A., Ph.D. thesis, Northwestern Univ., Evanston, Ill. (1965).

Manuscript received August 8, 1966; revision received March 20, 1967; paper accepted April 11, 1967. Paper presented at AIChE Houston meeting.

APPENDIX: DERIVATION OF THE ALGORITHM

Use is made of several well-known theorems of linear algebra.

Lemma 1: An extremum of a function $f(x)$ is minimum if the matrix of second partial derivatives is positive definite when evaluated at the extremum.

Lemma 2: any positive definite matrix has a unique nonnegative square root. Furthermore, if diagonal, the elements are just the positive square roots of the original elements.

Lemma 3: If B is an $n \times m$ matrix of rank m, where $n \geq m$, then $B^T B$ is a positive matrix.

Lemma 4: A diagonal matrix N has the property $N = N^T$.

Consider the positive diagonal matrix Q, appearing in the criterion function of Equation (3). By Lemma 2

$$Q = N^2 = NN$$

where N is also diagonal. Thus, by Lemma 4

$$Q = N^T N$$

and Equation (5) becomes

$$\frac{\partial^2 J}{\partial M^2} = 2\Delta^T N^T N\Delta = 2(N\Delta)^T N\Delta$$

By Lemma 3, $\partial^2 J/\partial M^2$ is positive, provided that $N\Delta$ is of rank m. This implies that Δ must be of rank m, as stated previously. Finally, Lemma 1 indicates that the extremum is a minimum. In addition, since the Hessian matrix is a positive definite form composed of constants independent of position, the response surface is some sort of a hyperbowl.

It is important to recognize that the above development identifies the problem as a special case of the general quadratic programming problem. The general problem requires extremizing a function

$$f(\beta, M) = \beta p M^T + \frac{1}{2} M^T R M$$

subject to

$$\beta \geq 0 \quad M \geq 0 \quad BM = b$$

where R is an $m \times m$ positive semidefinite matrix $p = (p_1, p_2, \ldots, p_n)$, and β is a scalar.

The special case presented in the early part of this paper does not include a β parameter and has positive definite R equal to $\Delta^T Q \Delta$. The principal difference is the very special set of constraints, as given in Equation (4), which can be put in the form required by the general problem by translation and addition of slack variables. The special computational algorithm used for solution of the problem, as presented in Equations (3) and (4), requires two key theorems which are presented and proved below.

Theorem 1

Let $Z = Z(M) = Z(M_1, M_2, \ldots, M_n)$ be a positive quadratic function. The response (function value) along any straight line in the control space is concave up (positive).

Proof

Since Z is positive, there must exist some linear coordinate transformation $N = PM$ such that the Hessian matrix is diagonal and composed only of positive constants:

$$\frac{\partial^2 Z}{\partial N_i \partial N_j} = \begin{cases} a_i > 0 & i = j \\ 0 & i \neq j \end{cases}$$

Any straight line in the hyperspace may be characterized by the one parameter family

$$N_i = \lambda_i N_1 + k_i \quad (i = 2, 3, \ldots, n)$$

where λ_i and k_i are constants determined by the orientation of the line in the hyperspace. By chain rule differentiation

$$\frac{dZ}{dN_1} = \frac{\partial Z}{\partial N_1} + \sum_{i=2}^{n} \frac{\partial Z}{\partial N_i} \lambda_i$$

and

$$\frac{d^2 Z}{dN_1^2} = \frac{\partial^2 Z}{\partial N_1^2} + 2\sum_{i=2}^{n} \frac{\partial^2 Z}{\partial N_1 \partial N_i} \lambda_i + \sum_{i=2}^{n} \sum_{j=2}^{n} \frac{\partial^2 Z}{\partial N_j \partial N_j} \lambda_i \lambda_j$$

or, finally

$$\frac{d^2 Z}{dN_1^2} = a_1 + \sum_{i=2}^{n} \lambda_i^2 a_i > 0$$

The desired proof of Theorem 1 follows from this fact.

Theorem 2

Let $Z = Z(M) = Z(M_1, M_2, \ldots, M_n)$ be a positive quadratic function. Let $Z^* = Z^*(M^*) = Z^*(M_1^* M_2^*, \ldots, M_n^*)$ be the *unconstrained* minimum of Z. Consider the $2n$ constraints

$$M_{\min} \leq M \leq M_{\max}$$

Reorder and partition M^* such that $M^* = (M^*_{\min}, M^*_{\max}, M^*_a)$, where

(1) $M_i^* \subset M_a^*$ if and only if $M_{\max} > M_i^* > M_{\min}$ (true for $i = 1, 2, \ldots, k$)

(2) $M_i^* \subset M^*_{\max}$ if and only if $M_i^* \geq M_{i,\max}$ (true for $i = k+1, k+2, \ldots, k+j$)

(3) $M_i^* \subset M^*_{\min}$ if and only if $M_i^* \leq M_{i,\min}$ (true for $i = k+j+1, k+j+2, \ldots, n$)

The constraint set defines a hyperbox in the n space. The *constrained* minimum must

(*a*) Equal the unconstrained minimum if $M_a^* = M^*$

(*b*) Lie on a boundary if $M_a^* \neq M^*$

In addition

(*c*) The only faces which may contain the constrained minimum are those $n - k$ defined by the right-hand side of inequalities in (2) and (3) above.

Proof

Part (*a*) of the Theorem is obvious. If the global optimum is admissible, it is certainly the locally constrained optimum.

Part (*b*) follows directly from Theorem 1. Suppose that an interior point P_1 is the constrained optimum. Under the conditions of (*b*), a straight-line segment from M^* to this point must pierce a bounding face at some admissible point P_2. If this is so, Theorem 1 says

$$Z(P_1) < Z(P_2) < Z(M^*)$$

and the assumption of an optimum in the interior has led to a contradiction.

Part (*c*) is somewhat more involved. Consider an element M_t^* of $M^*_{\max}$. By definition $M_t^* > M_{t,\max}$. We must show that the bounding surface in the plane $M_t = M_{t,\min}$ cannot contain the unconstrained optimum.

Two possibilities exist. In the first case, all other constraints are satisfied at M^*, such that

(4) $M_{i,\max} \geq M_i \geq M_{i,\min} \quad i \neq t$

Consider a line from M^* to any potential optimum on the surface. To be eligible for the optimum, (4) must be satisfied at this point. By the property of linearity, it can be deduced that (4) must hold for all points on this line segment, since it holds at the two end points. However, from the hypothesis $M_t^* > M_{t,\max}$, the surface $M_t = M_{t,\max}$ must be pierced by this line segment. But, by the argument just completed, this point of intersection will be a candidate for the optimum and by Theorem 1 must yield a better result than the point on the surface $M_t = M_{t,\min}$. Similarly, all such points are eliminated.

The other case is where one or more of the conditions of (4) do not hold, and thus a line from M^* to $M_t = M_{t,\min}$ may pierce $M_t = M_{t,\max}$ at a point not satisfying these inequalities. But this implies that a group of controls has $M_k > M_{k,\max}$ or $M_k < M_{k,\min}$ at this point, called P'. The segment from P' to the point in $M_t = M_{t,\min}$ which is under consideration will, of course, satisfy the constraints in t. However, since all the conditions of (4) must be satisfied at $M_t = M_{t,\min}$ for the point in question, some boundary of all variables with the property of M_k must be crossed, in turn, before arriving at the surface $M_t = M_{t,\min}$ from the point M^*. The last such surface pierced is clearly a candidate for the bounded optimum, since all constraints are satisfied. By Theorem 1, this point must give a better response than the one under consideration. Hence, $M_t = M_{t,\min}$ need not be considered.

By similar arguments, the following facts may be established. Given that $M_t < M_{t,\min}$ in the unconstrained optimum, $M_t = M_{t,\max}$ can be eliminated as a candidate for the constrained optimum. Given that $M_{t,\min} < M_t < M_{t,\max}$ in the unconstrained optimum, both $M_t = M_{t,\min}$ and $M_t = M_{t,\max}$ can be eliminated as candidates for the constrained optimum. In addition, the same sort of reasoning rules out all interior points of the admissible space unless such a point is the unconstrained optimum.

Application of these two theorems results in the following conclusion. The only points which need to be searched to locate the constrained optimum are those contained in the hyperfaces corresponding to the subsets $M^*_{\max}$ and $M^*_{\min}$. This leads to the conclusion that the required boundary search for the optimum on any hyperface or edge of the constrained space is really a direct calculation, rather than a search. On any such $n - k$ dimensional surface, k control variables are fixed parameters at their extreme values. The remaining $n - k$ variables which are free over a range of admissible values now make up a new $n - k$ dimensional control vector, and an unconstrained optimum in such a surface may then be found directly. Reference 9 gives details of the necessary computer programs for those readers who may be interested in details.