STUDIES IN THE OPTIMUM DESIGN, CONTROL AND OPERATION OF CHEMICAL PLANT

Published work submitted for the degree of

DOCTOR OF SCIENCE

of

THE UNIVERSITY OF EDINBURGH

by

ROY JACKSON, M.A.

Introduction

The papers in this collection are concerned with problems of optimum design, control and operation in chemical plants. The order of presentation is not chronological; instead the papers are divided into four groups, each concerned with a different aspect of the subject, as follows:-

Group A

Optimization problems in the automatic control of chemical plant.

Group B

The optimum design and operation of systems of interconnected units.

Group C

Variational optimization problems in chemical reactor design.

Group D

Optimum continuous distillation.

The majority of the publications report the independent work of the present writer. However, publication C1 has a postgraduate student as co-author, while publications A3, B2, B3, B4, C8 and D1 have more senior colleagues as co-authors. The relative contributions of the present writer and his co-workers are described in the introductions to the separate groups of papers.

The publications of Group A describe work carried out while the writer was employed by Imperial Chemical Industries Ltd. Publications C4, C5, C6 and C8 describe work carried out during the tenure of a Visiting Professorship at Rice University, Houston, Texas, in the academic year 1965-66, and all the remaining publications describe work carried out at the University of Edinburgh.

Index to Publications

Group A

A1. Calculation of process controllability using the error-squared criterion.
Transactions of the Society of Instrument Technology, 10, 1958, p.68.

A2. A non-linear theory of the dynamical behaviour of pneumatic devices.
Transactions of the Society of Instrument Technology, 10, 1958, p.161.

A3. The behaviour of linear systems with inputs satisfying certain bounding conditions. (with B. J. Birch)
Journal of Electronics and Control, 6, 1959, p.366.

A4. The design of control systems with disturbances satisfying certain bounding conditions, with application to simple level control systems.
Proceedings of 1st Congress of the International Federation of Automatic Control, Moscow, 1960.

A5. Optimum sampled-data control.
Institution of Electrical Engineers Monograph No. 426M., 1961.
also: Proceedings of the Institution of Electrical Engineers, 108C, 1961, p.309.

Group B

B1. Comments on the paper: Optimum cross-current extraction with product recycle.
Chemical Engineering Science, 18, 1963, p.215.

B2. The discrete maximum principle (with F. Horn)
Industrial and Engineering Chemistry (Fundamentals), 4, 1965, p.110.

B3. The/

B3. The discrete maximum principle. (with F. Horn)
Industrial and Engineering Chemistry (Fundamentals), 4, 1965, p.487.

B4. On discrete analogues of Pontryagin's maximum principle. (with F. Horn)
International Journal of Control, 1, 1965, p.389.

B5. Some algebraic properties of optimization problems in complex chemical plants.
Chemical Engineering Science, 19, 1964, p.19.

B6. A generalised variational treatment of optimization problems in complex chemical plants.
Chemical Engineering Science, 19, 1964, p.253.

B7. A variational solution of unsteady state optimization problems in complex chemical plants.
Chemical Engineering Science, 20, 1965, p.405.

Group C

C1. Optimum temperature profiles in tubular reactors: an exploration of some difficulties in the use of Pontryagin's maximum principle. (with I. Coward)
Chemical Engineering Science, 20, 1965, p.911.

C2. Optimum startup procedures for an autothermic reaction system.
Chemical Engineering Science, 21, 1966, p.241.

C3. Optimum temperature gradients in tubular reactors with decaying catalyst.
A.I.Ch.E. - I.Chem.E. Symposium Series No. 4, 1965, p.33.

C4. An approach to the numerical solution of time-dependent optimization problems in two-phase contacting devices.
Transactions of the Institution of Chemical Engineers, 45, 1967, p. T160.

C5. Optimization problems in a class of systems described by hyperbolic partial differential equations. Part I. Variational theory.
International Journal of Control, 4, 1966, p.127.

C6. Optimization problems in a class of systems described by hyperbolic partial differential equations. Part II. A maximum principle.
International Journal of Control, 4, 1966, p.585.

C7. The optimal use of mixed catalysts for two successive chemical reactions.
Journal of Optimization Theory and Applications (in press)

C8. Reactor optimization problems for reversible exothermic reactions.
(with D. C. Dyson, F. Horn, and C. B. Schlesinger)
Canadian Journal of Chemical Engineering (in press)

Group D

D1. Energy requirements in the separation of mixtures by distillation.
(with J. R. Flower)
Transactions of the Institution of Chemical Engineers, 42, 1964, p. T249.

Group A

The publications of this group describe various general problems in the theory of automatic control, which arose out of specific problems encountered in designing control systems for projected chemical plants.

ˡhe quality of control which it is possible to attain in a given piece of equipment depends partly on the sophistication of the controlling device used, but is also inherently limited by the dynamic properties of the piece of plant to be controlled. Within a given class of controlling devices it should therefore be possible to define in a quantitative manner the controllability of a certain section of plant, representing the best control quality obtainable with the given plant and controlling devices from the specified class. Publication A1 is concerned with the development of a quantitative definition of controllability on these lines and with methods of calculating this quantity.

The actual devices with which control is implemented in chemical plants are frequently pneumatically operated and publication A2 presents an analysis of the dynamic behaviour of the basic elements from which pneumatic control systems are built up. It had frequently been assumed that the behaviour of these devices was closely analogous to that of electronic amplifiers, so that their response could be assumed to be approximately linear and methods of analysis such as frequency response testing could be applied to them. In publication A2 it is shown that this is not the case, and that their behaviour is inherently non linear and radically different from that of electronic amplifiers in unsteady states.

A failure to control certain variables in chemical plants within strictly defined bounds is frequently disastrous, but the commoner methods of control system design deal with quantities such as the time average of the square of the deviation of a variable from its desired value. Needless to say, control

of/

of such a quantity to a specified value is no guarantee that the variable in question may not take very large values for quite short periods of time, so one is led to seek a method of establishing absolute bounds for the values of the controlled variables in cases where the disturbances affecting the system satisfy certain conditions limiting their size and rate of change. So far as I am aware, publication A3 describes the first successful work on this problem, and publication A4 goes on to apply the theory to the design of systems for controlling the levels of liquids in vessels. This method was adopted as a standard design procedure by Imperial Chemical Industries Ltd. The theoretical development could perhaps be described as a family project in which the writer collaborated with his brother in law, Dr. B. J. Birch, whose contribution, as a mathematician was to the rigorous mathematical formulation of the derivations.

Finally publication A5 arose from the developing interest of the chemical industry in the use of intermittent analytical measurements, such as those obtained from automatic chromatographs or mass spectrometers, for the control of continuously varying quantities. Use is made of the concept of controllability, developed in publication A1, and Wiener's theory of spectral factorisation is applied to investigate the limitations in controllability due to the loss of information inherent in using a sampled signal for control purposes. This paper was awarded a premium by the Institution of Electrical Engineers.

A1

CALCULATION OF PROCESS CONTROLLABILITY USING THE ERROR-SQUARED CRITERION

By

R. Jackson, B.A.

SYNOPSIS

Using the integral or average of the square of the error as a criterion of control quality, the controllability of a loop with a given type of controller may be defined as the optimum quality which can be obtained by adjusting the controller settings. The dependence of the controllability on the parameters of the plant is illustrated by considering a system of three exponential transfer stages with a proportional-plus-integral controller of conventional type. The results, which are displayed graphically in the form of contour charts, are compared with those obtained from other criteria.

63

CONTROL QUALITY AND CONTROLLABILITY

In a feedback control system, the fluctuations of an output variable produced by some disturbance are reduced to an acceptable size by using measurements of the output variable itself to control a correcting variable. As a quantitative measure of what is meant by ' an acceptable size ', it is necessary to adopt some criterion of control quality. The appropriate choice will obviously depend on the process involved, but, in the control of continuous processes, the following two cases cover a large proportion of the possible situations :

(i) The output variable must on no account pass outside certain limits. A good example of this type would be a level-control system with a pumped outflow of liquid. Too high a level would then lead to carry-over of liquid into parts of the system which should contain only gas, while too low a level would lead to loss of suction on the pumps. The appropriate measure of control quality in this case is the magnitude of the maximum deviation of the output from its desired value. It should be noted that it is only possible to specify this if bounds are given for the variation of the disturbance. This case will not be considered further in the present work

(ii) No absolute limitation on the tolerable magnitude of the output disturbance is given, but it is desirable that it should be as small as possible for as large a proportion of the time as possible. This immediately suggests the use of the time integral of some even function of the deviation of the output from its desired value, the usual choice being the integral of the square of this deviation, since this is relatively easy to compute.[1, 2, 6, 8]

The components of the control loop can conveniently be divided into two groups, the plant and the controller (corresponding to the physical division provided by the controller box), since in process control the usual practice is to use a standard type of controller in all applications, relying on the adjustments provided to match its transfer characteristics to those of the remainder of the loop. The control quality obtained with any given arrangement will therefore depend both on those parameters of the plant which determine its dynamic behaviour and on the controller settings. It is also generally true to say that improvements in control quality obtained by adjusting the controller will lead to greater demands on the range and speed of action of the correcting element. Assuming, however, that it is possible to provide equipment capable of applying the correcting signals called for by the controller, it is important to know whether there is any limitation imposed on the attainable control quality by the nature of the plant, or whether it is possible to improve the quality without bound by suitable adjustments of the controller.

The answer to this question depends on the degree of flexibility permitted in the dynamic characteristics of the controller, for it can be shown that if *any* combination of proportional, repeated derivative, and repeated integral terms may be used, it is possible to obtain any control quality desired, whatever the transfer function of the remainder of the loop, provided no true distance-velocity lag is present. However, apart from the fact that a control function of this type is not physically realizable, one is in practice limited to a commercially available type of controller giving, at most, proportional, integral, and derivative terms. Even with this restricted class of controllers, it is possible to improve the control quality without bound for certain simple forms of the plant transfer function, the most obvious case being a single exponential transfer stage, with which a simple proportional controller suffices to give any desired quality. In general, however, the attainable quality is limited, and there exist certain controller adjustments which, with the given plant, give better control quality than any others. This best attainable quality, which is a function of the plant parameters only (for a given class of controllers), may be called the *controllability* of the plant. It is clearly important to know how this quantity depends on the plant para-

Mr. Jackson is with the Billingham Division of Imperial Chemical Industries Ltd.

meters, since this indicates which features of the design impose limitations on the attainable control performance and allows one to evaluate modifications intended to improve this performance.

By minimizing the integral of the square of the error following a unit step disturbance, Hazebroek and van der Waerden[1] determined the optimum settings for a proportional-plus-integral controller and several types of plant transfer function, but the corresponding minimum values of the integrals as functions of the plant parameters are not quoted. It is felt that too much emphasis can be placed on the theoretical prediction of optimum controller settings, since their practical determination hardly ever presents serious difficulties, and the minima, as found for example by the above authors, are often very flat. It is of much greater interest to know how the minimum value varies with the parameters of the *plant*, as this provides the measure of controllability discussed above and gives a basis for comparing the suitability of different designs for automatic control.

In the present paper a simple example is taken to illustrate the calculation of controllability and the type of conclusions which can be drawn from such a calculation. A plant transfer function corresponding to three exponential transfer stages is considered, with the integral of the square of the error (case (ii) above) as the criterion of control quality. This transfer function is sufficiently simple to be perfectly controllable if a three-term controller is considered, so attention is limited to proportional-plus-integral controllers. The commonly used unit step disturbance is considered first, but the treatment can equally easily be applied to a stationary time series used to provide a more realistic representation of actual disturbances. A simple class of stationary disturbances is therefore dealt with to illustrate how the form of the dependence of controllability on the plant parameters alters as the spectrum of disturbance changes.

Although it is possible to evaluate explicitly all the integrals arising in the simple examples discussed here, it should be emphasized that the method is in no way limited to the treatment of cases where this is possible. The basic minimization involved in determining the controllability is carried out on a digital computer, as described in Appendix I, using a subsidiary routine to evaluate the integral which is to be minimized. In the present case this merely has to evaluate an explicit form for the integral, but it could equally well be a routine for evaluating the integral numerically in the case of a more complicated transfer function.

INTEGRAL OF THE SQUARE OF THE ERROR AFTER A UNIT STEP DISTURBANCE

In a simple control loop, $d(t)$ is the disturbance, which affects the output $o(t)$ through a section of plant with transfer function $X(s)$, $Y(s)$ is the transfer function of the part of the plant included in the control loop, and $C(s)$ is the transfer function of the controller. All quantities are measured in terms of potential changes in o, so that $X(s)$, $Y(s)$, $\rightarrow 1$ as $s \rightarrow 0$. When $d(t)$ is a unit step function at $t = 0$, it possesses a Laplace transform $d(s) = 1/s$, and $o(t)$ also possesses a Laplace transform, related to $d(s)$ by the well-known equation :

$$o(s) = \frac{X(s)\,d(s)}{1 + C(s)\,Y(s)} = Z(s)\,d(s) \quad \text{(say)} \quad \ldots\ldots\ldots\ldots\ldots(1)$$

Now it can be shown[1] that, if I is the integral of the square of the output after a step disturbance

$$I = \int_o^\infty o^2(t)dt = \frac{1}{\pi}\int_o^\infty \frac{|X(j\omega)|^2}{\omega^2} \cdot \frac{d\omega}{|1 + C(j\omega)Y(j\omega)|^2} \quad \ldots\ldots(2)$$

which may be given a geometrical interpretation by noting that $C(j\omega)\ Y(j\omega)$, plotted in the complex plane with ω as a parameter, is the Nyquist locus of the system, and $\rho(\omega) = |1 + C(j\omega)Y(j\omega)|$ is the distance of the current point on this locus from $(-1, 0)$. Equation (2) may therefore be written :

$$I = \frac{1}{\pi}\int_o^\infty \frac{W(\omega)d\omega}{\rho^2(\omega)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

with $W(\omega) = |X(j\omega)|^2/\omega^2$. This may be interpreted as an integral along the Nyquist locus of a density $W(\omega)$ divided by the square of the distance of the current point from $(-1, 0)$. It provides a basis for the principle that a Nyquist locus corresponds to a system with good performance provided it keeps well away from the 'danger' region near $(-1, 0)$, and is useful in drawing qualitative conclusions for more complicated systems.* An alternative representation, very convenient for graphical evaluation of the integral has been described by Rosenbrock.[6]

In the particular example considered here, the disturbance affects the output through a single exponential transfer stage of time-constant τ, while the output is fed back through the controller and two additional time-constants, which may be written $m\tau$ and $n\tau$, in series with τ. Thus :

$$X(j\omega) = \frac{1}{1 + j\omega\tau}\ ;\ Y(j\omega) = \frac{1}{(1 + j\omega\tau)(1 + j\omega m\tau)(1 + j\omega n\tau)} \ldots(4)$$

and

$$C(j\omega) = \mu + \alpha j\omega + \beta/j\omega \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

for a conventional three-term controller. Substituting these expressions into equation (2) gives the integral of a rational function of ω, which may be evaluated by a well known method due to Phillips,[2] giving, after some reduction

$$\frac{2I}{\tau} = \frac{A + By + (C - Dy)/\eta}{y(C - Dy) - E\eta} = U \quad \text{(say)} \ldots\ldots\ldots\ldots\ldots(6)$$

where :

$$\left.\begin{array}{llll} A = (m^2 + n^2)(m + n + mn) & & & \\ B = m^2 n^2 & y & = & 1 + \mu \\ C = (1 + m + n + \epsilon)(m + n + mn) & \epsilon & = & \alpha/\tau \\ D = mn & \eta & = & \beta\tau \\ E = (m + n + mn)^2 & & & \end{array}\right\} \ldots(7)$$

It is easy to show that U may be made as small as we please with a three-term controller, as mentioned in the introduction. Limiting attention, therefore, to a proportional-plus-integral controller ($\epsilon = 0$), it is obvious that $I \rightarrow \infty$ on

$$\eta = \frac{y\,(C - Dy)}{E}$$

*This geometrical interpretation was first brought to the author's notice by Dr. P. C. Price.

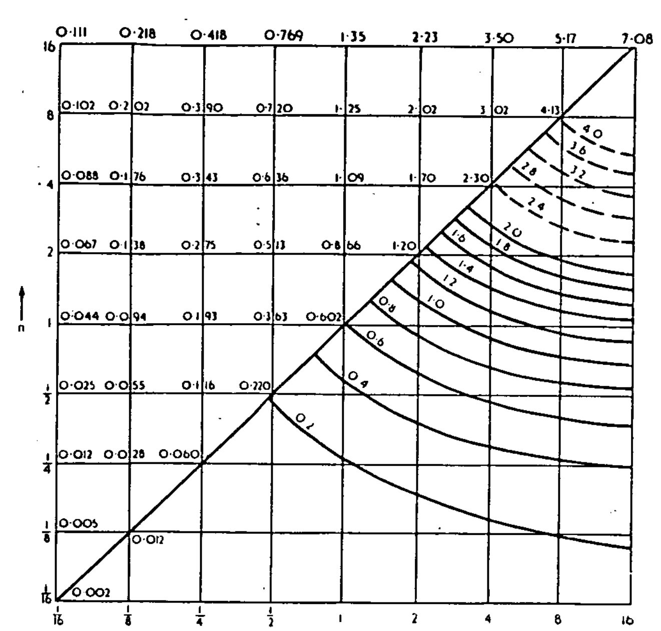


Fig. 1—Contours of $I_{min.}/\tau$

which is simply the boundary of stability as given by the Hurwitz criterion, and also on the line $\eta = 0$, when there is no integration in the loop. I must therefore have at least one minimum in the region enclosed between these curves, and this minimum must be a stationary value with respect to y and η, since I is well behaved. Consideration of the highest degree term in the equation $I =$ constant shows that I cannot have more than one minimum in this region, so a numerical search for a stationary value of I must lead to the required $I_{min.}$

A procedure for minimizing I with respect to y and η was programmed for an Elliott 402 computer (*see* Appendix I), which printed out values of $I_{min.}/\tau$ and the corresponding gain and integral action time, $\mu_{min.}$ and $(\tau I/\tau)_{min.}$ $[= (\mu/\eta)_{min.}]$. The results are shown in Figs. 1–3, with contours interpolated to show more clearly the form of the dependence on m and n. The charts are symmetrical about the diagonal, so that the contours are given only in the lower half, the computed ' spot heights ' being indicated in the upper half.

A simple modification of the program makes it possible to print values of I/τ as a function of y and η for fixed values of m and n, so as to investigate the behaviour of the system for controller settings other than the optimum. The results for $m = n = 1$ are shown in Fig. 4, which closely resembles a diagram given by Hazebroek and van der Waerden.[1] It is seen that the minimum is very flat, so that the control quality is not at all sensitive to the controller settings in the neighbourhood of the minimum.

AVERAGE OF THE SQUARE OF THE ERROR WITH A STATIONARY DISTURBANCE

While a step function disturbance provides quite a severe test of the dynamical behaviour of a system, it may bear little resemblance to the actual disturbances to which the system will be subjected. Knowledge of these disturbances must be based on observations of the past behaviour of the plant, or of similar plants, which allow statistical predictions to be made about the future behaviour of $d(t)$, rather than giving its complete functional form. In this situation the most appropriate representation of $d(t)$ is a member function of a stationary ergodic random process, whose statistical properties are consistent with those predictable from previous experience as described above.*

In the case of a stationary disturbance, the appropriate measure of control quality is the average of the square of the error rather than its integral, which clearly diverges. It is well known that the mean square value of a stationary time series is completely determined by its autocorrelation function, or alternatively by the power spectrum. $\overline{o^2}$ is, in fact, given by :

$$\overline{o^2} = \int_0^\infty O(f)\,df \quad \text{...(8)}$$

where $O(f)$ is the power spectrum of $o(t)$, and f represents frequency. Further, $O(f)$ is related to $D(f)$, the power spectrum of the disturbance, by :

$$O(f) = |Z(2\pi jf)|^2 D(f) \quad \text{...............................(9)}$$

Equations (8) and (9) determine $\overline{o^2}$ in terms of Z and D.

The class of disturbance spectra considered here is defined by :

$$D_0(\omega) = \frac{K}{1 + l^2\omega^2} \quad \text{.....................................(10)}$$

* In this section various definitions and well known results from the theory of stationary random processes are taken for granted. A more complete discussion and proofs can be found elsewhere.[2,3]

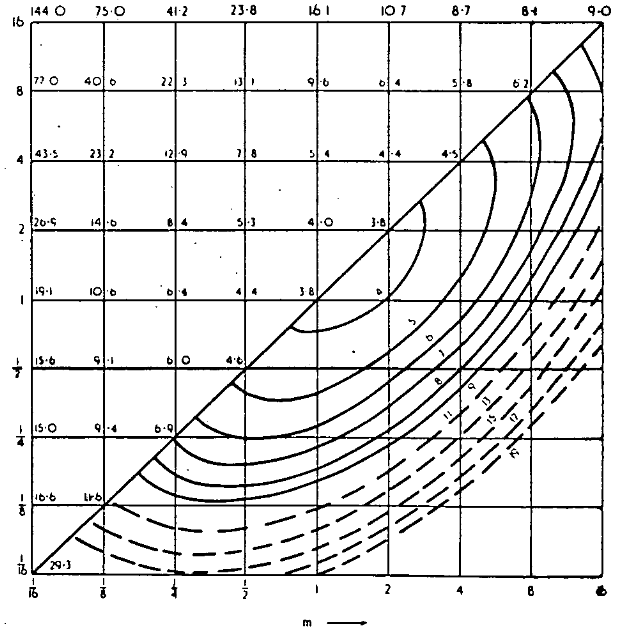


Fig. 2—Contours of $\mu_{min.}$

where $\omega = 2\pi f$. These are monotonic spectra with cut-off frequency $\omega = 1/l$, so by varying l it is possible to study the effect of this frequency on the controllability. $D_o(\omega)$ has a simple physical interpretation since it can be shown to be the power spectrum of a disturbance $d_o(t)$, which alternates between constant positive and negative values, the changeover points being placed at random on the time axis, with mean spacing l. The precise form of the distribution of the amplitudes of the segments does not matter, provided it has zero mean, and mean square given by :

$$\overline{a^2} = \frac{K}{4l}$$

Using the power spectrum of equation (10) is therefore equivalent to testing the system with a randomly spaced sequence of step disturbances rather than a single step ; the effects of successive steps will not be independent if their mean spacing is so close that the transient following one has not died out before the next arrives.

In the case of stationary disturbances, finite results can be obtained with a proportional-only controller ; in the interest of simplicity, therefore, this type is considered, though this leads to rather anomalous results for large values of l as shown below. Inserting $C(j\omega) = \mu$, together with equations (1) and (10) for Z and D, into equation (9) and combining this with equation (8) gives an expression for $\overline{o^2}$ which contains p as a factor, where $p = l/\tau$, and tends to zero as $p \to 0$. This simply means that the control quality becomes very good for small p, which is to be expected because of the smoothing of the disturbance by the time-constant τ. However, the numerical value of $\overline{o^2}$ is rather small for convenient use when $p \to 0$, so consider instead the ratio $\overline{o^2}/\overline{o_o^2}$, where $\overline{o_o^2}$ is the value taken by $\overline{o^2}$ when the feedback loop is disconnected. This ratio, which will be denoted by ψ, is a measure of the improvement in behaviour produced by connecting the feedback loop. It is easily shown that :

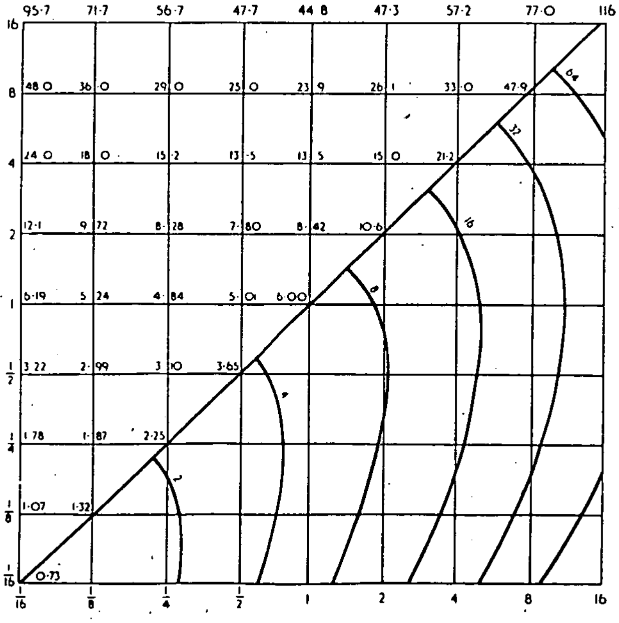


Fig. 3—Contours of $(\tau_I/\tau)_{min}$.

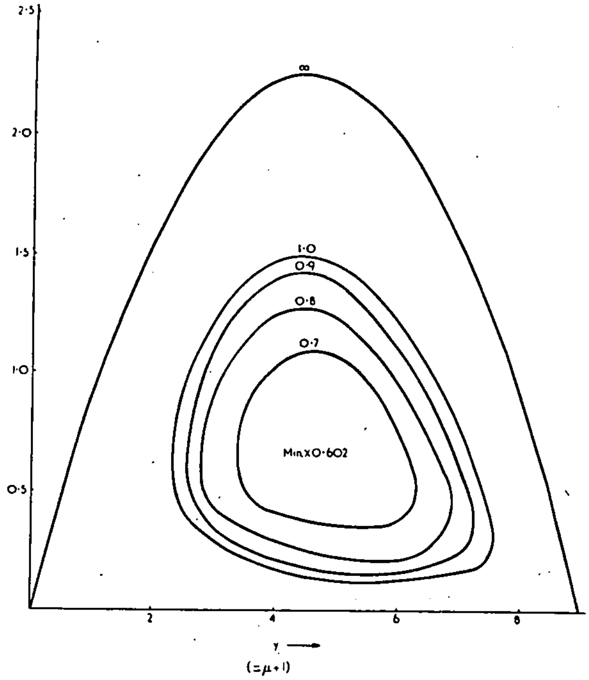


Fig. 4—l/τ as a function of y and n for $m = n = 1$

$$\overline{o_o^2} = \frac{p}{p+1} \qquad \text{...(11)}$$

so in ψ the factor p is replaced by $p + 1$, and ψ remains finite when $p \to 0$. ψ is the integral of a rational function of ω of the same type as that dealt with in the previous section, and can be evaluated explicitly in the same way, giving :

$$\psi = \frac{1+p}{y} \cdot \frac{A + By + Cy^2}{(D - Ey)(F + Gy)} \qquad \text{.....................(12)}$$

where $y = 1 + \mu$ and

$$\left.\begin{aligned} A &= (m + n + mn)\,[mn + p(m + n + mn) \\ &\qquad + p^2(1 + m + n)] \\ B &= mn\,[m^2 + n^2 + mn\,(1 + m + n)] + \\ &\qquad p\,(m^2 + n^2)\,(m + n + mn) - p^2mn \\ C &= pm^2n^2 \\ D &= (1 + m + n)\,(m + n + mn) \\ E &= mn \\ F &= mn + p\,(m + n + mn) + p^2(1 + m + n) \\ G &= p^3 \end{aligned}\right\} \text{......(13)}$$

(The use of A, B, C, D, and E here to represent different quantities from those in the previous section will not lead to any ambiguity.) It is seen that $\psi \to \infty$ through positive values as $y \to 0$ and as $y \to D/E$ from the interior of the interval $0 \to D/E$, so it passes through at least one minimum between these points, and since ψ = constant gives a cubic equation in y, it cannot pass through more than one. It can further be proved that the minimum value occurs for some $y > 1$, that is for some $\mu > 0$, as would be intuitively expected.

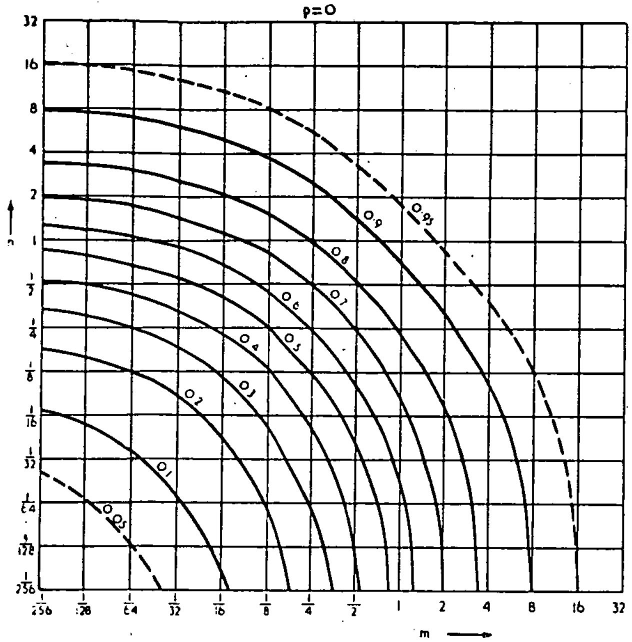


Fig. 5—$\psi_{min.}$ as a function of m and n for $p = 0$

The dependence of $\psi_{min.}$ on m, n, and p is best indicated by plotting contours of $\psi_{min.}$ in the (m, n)-plane for various values of p, but a considerable amount of labour would be involved in plotting these contours from a set of 'spot heights', as in the previous section, because of the number of diagrams required. It was therefore thought worthwhile to use an automatic contour-plotting program on the computer, making use of subsidiary routines to compute and minimize ψ. In this way contour diagrams were prepared for $p = 0$, $\frac{1}{8}$, $\frac{1}{2}$, 1, 8, 32, and 5000, the cases $p = 0$, 32, and 5000 being reproduced here as Figs. 5–7. For intermediate values of p, the charts undergo a continuous transition between the forms shown in Figs. 5 and 6.

To show the effect of the gain on ψ, the program was slightly modified to plot contours of ψ as a function of μ and m for fixed values of n and p. The resulting charts show a prominent 'valley' whose floor gives the relationship between μ and m, which minimizes ψ. The wall of the valley on the side corresponding to large μ rises steeply, so that the height becomes infinite on the stability boundary.

DISCUSSION OF THE RESULTS

Step Disturbance

It is seen from Fig. 1 that $I_{min.}/\tau$ is a monotonic increasing function of both m and n in the region investigated, and the slope along the diagonal $m = n$ increases monotonically with m and n (on the logarithmic scales used). This increase continues up to $m = n = 2^{14}$, to which the computations were extended, so it is certainly safe to say that $I_{min.}/\tau$ shows no sign of flattening out in any region of practical interest. Another interesting feature of Fig. 1 is that, for $m > n$, the line of steepest descent across the contours makes a larger angle with the m-axis than with the n-axis; in other words a greater improvement in controllability can be obtained by reducing the smaller time-constant in a given ratio than by reducing the larger in the same ratio.

$I_{min.}$ is, of course, a function of the three time-constants τ_1, τ_2, and τ_3, which has been obtained in the form:

$$I_{min.} = \tau_1 f(m,n) = \tau_1 f(\tau_2/\tau_1, \tau_3/\tau_1) \quad \text{.................(14)}$$

Fig. 1 is a section on the plane $\tau_1 = 1$ through the three-dimensional space (τ_1, τ_2, τ_3), and other sections parallel to this can be obtained merely by scaling the variables. Using equation (14) with Fig. 1, it is also possible to plot two-dimentional contour diagrams for sections perpendicular to each of the other two axes, though in fact, because of the symmetry in m and n, it is only necessary to plot one such set. In this way the function $I_{min.}$ (τ_1, τ_2, τ_3) has been mapped through the interior of the cube:

$$1/4 \leq \tau_1, \tau_2, \tau_3 \leq 4 \quad \text{..(15)}$$

by drawing two-dimensional contour diagrams on sections spaced at equal intervals perpendicular to the τ_1 and τ_3 axes. Although the sections perpendicular to the τ_1 axis are similar in appearance to Fig. 1, the orthogonal set reveals the interesting fact that $I_{min.}$ passes through a maximum value as τ_1 is varied, τ_2 and τ_3 being held constant, so that, for each τ_2, τ_3, there is a value of τ_1 which gives poorest controllability. It is easy to see qualitatively why this should be so. When τ_1 is very small, the system approximates to a loop with two time-constants in which the disturbance affects the measured value directly, giving rise to a transient with quite large initial deviation but heavy damping and fairly high frequency. On the other hand, when τ_1 is very large, the system approximates to a loop with one time-constant, through which both disturbance and correction are applied,

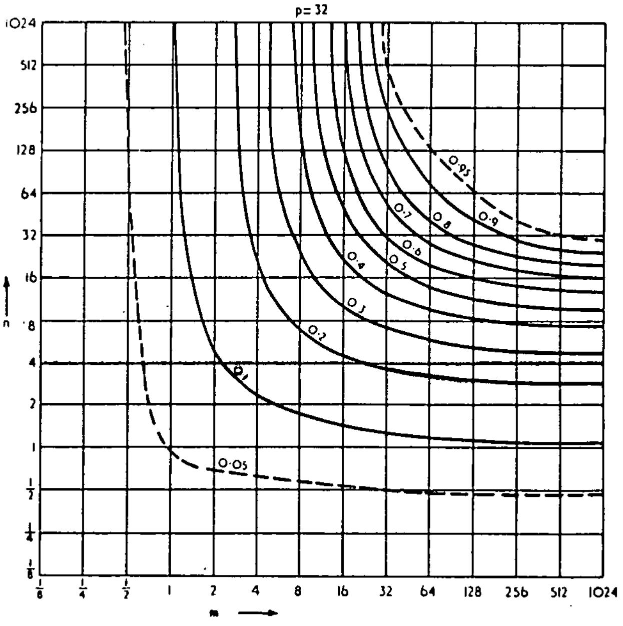


Fig. 6—$\psi_{min.}$ as a function of m and n for $p = 32$

giving a transient of low frequency with very small initial deviation and heavy damping. Both these cases would be expected to be better, judged by the present criterion, than a transient with fairly large initial deviation and rather poor damping such as would be obtained with an intermediate value of τ_1.

The limiting gain for stability with proportional feedback μ_l has often been used as a criterion of controllability,[4] and it is interesting to compare it with the present one. μ_l depends only on the time-constant ratios m and n and can be plotted as a contour diagram in the m,n-plane (Fig. 8), from which it is seen that μ_l is a minimum when all three time-constants are equal. Considered as a function of any one of τ_1, τ_2, τ_3 for fixed values of the remaining two, it passes through a minimum (corresponding to poorest controllability) when the time-constant in question is the geometric mean of the other two. Using $I_{min.}$ as a criterion, however, the controllability varies montonically with τ_2 and τ_3, and it is only when considered as a function of τ_1 that it exhibits a poorest value for some finite τ_1, as discussed above. It is curious that, over the limited range investigated in equation (15), the value of τ_1 which gives poorest controllability is approximately proportional to the geometric mean of τ_2 and τ_3.

One well-known method for determining the settings of a proportional-plus-integral controller[5] is to adjust the gain and integralaction time to give a transient with $e : 1$ subsidence ratio and integral action time equal to the period. These settings, denoted by μ_e and τ_e can be calculated approximately without great difficulty for a plant with three exponential transfer stages, and plotted as contour diagrams in the m,n-plane. The μ_e diagram obtained in this way is similar in form to Fig. 2, which gives $\mu_{min.}$, but the τ_e/τ diagram bears very little resemblance to Fig. 3, giving $(\tau_I/\tau)_{min.}$ The discrepancy is greatest in the neighbourhood of $m = 16$, $n = 1/16$, where $(\tau_I/\tau)_{min.} \doteqdot 96$ and $\tau_e/\tau \doteqdot 2{\cdot}8$! This is perhaps hardly surprising as the choice of τ_e equal to the period of the transient is more or less arbitrary. Further, the value of I/τ corresponding to the settings μ_e, τ_e is 0·27, compared with an optimum value 0·111 for $I_{min.}/\tau$. The relatively small difference between these figures for such widely differing controller settings is comforting confirmation of the flatness of the minimum corresponding to optimum control quality. An investigation of the actual form of the step responses shows that the 'optimum' settings, determined by the present method, give a transient with smaller overshoot, slightly higher frequency, and rather poorer damping than that given by the settings μ_e, τ_e.[6,7]

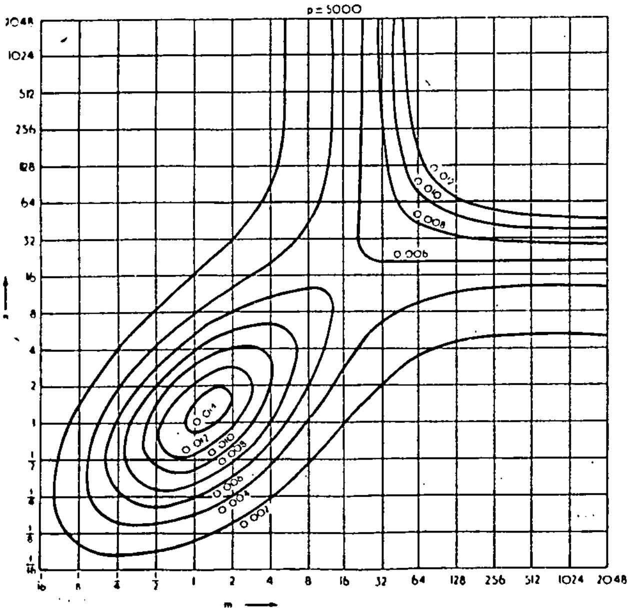


Fig. 7—$\psi_{min.}$ as a function of m and n for $p = 5000$

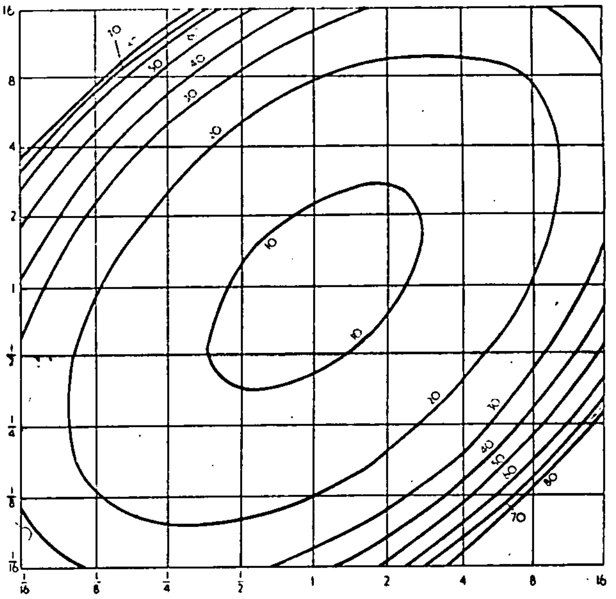


Fig. 8—Limiting gain, μ_l, as a function of m and n

The flatness of the minimum is shown clearly by Fig. 4, from which it is seen that μ and η can each be changed by about 50% from their optimum values without increasing I/τ beyond 0·8, compared with its minimum value of 0·602. This may be compared with a similar result quoted by Hazebroek and van der Waerden.[1] Figure 4 is very similar to these authors' Fig. 3.

Stationary Time Series Disturbances

The most obvious difference between Fig. 1 and the controllability charts for stationary disturbances given in Figs. 5 and 6 (excluding for the moment the case $p = 5000$ illustrated in Fig. 7) is that the value of ψ does not continue to increase as m and n are increased. Instead there is an escarpment separating two plateaux, one in the region of small m and n, corresponding to good controllability, and one in the region of large m and n, corresponding to poor controllability. For the larger p the contours are similar in shape to those of Fig. 1, which is to be expected, since the separate step changes of the stationary disturbance are widely spaced, and the response after each step corresponds more closely to that following a single step disturbance, provided the time-constants are not too long. For small values of p, however, corresponding to disturbances with high cut-off frequencies, the curvature of the contours reverses in parts of the chart ; in particular, for the limiting case $p = 0$ (Fig. 5), it is easy to show that the contours are throughout concave towards the origin and are asymptotic to lines parallel to the axes. The contour $\psi = 0{\cdot}5$, for instance, is asymptotic to $m = 1$ and to $n = 1$. In regions where the curvature is reversed in this way, it clearly pays to reduce the *larger* of m and n in a given ratio rather than the smaller, contrary to the conclusion reached above when discussing Fig. 1.

The dependence of ψ on the three time-constants τ_1, τ_2, and τ_3 can be obtained by a procedure similar to that used in the case of $I_{min.}$ above, and sections through a three-dimensional region can be constructed in the same way. The quantity $\overline{{}_o^{2}}_{min.}$, measuring the actual optimum control quality rather than the improvement produced by the feedback loop, can be obtained simply by multiplying $\psi_{min.}$ by $\dfrac{p}{p+1}$.

It remains to discuss the chart for $p = 5000$ (Fig. 7) which is rather different in appearance from Figs. 5 and 6. A closer inspection, however, reveals that it has essentially the same form, the edge of the lower plateau lying in the upper right-hand corner. The vertical scale covered is much smaller than in the case of the other charts, so the prominent hill near the point (1, 1) is, in fact, merely a small bump which projects from the escarpment into the lower plateau. On this chart

$$p \gg 1, m, n$$

so that it is possible to write $p = 1/q$ in equations (12) and (13) and neglect terms of $O(q^2)$. This gives :

$$\psi \simeq \frac{1}{y\,[y+q\,(1+m+n)\,]}\left\{\,1+q+\frac{q\,(m+n+mn+m^2y)(m+n+mn+n^2y)}{D-Ey}\right\} \quad \ldots\ldots\ldots(16)$$

For small q this has a minimum value for y very nearly equal to D/E and it is a good approximation to take

$$\psi_{min.} = \psi'_{min.} = \left[\frac{1}{y\left\{y+q\,(1+m+n)\right\}}\right]_{\substack{y\,=\,D/E\\ q\,=\,O}}$$

But $y = 1 + \mu$, so $D/E = \mu_l + 1$ and

$$\psi_{min.} = \frac{1}{(\mu_l + 1)^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(17)$$

where μ_l is the limiting gain for stability, as above. Thus the contours of $\psi_{min.}$ approach the limiting gain contours (Fig. 8) when p is sufficiently large compared with m and n. The hill near (1, 1) in Fig. 7 therefore corresponds to the depression in Fig. 8 ; it is seen that their heights are comparable, as expected.

The physical interpretation of this behaviour of the contours is extremely simple. For large p, the system is effectively being tested with very widely spaced step disturbances, and so it spends the majority of the time very near to a steady state with deviation $1/(1+\mu)$ from the desired value, while the transient oscillations following each step occupy only a very small fraction of the total time. In averaging the square of the deviation with respect to time, the steady state is therefore weighted very heavily compared with the transients, and the result is very nearly $1/(1+\mu)^2$ which is minimized by making μ almost equal to μ_l. However, if m and n are increased with a fixed value of p, the duration of the transient after each step increases until it is given weight comparable with the final deviation. The chart then begins to resemble those drawn for smaller values of p, as can be seen in the top right-hand corner of Fig. 7. When p is as large as 5000, for practical values of m and n it is clearly more realistic to regard the disturbance as a sequence of independent steps rather than a stationary time series, and to discuss the controllability in terms of $I_{min.}/\tau$ as given in Fig. 1.

CONCLUSIONS

In cases where the integral or average of the square of the error is an appropriate measure of control quality, the method of estimating controllability described here provides a useful means of assessing the limitations imposed by the structure of the plant on the control quality attainable with a given class of controllers. In particular, it provides a method of comparing the values of alternative designs aimed at improving the controllability.

When the dynamical behaviour of the plant can be specified in terms of two parameters, as in the simple case treated here, a geometrical representation in the form of contour charts is very useful in suggesting those modifications to the parameters which would be of greatest value ; if more than two parameters are involved, a complete geometrical representation is no longer possible but the method can still be used for comparing a number of alternative designs.

As was emphasized at the beginning of the paper, there is no fundamental limitation to cases in which the integrals involved can be evaluated explicitly. The time taken for a calculation compares very favourably with the time for graphical methods[5] ; in the present case, with a proportional-plus-integral controller, $I_{min.}/\tau$ was calculated in about 1½ min from given values of m and n. The programs were written in a slow interpretive code, and this time could be reduced considerably by writing in machine code for one of the faster machines.

An important feature of the method is that stationary time series disturbances can be handled as easily as the usual test disturbances, which is likely to prove valuable as more information about the nature of disturbances affecting process plants becomes available.

Acknowledgments

The author wishes to thank Mr. E. N. Martin for many helpful discussions which contributed materially to the final form of this work, Mr. C. M. Reeves whose contour-plotting routine was used in constructing the ψ-charts and who devised the auxiliary minimization program for this case, and Mr. R. Johnson who undertook many of the auxiliary computations which were not programmed.

References

1. P. Hazebroek and B. L. van der Waerden : *Trans. A.S.M.E.*, 1950, vol. 72, p. 309.
2. H. M. James, N. B. Nichols, and R. S. Phillips : 'Theory of Servomechanisms', Ch. 7 : 1947, McGraw-Hill.
3. J. H. Laning and R. H. Battin : 'Random Processes in Automatic Control', 1956, McGraw-Hill.
4. D. G. Prinz : *J. Sci. Instr.*, 1944, vol. 21, No. 4, p. 53.
5. A. J. Young : 'An Introduction to Process Control System 'Design', 1955, Longmans.
6. H. H. Rosenbrock : *Proc. I.E.E.*, 1955, vol. 102, p. 602.
7. J. H. Westcott : *Proc. I.E.E.*, 1954, vol. 101, p. 471.
8. J. M. L. Janssen and R. P. Offereins : *Trans. Soc. Instr. Techn.*, 1955, vol. 7, p. 111.

APPENDIX I

Computer Programs

Considering first the case of step disturbances already dealt with, the problem is to miminize the expression given in equation (6), in the region of the (y, η)-plane enclosed between the axis $\eta = O$ and the parabola $\eta = y\frac{(C-Dy)}{E}$.

The program was written in two parts, a main routine to perform the minimization and a subsidiary routine to compute the expression on the right-hand side of equation (6). The subsidiary routine is straightforward and of little interest and it will suffice to describe the minimization procedure used.

With an expression for I as simple as equation (6), the stationary point with respect to η for a fixed value of y can be found explicitly. Equating $\frac{d}{d\eta}\left(\frac{2I}{\tau}\right)$ to zero, and taking the root of the resulting quadratic in η lying in $\eta > 0$, gives :

$$\eta = \left(\frac{C - Dy}{A + By}\right)\left[\left\{1 + y\frac{(A + By)}{E}\right\}^{\frac{1}{2}} - 1\right] \quad \text{......(18)}$$

Substituting for n from equation (18) into (6) gives a function $f(y)$ of a single variable y, and it is known that the minimum lies in

$$0 < y < C/D$$

with the function unbounded at the end points of this interval.

The following numerical procedure was used to locate the minimum :

(i) The interval $0 < y < C/D$ is divided into p equal parts by points :

$$y_1 = \frac{1}{p}.\frac{C}{D}, y_2 = \frac{2}{p}.\frac{C}{D}, \ldots y_{p-1} = \frac{p-1}{p}.\frac{C}{D}$$

Then it is clear that, if $f(y_n) \geq f(y_{n+1})$, the minimum cannot lie in $y \leq y_n$, while, if $f(y_n) \leq f(y_{n+1})$, the minimum cannot lie in $y \geq y_{n+1}$. If values of the function are calculated at each y_i successively and $f(y_i) - f(y_{i-1})$ is found at each stage and tested for sign, then either there exists some y_q such that $f(y_q) \leq f(y_{q-1})$ and $f(y_q) \leq f(y_{q+1})$, in which case :

$$y_{q-1} < y_{min.} < y_{q+1}$$

or $f(y_{p-1}) \leq f(y_{p-2})$ in which case :

$$y_{p-2} < y_{min.} < C/D$$

since f is unbounded when $y \rightarrow C/D$.

At each stage, therefore, it is necessary to test $f(y_i)$ against $f(y_{i-1})$ and to test y_i against y_{p-1}, and by this process an interval of length $\frac{2}{p}.\frac{C}{D}$ is found which contains y_{min}.

(ii) This interval is in turn divided into p parts and the process is repeated, finding an interval of length $\left(\frac{2}{p}\right)^2 . \frac{C}{D}$ which contains y_{min}.

(iii) The intervals containing $y_{min.}$ are repeatedly divided into p equal parts until, after n repetitions, the inequality $\left(\frac{2}{p}\right)^n < \epsilon$ is satisfied, where ϵ is a specified small number. The position of $y_{min.}$ is then determined within an interval of length $\epsilon . \frac{C}{D}$ The mid-point of this interval is then taken as $y_{min.}$, and the corresponding value of I/τ as $I_{min.}/\tau$.

The choice of p which locates the minimum to a given accuracy in the smallest number of calculations depends, of course, on the position of the minimum. If it is assumed that the maximum number of calculations is needed to locate $y_{min.}$ at each reduction of the interval length, it is easy to show that the best value is $p = 4$. If the situation is less unfavourable than this, the best p is larger, and it was thought reasonable to choose $p = 5$. Values of p and ϵ can be set by the input tape, and with $p = 5$ and $\epsilon = 1/200$ the time required for minimization varied between 1 min 30 s and 1 min 45 s. On completing the minimization for a given pair of values of m and n, the values of $I_{min.}/\tau$, $\mu_{min.}$ and $(\tau_I/\tau)_{min.}$ are printed on a new line, and the machine is directed to read the next pair of values of m and n (or a halt order) from the input tape.

With a more complicated transfer function it would probably be best not to attempt to derive explicitly the value of η which minimizes I, for each y, but instead to carry out the type of search process just described in two dimensions. Assuming $p = 5$ and $\epsilon = 1/200$ to be used for both dimensions, this would probably increase the computation time by a factor of about 30 if the time for computation of I with given η and y remained the same. With a more complicated I it would be even longer, but it must be remembered that the present programs were written in a very slow interpretive code for convenience in programming. By using the machine code a very substantial reduction in the times just quoted could be obtained. In carrying out a two-dimensional search, it would also probably be more efficient to use a 'steepest descent' type of process rather than the 'rectangular grid' procedure just described.

The programs for the case of the stationary disturbance treated earlier in the paper are straightforward. The stationary value with respect to y was found by equating the explicit expression for $d\psi/dy$ to zero and solving the resulting equation numerically. The contour-plotting routine was a recently developed library subroutine for the computer used. It printed co-ordinates of points spaced along a contour ψ = constant at equal chord separations. The time taken to produce a complete chart, of the type shown in Figs. 5–7, varied between ½ h and 1 h.

A2

Reprinted from
Transactions of The Society of Instrument Technology
Vol. 10 No. 4 December, 1958

A NON-LINEAR THEORY OF THE DYNAMICAL BEHAVIOUR OF PNEUMATIC DEVICES

By R. JACKSON, B.A.

SYNOPSIS

Certain features of the response of pneumatic devices to large and rapidly varying inputs are of a type that cannot be accounted for by any linear theory. A non-linear theory which takes account of saturation effects in pneumatic amplifiers has therefore been developed and applied to some fairly simple systems, and the results have been found to agree well with experiment. With one extreme type of approximation the theory degenerates into the usual linear theory, but the other extreme approximation, which has been called the switching approximation, proves to be more suitable in some cases. The phase-plane representation, well known in non-linear mechanics, provides a very useful method of analysing and exhibiting the behaviour of the systems investigated. 106

INTRODUCTION

DYNAMICAL effects in pneumatic systems fall naturally into two classes: (*a*) *inherent* dynamics of the system, due to the modes of operation of flapper-nozzles and relays and to the capacities of chambers, connections, and bellows; and (*b*) *imposed* dynamics, due to the introduction of adjustable restrictors (derivative and integral valves) for the purpose of modifying the dynamical behaviour to give desired control actions.

Imposed dynamical effects arise only in controllers, of course; the dynamics of devices such as differential-pressure transmitters are entirely inherent.

Theoretical treatments of the dynamics of pneumatic controllers have been given by a number of authors[1,2] who were principally concerned with the effect of the derivative and integral action times on the response, and hence with the imposed dynamics in the sense defined above. Linear theory and the frequency-response approach were used throughout. The frequency-response method was extended by Gould and Smith,[8] and by Westcott,[3] to take account of inherent dynamical effects arising from lags in the forward loop of the controller. The flapper-nozzle and relay were treated as linear amplifiers with associated time lags, and on this basis a method of estimating the gains and the time constants in the forward loop was devised.

The effect of non-linearity in the flapper-nozzle characteristic on the static behaviour of the system has been discussed by Kirk[4], who suggests design modifications to improve the linearity and to make it behave more closely as a perfect null-balance detector, while a very thorough treatment of the effect of certain non-linearities on the dynamical behaviour has been given by Webb[5]. Ream, Tizard, and Townend[9] have noted a practical case where saturation effects in the relay of a controller had a significant effect.

The present work arose from an investigation of the speed of operation of diaphragm valve motors in the course of which it became clear that many effects observed when the controller input changed rapidly could not be explained by any linear theory. The well-known and often troublesome sustained oscillation known as ‘pumping’, which occurs in certain pneumatic systems, was also found to exhibit some of the characteristic features of non-linear oscillations. To account for these observations and, at the same time, for the success of linear theory when applied to small or slow input variations, a piecewise-linear theory has been developed which takes account of the narrowness of the effectively linear region of operation of pneumatic amplifiers. In the present paper this is applied to the simple case of a valve motor driven directly by a controller, with a negligible length of intervening pneumatic line. The predictions of the theory are shown to be in good agreement with experimental step responses, and to lead to a method of estimating the time constants in the forward loop of the controller which may be regarded as an alternative to that of Westcott. Under certain conditions it is shown that a ‘switching approximation’, in which the amplifiers are regarded

In addition to the references cited, the following may be found a useful introduction to phase-plane analysis: G. D. S. MACLELLAN, ‘Phase-plane methods in process control system design,’ *Trans. Soc. Instr. Tech.*, 1957, vol. 9, pp. 62–71.

Mr Jackson is with the Engineering Research Department at the Billingham Division of I.C.I.

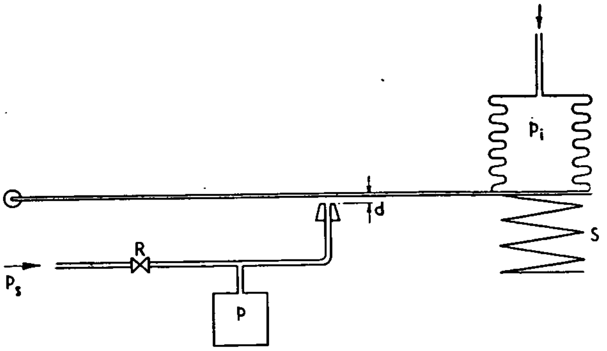


Fig. 1—Flapper-nozzle system

as pure switches with no zone of linear operation, gives a good approximation to the behaviour. This approximation and the usual linear theory may be regarded as the two extreme cases of the complete piecewise-linear theory.

The response of the system to sinusoidal inputs of large amplitude and high frequency can be treated by the switching approximation, which gives an almost triangular output waveform in agreement with observations of the author and others.[6] Frequency response curves can be predicted, but both phase lag and attenuation depend on the amplitude of the input as well as its frequency. The switching approximation is also used to predict the amplitude and frequency of sustained oscillations in a more complicated system, and is found to give excellent agreement with the observed results.

PNEUMATIC AMPLIFIERS

In analysing the dynamical behaviour of pneumatic systems, wide use has been made of the analogy between pneumatic devices, such as flapper-nozzles, and electronic amplifiers. In the simplest pneumatic units an amplifying device feeds a volume load, and it has been usual to represent this by a linear amplifier feeding a single time-constant, giving a transfer function of the form:

$$\frac{K}{1 + \tau s}$$

where K is the gain of the amplifier and τ the time constant associated with the volume load.[3]

That the actual situation is more complicated than this may be seen by considering in greater detail the operation of a flapper-nozzle system feeding a volume load. Figure 1 shows the essential features of such a system. A pneumatic input p_i is applied by means of a bellows and opposed by a spring S, which represents the collective stiffness of all bellows and springs attached to the flapper in an actual controller. It is convenient to measure pressures from the centre of the working range as origin, so with standard devices working in the range 3 to 15 lb/in^2 gauge the origin will be chosen at 9 lb/in^2 gauge.* With this convention it is convenient to choose the origin in measuring the flapper displacement, d, at the position where both input pressure and equilibrium output pressure are 9 lb/in^2 gauge. Then assuming that the displacement of the input bellows is proportional to p_i, we have

$$d = gp_i \qquad \text{(1)}$$

where g is a constant depending on the stiffness of S, and the effective area on which p_i acts. If p_s is the supply pressure and p_o the pressure on the downstream side of the nozzle (often atmospheric pressure), the mass flows through the fixed restrictor R and the nozzle are given respectively by:

$$\mathrm{f}_R = R\,\mathrm{f}(p_s, p), \quad \mathrm{f}_N = N\,\mathrm{h}(p, p_o) \qquad \text{(2)}$$

where the precise forms of the functions f and h depend on the flow regimes, but in all cases they are monotone increasing in the first variable and monotone decreasing in the second. The constants R and N depend on the restrictions to flow provided by the fixed restrictor and the flapper-nozzle respectively. N will clearly be a monotone increasing function of d, and hence of p_i:

$$N = \mathrm{N}(d) = \mathrm{N}(gp_i) \qquad \text{(3)}$$

Then if V is the load volume and $\boldsymbol{R}$ and T are the gas constant per gram and the absolute temperature respectively, the rate of change of load pressure is given by:

$$\frac{V}{\boldsymbol{R}T}\,\frac{\mathrm{d}p}{\mathrm{d}t} = R\,\mathrm{f}(p_s, p) - \mathrm{N}(gp_i)\,\mathrm{h}\,(p, p_o) \qquad \text{(4)}$$

Linearization of this equation in the usual way for small variations of p_i and p in the neighbourhood of values P_i and P, for which the system is in equilibrium, leads to the usual representation as a linear amplifier feeding a single time-constant.

On examining (4) it is seen that there are three sources of non-linearity:

(i) The non-linear dependence of $\mathrm{N}(gp_i)$ on p_i

(ii) The non-linear form of the functions $\mathrm{f}(p, p')$ and $\mathrm{h}(p, p')$ relating the flow of gas through the restrictions to the upstream and downstream pressures

(iii) The appearance of a product of $\mathrm{N}(gp_i)$ and $\mathrm{h}(p, p_o)$ in the second term on the right-hand side.

The agreement with experiment obtained with the theory to be developed shows that only (i) is sufficiently important to affect the qualitative nature of the transients, while (ii) and (iii) may be regarded as small corrections to their precise shape.

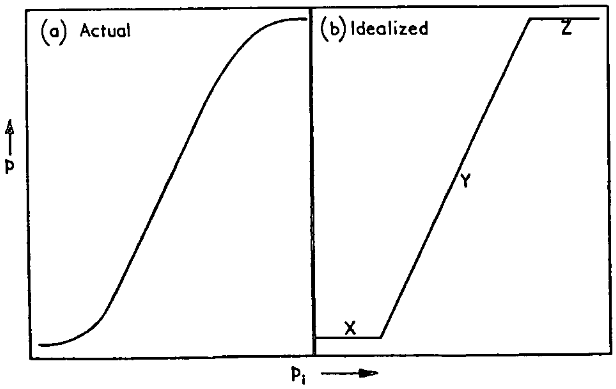


Fig. 2—Gain characteristics of pneumatic amplifiers

* To avoid confusion in the following, a pressure P measured with respect to 9 lb/in^2 gauge as origin will be written simply as 'P lb/in^2', while the same pressure measured with respect to atmospheric pressure as origin will be written 'P lb/in^2 gauge', if it is desired to indicate the origin explicitly.

Dealing then with (i), it is known that with a flapper-nozzle system of the usual design, quite a small change in d, of the order of one or two thousandths of an inch, is sufficient to cause the equilibrium value of p to traverse the whole working range. Consequently, unless the spring stiffness of the flapper assembly is very large, the output pressure traverses its complete range for quite a small change in input pressure. For one widely used ' floating disc ' type of controller, for instance, the change in input pressure required to change the nozzle pressure from 3 to 15 lb/in^2 gauge varies between $\frac{1}{4}$ and $1\frac{1}{4}$ lb/in^2 depending on the proportional-band setting. Outside this range the system rapidly saturates, giving a relationship between p and p_i of the form shown in Fig. 2 *a*. For the purpose of analysis, it is proposed to represent this by a piecewise-linear characteristic as shown in Fig. 2 *b*, which reproduces the saturating features of the true characteristic, and gives a zone of linear operation of the correct width. Two extreme types of approximation to a system of this type may be recognized, first a linear approximation, in which the saturation effects are neglected and the width of the linear band is considered to be effectively infinite, and second a switching approximation, in which the linear band is neglected and the system is regarded as a pure switch. The p–p_i characteristic would then be of the type shown in Fig. 2 *b* with a vertical central segment. Although it is not always easy to form a prior judgment as to which of these extremes would best represent the behaviour in a given case, it is fortunately possible to treat the intermediate case (Fig. 2 *b*) for some simple systems, and so to investigate the transition between the two extremes.

When the flapper is in the saturation region Z (Fig. 2 *b*), very close to the nozzle, p aims at the supply pressure p_s in a manner determined by the form of the functions f and h in equation (4). Whatever the precise form of these functions, however, N can be assumed to be almost zero since the nozzle is sealed off by the flapper, and p will rise monotonically towards p_s at a speed determined by R, the fixed restriction. If f (p_s, p) were simply p_s–p, the rise would be exponential and R would determine the time constant. It is proposed here to assume that this is the case and to choose a time constant which gives a reasonable fit to the true rising-pressure curve. Thus the neglect of the detailed form of f (p_s, p) will affect only the precise shape of the response of p. In a similar way, when the flapper is in the saturation region X, p falls to some equilibrium pressure near to p_0. In this case, however, the effective time constant is determined by R and the open nozzle in parallel, so it must be shorter than the time constant for rising pressure. It follows, therefore, from the nature of the flapper-nozzle arrangement, that *two* time constants are needed, strictly speaking, to describe its saturated behaviour.

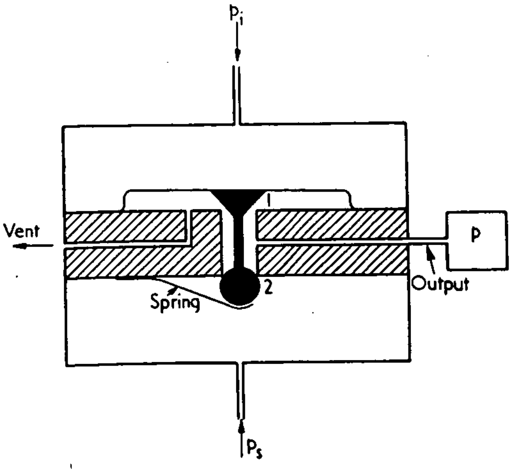
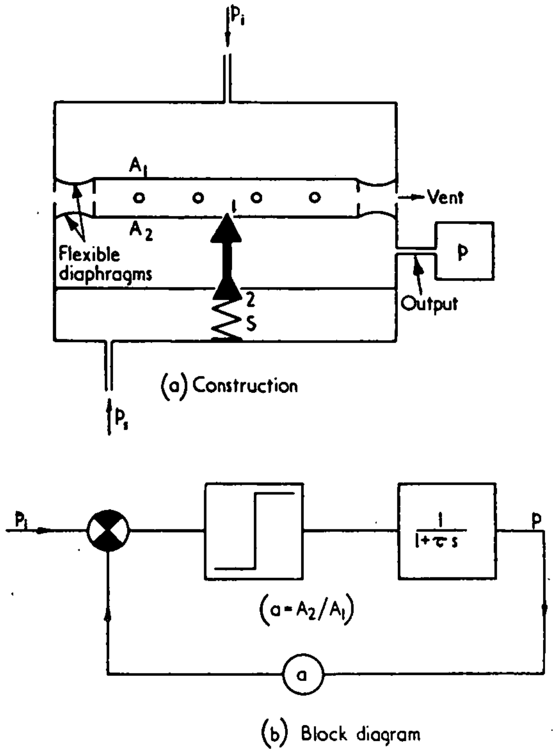


Fig. 3—Continuous-bleed, high-gain relay



(a) Construction

(b) Block diagram

Fig. 4—Non-bleed, low-gain relay

Figure 3 shows an arrangement which is typical of a second class of pneumatic amplifiers, commonly used as power relays. This is essentially the same as the flapper-nozzle system, except that both restrictions vary as p_i varies. The restrictions in question are the valves 1 and 2, which work in opposition. In the saturation region with rising p, 1 is closed and 2 is open, whereas with falling pressure 2 is closed and 1 is open. With suitably-matched valves there is no reason, in this case, why the effective time constants for rising and falling pressure should not be equal, though in practice they seldom are.

There is a continuous bleed of air from the supply in the balance condition, since the system balances with both valves 1 and 2 partially open. Accordingly it will be referred to as a continuous-bleed, high-gain relay.

In the following sections both this type of amplifier and the flapper-nozzle will be treated in the same way, as an amplifier with gain-characteristic of the type shown in Fig. 2 *b* feeding a single time-constant,

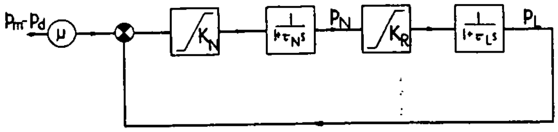


Fig. 5—Block diagram for controller driving simple load

the load. The difference between the time constants for rising and falling pressures in the case of the flapper-nozzle will often be neglected in the interest of simplicity. It will be clear from the treatment that there is no difficulty in taking account of this if necessary.

A third common type of amplifier is the low-gain, non-bleed relay. One common arrangement of this type is illustrated in Fig. 4 *a*. The output pressure p acts on the diaphragm in opposition to p_i, so if A_1 and A_2 are the effective areas of the upper and lower surfaces of the assembly, the balance condition is:

$$A_1 p_i = A_2 p$$

and the static gain is A_1/A_2. In several cases $A_1 = A_2$, giving unit gain, and the ratio seldom exceeds three or four. It is important to notice that valve 2 can open only after 1 has closed, for it is the pressure of the diaphragm on 1 which causes 2 to open. Conversely, 1 can open only after 2 has closed, for it is the seating of 2 that lifts 1 away from the diaphragm assembly. Neglecting the off-balance forces due to the pressure differences at the plugs and the spring S, the system would therefore be expected to act as a perfect switch, connecting the load either to the supply pressure through 2 or to atmosphere through 1. In practice, the off-balance forces referred to above give rise to a finite dead band in the operation of the switch, while imperfect seating of valves 1 and 2 gives rise to a small range of 'continuous bleed' operation. Although these features have been found to produce quite marked dynamical effects in certain cases,* it is possible to treat many phenomena using the simple representation as a perfect switch, so the representation shown in Fig. 4 *b* will be used in the present paper.

APPLICATION TO CONTROLLERS DRIVING SIMPLE LOADS

The simplest pneumatic systems of practical interest are those in which a volume load is driven directly by a proportional controller or transmitter. The load volume may be fixed or, more commonly, variable as in the case of a receiving bellows or diaphragm motor. The variation of the load volume with pressure will be a further cause of deviation in detailed shape from simple exponential curves, but in most cases where only the qualitative form of the response is to be discussed it will be neglected. It is, however, assumed that none of the connecting lines is sufficiently long to have an appreciable effect on the dynamical behaviour of the system.

Consider first a proportional-only controller of the floating-disc type, with a continuous-bleed power relay as shown in Fig. 3. The lines connecting the relay output, the load, and the feedback bellows are short and present no significant restriction to air flow compared with the resistance of the relay ports, so the load, feedback bellows, and connecting lines may be regarded as a single volume, throughout which the pressure is equal to the load pressure p_L. There is, of course, no sense in which a signal is 'sent out' by the controller and 'received' by the valve, and it is not permissible to divide the system between these two and consider them as separate units, as has been emphasized by Buckley.[6] Using the approximations discussed in the previous section, the block diagram of the system is as shown in Fig. 5, where μ is the proportional gain (i.e. 100/proportional bandwidth), K_N and K_R the slopes of the linear segments of the flapper-nozzle and relay gain characteristics respectively, τ_N and τ_L the time constants associated with the flapper-nozzle and with the relay and load respectively, and p_m and p_d the pressures representing the measured variable and the desired value. The same time-constant (τ_N or τ_L) is used for both rising and falling pressures in the interests of simplicity. Because of the geometry of the floating-disc arrangement, K_N is a function of μ rather than a constant, and the expression

$$K_N(\mu) = K_N(1).\frac{\mu + 1}{\mu^2 + 1} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

is a good approximation for proportional bands wider than about 2%. It is convenient to choose a supply pressure of 18 lb/in² gauge, so that the centre of the working range lies midway between atmospheric pressure and the supply pressure.

The behaviour of the system can then be represented by a modification of the usual phase-plane method of non-linear mechanics[7]. p_N and p_L, the two variables determining the states of the non-linear elements, are plotted against each other on rectangular axes, with time as a parameter. With the above value of the supply pressure, and origins for the pressures chosen at 9 lb/in² gauge, the behaviour of the system can then be represented within the square:

$$-9\,p.s.i. \leqslant p_N \leqslant +9\,p.s.i., -9\,p.s.i. \leqslant p_L \leqslant +9\,p.s.i. \ldots(6)$$

as shown in Fig. 6. There is a zone:

$$-\Delta_R < p_N < \Delta_R \text{ where } \Delta_R = 9/K_R \quad \ldots\ldots\ldots\ldots\ldots\ldots(7)$$

in which the relay operates in its linear band, and a zone:

$$-\Delta_N < p_L - \mu\ (p_m - p_d) < \Delta_N \text{ where } \Delta_N = 9/K_N \quad \ldots\ldots(8)$$

in which the flapper-nozzle operates in its linear band. The intersection of these gives a rectangle about the equilibrium point, P, in which the behaviour is completely linear. In each of the regions into which the (p_L, p_N)-plane is divided by these zone boundaries, the system is described by a set of linear differential equations, as discussed below, and the solutions can be represented as trajectories in the (p_L, p_N)-plane with time as a parameter. The differential equations will, of course, be different in different regions, but it is easy to show that the separate segments must join up continuously, with continuous gradients across all zone boundaries.

* An anomaly in the waveform of the sine wave response of a certain differential pressure transmitter, not attributable to the saturation effects dealt with later, can be accounted for in this way. The author is indebted to Mr D. M. Bishop for drawing his attention to this case.

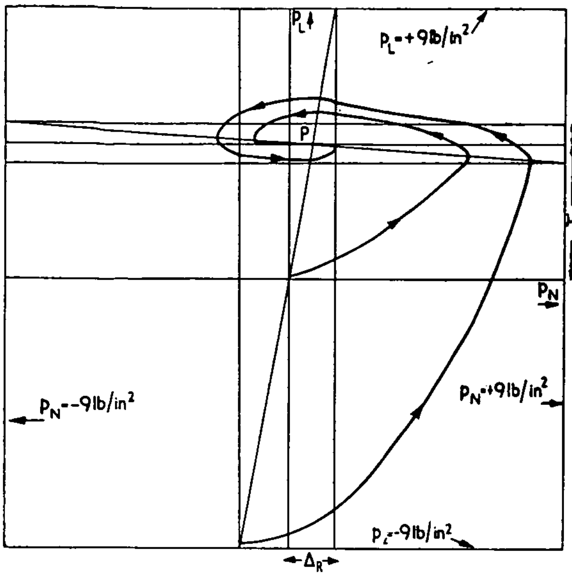


Fig. 6—Phase trajectories for controller driving simple load

The equations will now be considered for each region in turn.

Region of no saturation

$$\tau_N \frac{dp_N}{dt} = K_N(\mu p_i - p_L) - p_N \qquad (9)$$

$$\tau_L \frac{dp_L}{dt} = K_R\, p_N - p_L \qquad (10)$$

(writing $p_m - p_d = p_i$ for brevity). The differential equation of the trajectories is thus:

$$\frac{dp_L}{dp_N} = \frac{\tau_N}{\tau_L} \cdot \frac{p_L - K_R p_N}{K_N(p_L - \mu p_i) + p_N} \qquad (11)$$

and it is seen that the equilibrium point, P, is at the intersection of the lines:

$$p_N = -K_N(p_L - \mu p_i) \qquad (12)$$

and

$$p_L = K_R p_N \qquad (13)$$

as indicated in Fig. 6. P is a focal point if

$$1 + 2K_RK_N - 2\sqrt{\{K_RK_N(1 + K_RK_N)\}} < \frac{\tau_N}{\tau_L} < 1 + 2K_RK_N + 2\sqrt{\{K_RK_N(1 + K_RK_N)\}} \qquad (14)$$

and a nodal point otherwise. (For definitions of focal and nodal points see reference 7). Now $K_R\ K_N$ is normally quite large (in one widely used controller of this type, $K_R \bumpeq 9$ and $K_N \bumpeq 24$ at 100% proportional band) so the above inequalities reduce approximately to

$$\frac{1}{4K_RK_N} < \frac{\tau_N}{\tau_L} < 4\ K_RK_N \qquad (15)$$

When the controller is driving a large load such as a valve motor it is usually found that $\tau_N/\tau_L \ll 1$, but when the load is a receiving bellows τ_L is much smaller. Whether or not (15) is satisfied, and hence whether the response is oscillatory or over-damped, depends on the load, the construction of the particular controller, and the proportional-band setting at which it is operating. To be definite it will be assumed that P is a nodal point. The analysis for the case of a focal point is very similar. When P is a nodal point, the two asymptotic trajectories are straight lines through P with gradients

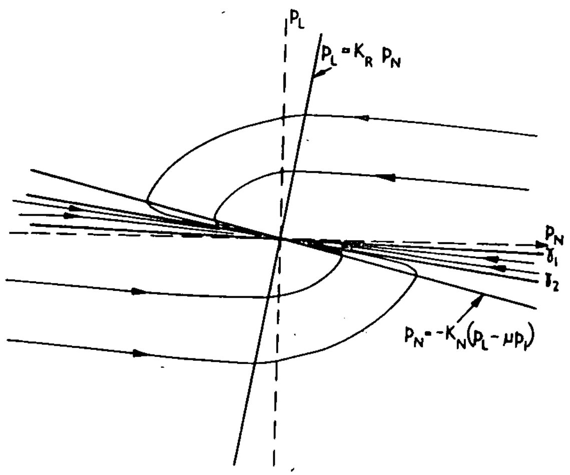


Fig. 7—Behaviour of the trajectories in the linear region

$$r_{1,2} = \frac{-\left(1-\frac{\tau_N}{\tau_L}\right) \pm \sqrt{\left\{\left(1-\frac{\tau_N}{\tau_L}\right)^2 - 4K_RK_N\frac{\tau_N}{\tau_L}\right\}}}{2K_N} \qquad (16)$$

and the behaviour of the trajectories in the neighbourhood of P is shown in Fig. 7. They can be sketched by noting that they must be tangential to the line r_2 at P and parallel to the line r_1 at infinity, and must cross (13) horizontally and (12) vertically.

Regions of saturation of both flapper-nozzle and relay

Consider the region

$$\mu p_i - \Delta_N > p_L \ , \ p_N > \Delta_R$$

in the lower right-hand corner of Fig. 6 as typical. The differential equations are

$$\tau_L \frac{dp_L}{dt} = 9 - p_L \ , \ \tau_N \frac{dp_N}{dt} = 9 - p_N \qquad (17)$$

and the trajectories are given by

$$\frac{dp_L}{dp_N} = \frac{\tau_N}{\tau_L} \cdot \frac{9 - p_L}{9 - p_N} \qquad (18)$$

They can be plotted by the method of isoclines[7] without much labour, since the isoclines are straight lines through $p_L = p_N = 9$, as can be seen from (18). The trajectories approach $p_L = p_N = 9$ as $t \to \infty$, and coincide with the isoclines if $\tau_N = \tau_L$. For $\tau_N < \tau_L$ they are convex in the direction of increasing p_N, while for $\tau_N > \tau_L$ they are convex in the direction of increasing p_L.

The differential equations in the remaining three saturation regions can be obtained from (17) simply by replacing +9 by —9 in one or both, so the trajectories have the same form but aim at one of the remaining three corners of the operating square. If the trajectories for one region are plotted on tracing paper, those for the remaining regions can be obtained conveniently by appropriate rotations and reversals of the tracing.

Regions of saturation of flapper-nozzle or relay

The differential equations in each of these regions can be written down in the same way, and the trajectories plotted by isoclines. Their main features are

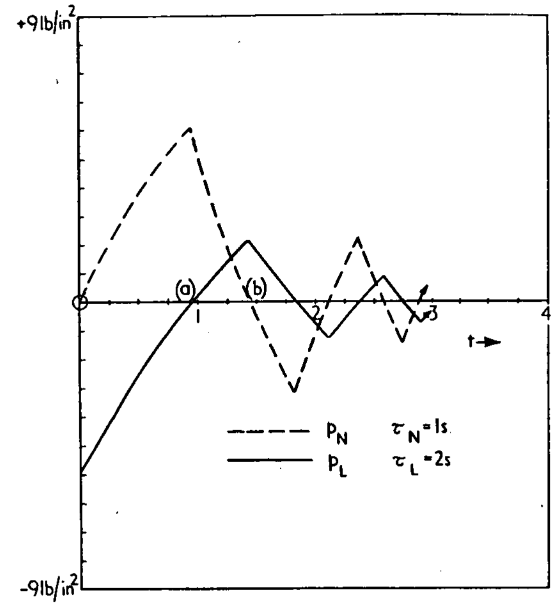


Fig. 8—Step response from the switching approximation

determined by the requirements that they must be horizontal when they cross (13) where the numerator of (11) vanishes, and vertical when they cross (12) where the denominator of (11) vanishes, and that they must join up smoothly with the trajectories in adjacent regions.

Having discussed the trajectories in the separate regions, it is now possible to join them together and deduce the complete transient response following a step change in p_i. Suppose that the system was in equilibrium with some value p_{io} of p_i for $t < 0$, and that p_i is suddenly changed at $t = 0$. The transient behaviour is then obtained by tracing out the trajectory starting from the initial equilibrium point, as shown in Fig. 6, where two trajectories corresponding to input steps of different sizes are shown. Although the time scale cannot immediately be deduced from the diagram, the principal features of the transient, such as the number and magnitude of overshoots before the variables settle to their final values, can be seen at a glance. For the larger step, p_L has two overshoots and p_N three overshoots before settling, while for the smaller step p_L has only one and p_N has two. It is clear from the construction that the numbers of overshoots will depend on the widths of the linear bands, and will increase as these widths decrease.

This behaviour can now be compared with the linear approximation, obtained by expanding the rectangle of no saturation about P until it covers the whole working region. In this case it is seen from Fig. 7 that p_L always tends monotonically to its final value, while p_N can have, at the most, one overshoot. The predictions of the two theories are therefore quite different. For any given transient it would, of course, be possible to choose the values of the gains K_N and K_R and the time constants τ_N and τ_L to give an oscillatory response in the linear approximation, with the correct value of the initial overshoot for p_L or p_N. Although the transient obtained in this way would have an infinite number of oscillations rather than the finite number predicted by the piecewise-linear theory, it might be very difficult to distinguish between the two experimentally if the damping were heavy. However, it will be shown that there are important discrepancies between the predictions of the linear theory and the present theory which can be decisively tested by experiment, and the results show that, during the first part of the transient at least, the predictions of linear theory bear no relation whatever to the observed facts.

Inspection of Fig. 6 suggests that for step disturbances of magnitude considerably greater than the width of the flapper-nozzle linear band, a good approximation to the transient, at least as far as the first overshoot, could be obtained by neglecting the finite widths of the linear bands completely and regarding the flapper-nozzle and relay as pure switches. This will be referred to as the switching approximation. It was noted above that $K_N \bumpeq 24$ for one well-known type of controller, so $\Delta_N \bumpeq \frac{3}{8}$ lb/in² and the requirement that the input step should be large compared with Δ_N is not a very serious restriction. Using the switching approximation, it is easy to work directly in the time domain, since at all times p_L and p_N are aiming exponentially at ±9 lb/in² with their respective time-constants. The two basic exponential curves can therefore be plotted, and the whole transient obtained by tracing segments of them between the switching points at $p_L = \mu p_i$ and $p_N = 0$. The method of construction is shown in Fig. 8, in which (*a*) is the first switching point with $p_L = \mu p_i$ and (*b*) the first switching point with $p_N = 0$. All the rising segments of the p_N curve represent the variation of pressure in the inlet chamber of the relay when the nozzle is completely sealed off by the flapper, while all the falling segments represent the pressure variations when the flapper is fully raised. Similarly, the rising segments of the p_L curve represent the variation of pressure in the load when the relay plug is in its fully raised position, while the falling segments represent the load pressure variations when the plug is fully lowered. The relay switches between its two states when the pressure in its inlet chamber passes through zero, and the flapper switches between its two states when the load pressure passes through its equilibrium value. Comparison of the two extreme approximations with the full piecewise-linear theory illustrated in Fig. 6 suggests that the switching approximation should give a good description of the first part of the

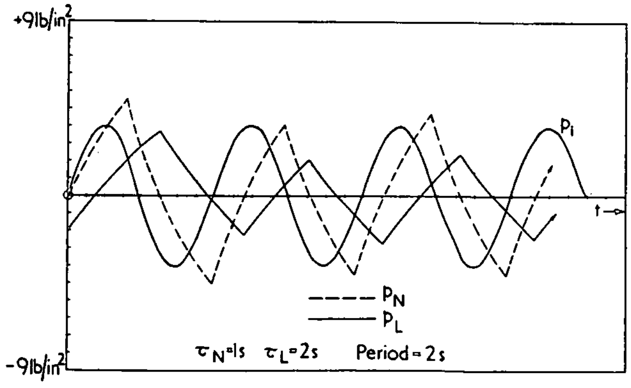


Fig. 9—Sine wave response from the switching approximation

transient following a large step disturbance, while the linear approximation should be applicable to the last part of the transient as the representative point approaches P.

The switching approximation can also be used to treat the response to a sinusoidal input of amplitude much larger than the flapper-nozzle linear bandwidth and frequency sufficiently high to call for rates of change of p_L significantly greater than those attainable with the relay in saturation. The construction is similar to that used for the step response, except that p_i, and hence the switching condition for the flapper, now varies sinusoidally with time. Any initial conditions may be chosen, since the initial transient is soon damped out, and the system settles into a periodic response with the same period as the input. The construction, which is largely self-explanatory, is shown in Fig. 9. The output waveform is almost triangular, as was remarked in the introduction. By repeating the construction for different input frequencies the attenuation and phase lag (suitably defined, for instance, in terms of the lag between corresponding peaks or zeros of input and output) can be obtained as functions of frequency, and both are found to increase with increasing frequency. This would also be the case with a linear theory of course. However, the construction can also be repeated with different amplitudes of input, when it is found that the attenuation and phase lag also increase with increasing amplitude, a result which could not be obtained from a linear theory.

All the constructions introduced in this section can equally well be applied to the case of a controller with a non-bleed relay of the type shown in Fig. 4 *a*. The only difference is that the relay switching condition is represented by a line of slope $1/a$ passing through the origin, rather than by the p_L axis. The existence of a finite dead band can be taken into account by drawing a strip about this switching line, and requiring that the trajectories should be horizontal straight lines inside it, since p_L cannot change within the dead band (assuming that both valves of the relay seat perfectly). The procedure follows that developed above so closely that it is not worth developing it in detail here.

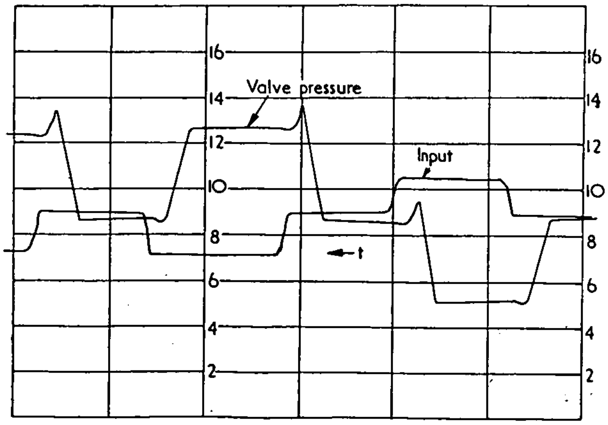


Fig. 10—Experimental step responses for controller driving simple load

TABLE I

First overshoot of p_L for various step changes

Initial Pressure, lb/in² gauge	Final Pressure, lb/in² gauge	Step Amplitude, lb/in²	Overshoot, lb/in²	Predicted Overshoot, lb/in²	
$5\frac{17}{32}$	9	$+3\frac{15}{32}$	0·94	0·93*	$\frac{\tau_N}{\tau_L} = 0{\cdot}18$
$3\frac{11}{32}$	9	$+5\frac{21}{32}$	1·03	0·99	
$6\frac{3}{8}$	9	$+2\frac{5}{8}$	0·81	0·87	
$12\frac{19}{32}$	9	$-3\frac{19}{32}$	0·25	0·25*	$\frac{\tau_N}{\tau_L} = 0{\cdot}04$
$14\frac{7}{8}$	9	$-5\frac{7}{8}$	0·25	0·25	
$14\frac{9}{16}$	9	$-5\frac{9}{16}$	0·31	0·25	

* Values used for fitting

COMPARISON WITH EXPERIMENT

Figure 10 is a photograph of responses to step changes of various magnitudes applied to the 'measured value' bellows of a floating-disc proportional controller with a continuous-bleed high-gain relay, feeding directly a load consisting of a No. 6* diaphragm valve motor. The volume of the motor varied between 25 in^3 and 55 in^3 as the pressure varied between 3 and 15 lb/in^2 gauge, and the pressure was recorded by connecting a bellows measuring-element to the motor with a short (less than 3 ft) length of $\frac{1}{4}$ in. internal dia. P.V.C. tube. All other connections were made with the same tubing and were of similar length. The apparent time lag between the input step and the response of the valve pressure is due to a physical separation of the recorder pens.

The general form of the transients is as predicted from Fig. 6, with three discernible overshoots for rising pressure and possibly two for falling pressure. Step responses were also obtained with very large input steps, so that the valve travelled through its whole range in both directions with the relay in a saturated condition. Comparison of these with the responses following smaller steps shows that the initial pressure variations in the transients of Fig. 10 take place with the relay in a saturated state.

Further confirmation of this is obtained from measurements of the first overshoot of p_L for step changes of various amplitudes. Some results are given in Table I for both rising and falling pressures, with a final pressure of 9 lb/in^2 gauge after each step to facilitate comparison with the predictions of the switching approximation. Pressures were estimated to $\frac{1}{32}$ lb/in^2 from the records, so quantities quoted as decimals are certainly not reliable to better than about ± 0·03 lb/in^2.

Any linear theory will predict an overshoot proportional to the amplitude of the step, but the observed overshoots are seen to vary much more slowly than this with step amplitude. The overshoot can easily be predicted from the switching approximation as illustrated in Fig. 8. On the first segment of the transient ($p_L<0$) the solutions for p_L and p_N which satisfy the appropriate initial conditions, $p_N = 0$ and

* The size number of a Foxboro-Yoxall motor.

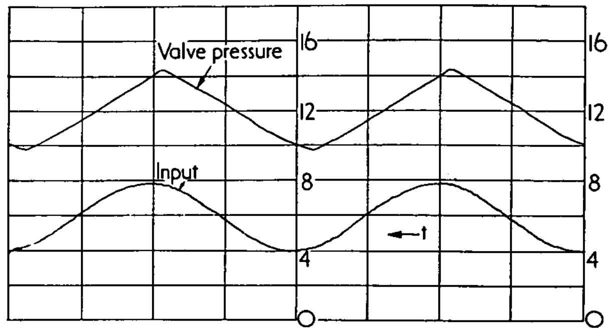


Fig. 11—Experimental sine wave response for controller driving simple load

p_L = —P for a step of amplitude $+P$, are clearly

$$p_N = 9(1—e^{-t/\tau_N}), \quad p_L = 9—(9 + P)e^{-t/\tau_L}$$

and $p_L = 0$ when

$$t = \tau_L \ln (1 + P/9)$$

If p_1 is the corresponding value of p_N, then

$$p_1 = 9 \left\{1—(1 + P/9)^{-\tau_L/\tau_N}\right\} \quad \text{.............................(19)}$$

Taking initial conditions $p_L = 0, p_N = p_1$ for the second segment, up to $p_N = 0$, the appropriate solution is

$$p_N = —9 + (9 + p_1)e^{-t/\tau_N}, \; p_L = 9(1 — e^{-t/\tau_L})$$

and $p_N = 0$ when

$$t = \tau_N \ln (1 + p_1/9)$$

which gives a value $p_L = p_L{}^0$ for the overshoot, where

$$p_L{}^0 = 9 \left\{1 — (1 + p_1/9)^{-\tau_N/\tau_L} \right\} \quad \text{.........................(20)}$$

Elimination of p_1 between (19) and (20) relates $p_L{}^0$ to P as required.

To compare these predictions with the experimental results given in the table, τ_N/τ_L must be known. τ_L can be estimated directly from the results of tests in which the valve is run through its full range with the relay saturated, but τ_N is not easy to measure directly because of the small size of the volume involved compared with the volume of any pressure-measuring element which might be attached to it. The procedure used, therefore, was to choose τ_N/τ_L to make the predicted and observed overshoots agree for one particular step amplitude, and to use the value obtained in this way to predict the overshoots for other amplitudes. Because of the inherent asymmetry of the flapper-nozzle, discussed above, it is necessary to take different values of τ_N/τ_L for rising and falling pressures. The fitted values of τ_N/τ_L are given to the right of the table and the experimental results used for fitting are indicated by asterisks. It is seen that the switching approximation is very successful in accounting for the relative insensitivity of the overshoot to the amplitude of the step change; in fact it predicts even less variation in overshoot than was observed. This might be expected, however, since the finite widths of the linear bands of both flapper-nozzle and relay have been neglected. When these bands extend over the whole working region, the system is linear and the overshoot is proportional to the amplitude of the step, so a result intermediate between the predictions of the linear approximation and the switching approximation might be expected from the complete treatment. This is confirmed by some graphical solutions.

Figure 11 is a photograph of the response to a large-amplitude sinusoidal input of period 4 seconds, from the same system as was used for Fig. 10. Comparison of the almost triangular waveform with the maximum speed responses for rising and falling pressures shows that the relay is saturated throughout except for short intervals in the neighbourhood of the maxima and minima, so the switching approximation would be appropriate for a theoretical treatment. As noted above, the width of the flapper-nozzle linear band in this case is about ⅔ lb/in^2, so a linear theory of the type used by Westcott[3] would only be strictly applicable for amplitudes less than ⅓ lb/in^2 if the frequency range of large phase lag is to be investigated. Unfortunately the phase-plane method illustrated in Fig. 6 does not lend itself to a treatment of sinusoidal inputs, so it has not proved possible to investigate theoretically the interesting transition between the regions of applicability of the two extreme approximations.

Provided sinusoidal inputs of sufficiently small amplitude are used, Westcott's procedure gives a perfectly valid method of obtaining the controller parameters which is probably much more accurate than the procedure described above, based on the step response. The latter, however, has the advantage of being very simple and quick to perform. However, although the frequency-response technique can be applied in the usual way for small amplitudes, where the linear approximation is valid, care must be exercised in using the results to predict the behaviour with other inputs. The well-known methods of deriving the step response from the frequency response, for instance, will give results which bear no relation whatever to the observed transients, except for the smallest step changes, and this situation cannot be improved by taking frequency-response measurements with sine waves of larger amplitude. The importance of the frequency response for a linear system is largely due to the fact that it is closely related to the response to any input of any magnitude, in fact the response to any input can theoretically be obtained from the frequency response. It therefore provides a convenient complete summary of the dynamical behaviour of the system. In the case of dynamical effects due to lags in the forward loop of a pneumatic controller, however, this property is lost, as has been shown, so the frequency response is of very limited value as a description of the dynamical behaviour.

SUSTAINED OSCILLATIONS

The switching approximation can give a good account of some features of the well known sustained oscillation or ' pumping ' which occurs in certain pneumatic systems. When a valve positioner feeds a valve through a booster relay, for instance, the system often oscillates with large amplitude unless special steps are taken to prevent this. To avoid possible complications due to friction in the gland of the valve, and backlash in the mechanical feedback linkage of a valve positioner, an equivalent system with pneumatic feedback was investigated, consisting of a controller with a high-gain continuous-bleed relay feeding a motor-valve through a booster relay.

The feedback to the controller was taken from the valve rather than the controller output, giving the arrangement shown in Fig. 12, which is essentially equivalent to the valve positioner with booster. The system oscillated with large amplitude and almost triangular waveform, and by comparing the oscillations with the maximum-speed responses of the valve driven by the booster, it can be seen that the booster was saturated except for short intervals in the neighbourhood of the maxima and minima.

The switching approximation should therefore be appropriate, and it can be applied using a construction method very similar to that used for Figs. 8 and 9, except that three pressures, p_N, p_B, and p_L, are now involved, so there are three fundamental response curves and three different switching conditions. By starting the construction with different initial conditions it can be shown that the final steady oscillations are independent of the initial conditions, as would be expected. A knowledge of the three time constants τ_N, τ_B, and τ_L is necessary to carry out the construction, though for quantitative predictions the observed full-speed responses of the valve for rising and falling pressures may be used instead of assuming exponential responses with time constant τ_L. For the controller used τ_N had already been estimated as described above, but τ_B was not known and could not, of course, be measured directly because of the small volume involved. The procedure used, therefore, was to choose τ_B so as to fit the predicted amplitude of oscillation to the observed amplitude with a No. 6 motor. Comparison of the observed and predicted periods then provided a test of the theory, but a better test could be obtained by using τ_B to predict both amplitude and period for a diffe,ent size of motor, and comparing the results with experiment. This was done for a No. 8 motor and both experimental and theorctical results for the two cases are given in Table II. The agreement is excellent. The observed amplitude was found to change very little with the proportional band, as would be expected since this only alters the width of the flapper-nozzle linear band to some extent.

A more complete treatment accounts for the way in which the stability of the system depends on τ_N, τ_B, and τ_L, both for the present system and for the corresponding system with a 1:1 non-bleed relay. In the latter case it is also possible, by taking account of the finite linear band of the flapper-nozzle, to account for the fact that the oscillations are found to be shock-excited rather than spontaneously excited. This is, of course, a phenomenon found only in non-linear systems.

TABLE II

Comparison of observed and predicted results with No. 6 and No. 8 motors

Motor	Observed Amplitude (peak-to-peak)	Observed Period	Predicted Amplitude (peak-to-peak)	Predicted Period
No. 6	4·0 lb/in^2	0·63 s	4·1 lb/in^2	0·67 s
No. 8	2·1 lb/in^2	1·0 s	2·3 lb/in^2	0·93 s

τ_N = 1/5 s (mean for rising and falling steps)

τ_B = 1/5 s (chosen to fit observed amplitude with No. 6 motor)

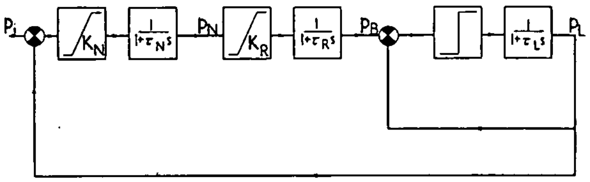


Fig. 12—Block diagram for oscillatory system

CONCLUSION

It has been shown that a piecewise-linear approximation, which takes account of the small width of the linear band of pneumatic amplifiers but neglects other sources of non-linearity, is capable of giving a good account of many features of the behaviour of simple pneumatic systems. The approximation in which the saturation effects are neglected leads to the usual linear theory, while the other extreme approximation, in which the linear band is neglected, is more appropriate for tieating large and rapid changes in input. The full treatment illustrates the transition between these two extremes and indicates which is the more appropriate in any given case.

The second, or switching, approximation has been shown to give a good account of the main features of the step response of a proportional controller driving a valve motor and also of the sustained oscillations which can occur in some more complicated systems.

It follows from this work that frequency response curves for a pneumatic system, obtained by using small-amplitude sine waves, do not give a complete summary of the dynamical properties of the system and can, indeed, lead to very misleading results if used to predict the effect of the inherent dynamics of the system on its behaviour.

Acknowledgments

The author wishes to thank Mr E. N. Martin for much useful discussion and constructive criticism, and Mr J. P. Mole for his careful attention to the experimental work. The work described is part of the research programme of Imperial Chemical Industries Limited, Billingham Division.

References

1. A. R. Aikman and C. I. Rutherford: 'The Characteristics of Air-Operated Controllers', Automatic and Manual Control, 1952, Butterworths.
2. J. M. L. Janssen: 'Analysis of Pneumatic Controllers', *Ibid.*
3. J. H. Westcott: *Trans. Soc. Instr. Tech.*, 1957, vol. 9, p. 79.
4. D. B. Kirk: *Trans. A.S.M.E.*, 1948, vol. 15, p. 111.
5. C. R. Webb: *Trans. Soc. Instr. Tech.*, 1955, vol. 7, p. 9.
6. P. S. Buckley: *Proc. I.S.A.*, 1955, vol. 10, No. 1.
7. N. Minorsky: 'Introduction to Non-Linear Mechanics', 1947, Edwards.
8. L. A. Gould and P. E. Smith jun.: 'Dynamic Behaviour of Pneumatic Devices', *Instruments*, 1953, vol. 26, Nos. 6 and 7.
9. N. Ream, R. H. Tizard, and D. S. Townend: 'Plant and Process Dynamic Characteristics', p. 145: 1957, Butterworths.

A3

The Behaviour of Linear Systems with Inputs satisfying certain Bounding Conditions†

By B. J. Birch
Trinity College, Cambridge
and R. Jackson
Imperial Chemical Industries, Ltd., Billingham Division

[Received March 17, 1959]

Abstract

Linear filters with bounded inputs give outputs which are also bounded. A method is described for obtaining the least upper bound of the output for the case where bounds are specified both for the magnitude of the input and its rate of change. The result, which is obtained in the form of a procedure for constructing that input (satisfying the bounding conditions) which gives rise to the largest output, has applications in the design of automatic control systems.

§ 1. Introduction

The purpose of the simplest type of automatic control system is to maintain some measured quantity $y(t)$ close to a specified desired value Y, in spite of the effect of a disturbance $i(t)$. At present, in process control applications, it is extremely difficult to obtain any but the most elementary properties of $i(t)$ at the design stage, so it is not usually possible to design to a specification such as the mean square error criterion, for which a knowledge of the disturbance power spectrum is required. It should, however, be possible to devise some design method in which the absence of desirable information about $i(t)$ does not lead to complete impotence, but rather to an overdesigned system in which the degree of overdesign (or 'safety factor') is a measure of the designer's ignorance of the detailed form of $i(t)$.

The present work describes the mathematical basis of a method of this type which uses only the elementary type of information about the disturbance likely to be available at the design stage. The desired performance is specified by giving a band $Y \pm y_m$ within which the measured variable must remain. In order to design to this specification it is necessary that the disturbance should be bounded, and if bounds $I \pm i_m$ for the variation of $i(t)$ are known, the method produces the most economical design which is safe in the sense that $y(t)$ will never pass outside the range $Y \pm y_m$ however $i(t)$ varies, provided $i(t)$ remains within the range $I \pm i_m$. If extra information that the rate of change of $i(t)$ must lie within the range $\pm i_m'$ is also

† Communicated by the Authors.

available, it can be incorporated in the design procedure to give a more economical system which is nevertheless still safe, in the sense that $y(t)$ will never pass outside the range $Y \pm y_m$ however $i(t)$ varies, provided it remains within $I \pm i_m$ and does not change at a rate greater than i_m'. This type of information about the disturbance (i.e. values of i_m and i_m') can often be estimated at the design stage from physical limitations in the plant and control apparatus (available pressure drops, resistances of pipes and fitments, speeds of valve motors, etc.).

The basic mathematical problem may be stated as follows: in a stable linear automatic control system, y at time t is related to values of $i(t)$ at all previous times by an equation of the form

$$y(t) = \int_0^\infty W(u) i(t-u)\, du \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

where the weighting function $W(u)$ characterizes the particular system considered, and the stability condition demands that $\int_0^\infty |W(u)| du$ should exist. It is assumed that origins for $y(t)$ and $i(t)$ are chosen so that $I = Y = 0$ and that the following limitations are imposed on the behaviour of $i(t)$:

$$|i(t)| \leqslant i_m \qquad \text{(all } t\text{)} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

$$\frac{|i(t_1) - i(t_2)|}{|t_1 - t_2|} \leqslant i_m' \qquad \text{(all } t_1, t_2\text{)} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (3)$$

and it is required to find the corresponding least bounds $\pm y_m$ for the variation of $y(t)$. The result will be obtained in the form of a procedure for constructing that $i(t)$ which makes y, at some specified time, as large as possible. The corresponding value of $y(=y_m)$ can then be obtained from (1).

§ 2. Conditions for the Extremal $i(t)$

When there is no restriction on the permissible rate of change of $i(t)$, a possible input is

$$\begin{aligned} i(t-u) &= +i_m \quad \text{when } W(u) > 0 \\ &= -i_m \quad \text{when } W(u) < 0 \end{aligned}$$

with which, from (1),

$$y = i_m \int_0^\infty |W(u)|\, du \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (4)$$

and the integral exists for a stable system, as already noted. But for any $i(t)$ satisfying (2), it follows from (1) that

$$y \leqslant i_m \int_0^\infty |W(u)|\, du$$

so (4) gives the required y_m. The problem is therefore trivial unless the additional constraint (3) is imposed.

In dealing with the general problem the choice of the particular value of t at which $y(t)$ is considered is clearly arbitrary, and it is convenient to take $t=0$, and to write $j(u)$ for $i(-u)$ and y for $y(0)$. (1) then becomes

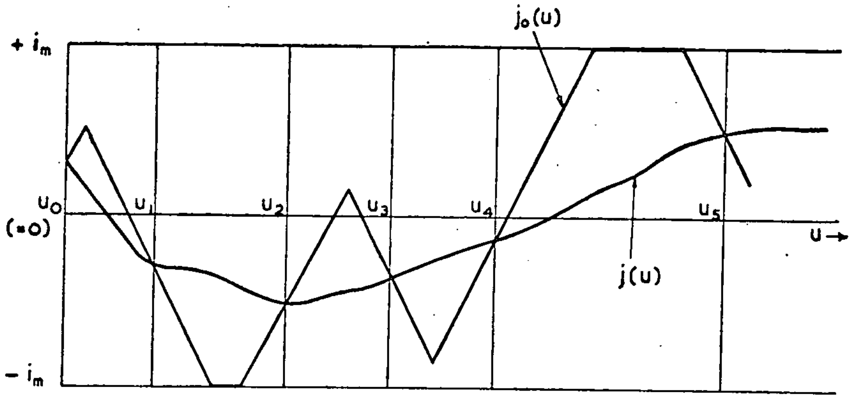
$$y=\int_0^\infty W(u)j(u)\,du. \qquad \qquad (5)$$

Consider an arbitrary $j(u)$ plotted as a function of u and let $u_0(=0)$, u_1, $u_2, \ldots$ be the points at which $W(u)$ changes sign, arranged in increasing order of magnitude. From each $j(u)$ satisfying (2) and (3) construct a new function $j_0(u)$ as follows:

(i) If $u_n \to u_{n+1}$ is an interval in which $W(u) \geqslant 0$, draw a line of slope $+i_m'$ through $[u_n, j(u_n)]$ and a line of slope $-i_m'$ through $[u_{n+1}, j(u_{n+1})]$. If these intersect at some $j \leqslant i_m$, $j_0(u)$ consists of these two line segments in $u_n \to u_{n+1}$; if they intersect for $j > i_m$, complete $j_0(u)$ in this interval by joining them with a line segment along $j=+i_m$.

(ii) If $u_n \to u_{n+1}$ is an interval in which $W(u) \leqslant 0$ proceed in a similar way but reverse the signs of the slopes of the two lines, and join along $j=-i_m$ if they intersect for some $j<-i_m$.

Fig. 1



The complete $j_0(u)$ obtained in this way is a sequence of straight line segments as shown in fig. 1, which illustrates the construction. Since it will be necessary to refer fairly frequently to the salient features of functions of this type, it will be convenient at this stage to give them short descriptive names. Segments of slopes $\pm i_m'$ will be called 'zigs' and 'zags' respectively, while horizontal segments at $\pm i_m$ will be called 'plateaux' and 'valleys'. A zig may end at the start of a plateau or at the start of a zag, the sharp point formed in the second case being known as a 'peak', while the corresponding feature when a zag ends at the beginning of a zig will be called a 'ditch'. A sequence of zigs and zags without any intervening plateaux or valleys will be called a 'zigzag'.

Clearly each $j_0(u)$ satisfies (2) and (3) and the set $\{j_0(u)\}$ is a subset of the set $\{j(u)\}$ of all permissible inputs; the same $j_0(u)$ is obtained from all $j(u)$'s which are equal at the points $u=u_n$. Further, any zigzag $j(u)$

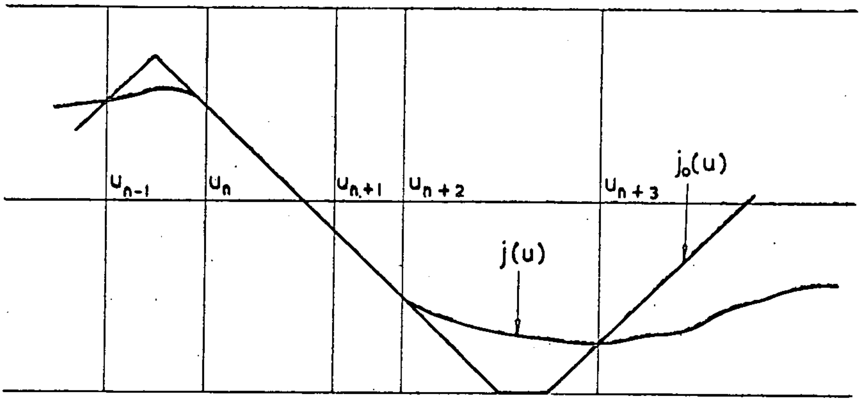
obtained by drawing straight line segments of alternate slopes $+i_m'$ and $-i_m'$, and joining them by horizontal segments at $j = \pm i_m$ if necessary, is a $j_0(u)$ provided:

(*a*) $j'(u_n) = +i_m'$ if $W(u_n+\epsilon) > 0$ for sufficiently small $\epsilon > 0$.

(*b*) $j'(u_n) = -i_m'$ if $W(u_n+\epsilon) < 0$ for sufficiently small $\epsilon > 0$.

(*c*) Between successive points u_n and u_{n+1} there is only one segment with slope $+i_m'$ and one with slope $-i_m'$. (In some cases the segment of slope $\pm i_m'$ associated with some u_n may appear to be missing; however, for formal reasons we regard it as present but of zero length. Figure 2 shows how this may occur in constructing the $j_0(u)$ corresponding to a $j(u)$ which maintains a constant slope $-i_m'$ over two adjacent intervals $u_n \to u_{n+1}$ and $u_{n+1} \to u_{n+2}$.)

Fig. 2



A particular $j_0(u)$ is uniquely determined by (*a*), (*b*), (*c*) and the positions of its sloping segments, specified for instance by giving their intercepts (produced if necessary) with any line.

If $j_0(u)$ is constructed from $j(u)$ as described it is seen that

$$j_0(u) \geqslant j(u) \quad \text{whenever } W(u) > 0$$

$$j_0(u) \leqslant j(u) \quad \text{whenever } W(u) < 0$$

so from (5) we have the result:

Theorem 1

If y is the output (at $t=0$) given by any input $j(u)$ satisfying (2) *and* (3), *and y_0 the output given by the corresponding $j_0(u)$, then $y_0 \geqslant y$.*

In seeking the greatest y we may therefore restrict attention to the subset $\{y_0\}$ generated by the inputs $\{j_0(u)\}$. It has already been noted that, for stability, $\int_0^\infty |W(u)|\,du$ exists, so there exists some U such that $\int_U^\infty |W(u)|\,du < \epsilon$, with arbitrarily small ϵ. The error produced in y by truncating the integral (5) at $u = U$ will certainly not exceed ϵi_m for any

$j(u)$ satisfying (2); thus, the integral may be truncated after a finite interval, with arbitrarily small loss in accuracy. For convenience, we will suppose from now on that this has been done, and that $W(u)$ has only finitely many zeros in the range $0 \leqslant u \leqslant U$, namely $u_1, u_2, \ldots u_N$.

A $j_0(u)$ is uniquely specified by the positions of all its zigs and zags; it is convenient to define the position of the nth zig or zag (that is the one corresponding to the zero u_n of W; as we remarked after (*c*), this may be of zero length) by giving its intercept x_n with the line $i=0$, measured from u_n as origin. Since u_n must lie on this sloping segment,

$$-\phi \leqslant x_n \leqslant \phi \quad \text{for } n=1, 2, \ldots N \qquad \ldots\ldots \quad (6)$$

where $\phi = i_m/i_m'$. Since the sloping segments corresponding to u_n and u_{n+1} must meet between u_n and u_{n+1},

$$u_n \leqslant \tfrac{1}{2}(u_n + u_{n+1} + x_n + x_{n+1}) \leqslant u_{n+1}. \qquad \ldots\ldots \quad (7)$$

Conversely, to any $(x_1, \ldots . x_N)$ satisfying (6) and (7) there corresponds a $j_0(u)$. Accordingly, $j_0(u)$ may be represented by a point $(x_1, \ldots . x_N)$ of an N-dimensional cube of side 2ϕ, and indeed points corresponding to j_0's fill up a convex subset of this cube (that is, if $(x_1, \ldots . x_N)$ and $(y_1, \ldots . y_N)$ both give j_0's, then all the points of the segment joining them give j_0's). The problem is therefore reduced to that of finding the greatest value of a function of N variables whose variation is restricted by (6) and (7).

It is certainly possible to find an extremal $j_0(u)$, giving a stationary value of y in our truncated problem (indeed, it may be proved that an extremal exists even if the integral (5) is left untruncated). As a first step, conditions for an extremal $j_0(u)$ will be found, and then later we will show that every extremal must give the maximal value for y.

Theorem 2

$j_0(u)$ is extremal if and only if

$$\int_a^b W(u)\,du = 0$$

whenever a and b are the values of u at the beginning and end of any zig or zag.

Proof

Figure 3 (1) shows an adjacent zig and zag of some $j_0(u)$, meeting in a peak. The end points of the sloping segments are a, b and c as indicated. Consider a small variation to give $\tilde{j}_0(u)$, in which the zig is displaced by δx to the right, and the zag by δy, but $j_0(u)$ and $\tilde{j}_0(u)$ are otherwise identical. $\tilde{j}_0(u)$ is indicated in fig. 3 (1) by broken lines, and the difference $\tilde{j}_0(u) - j_0(u)$ is plotted in fig. 3 (2). This is split into two components in figs. 3 (3) and 3 (4); it is obvious that the sum of these gives $\tilde{j}_0(u) - j_0(u)$. When

$$\int_0^\infty W(u)[\tilde{j}_0(u) - j_0(u)]\,du$$

is formed, the component in fig. 3 (3) gives an expression of first order in δx, δy:

$$-i_m'\,\delta x \int_a^b W(u)\,du + i_m'\,\delta y \int_b^c W(u)\,du \qquad \ldots \quad (8)$$

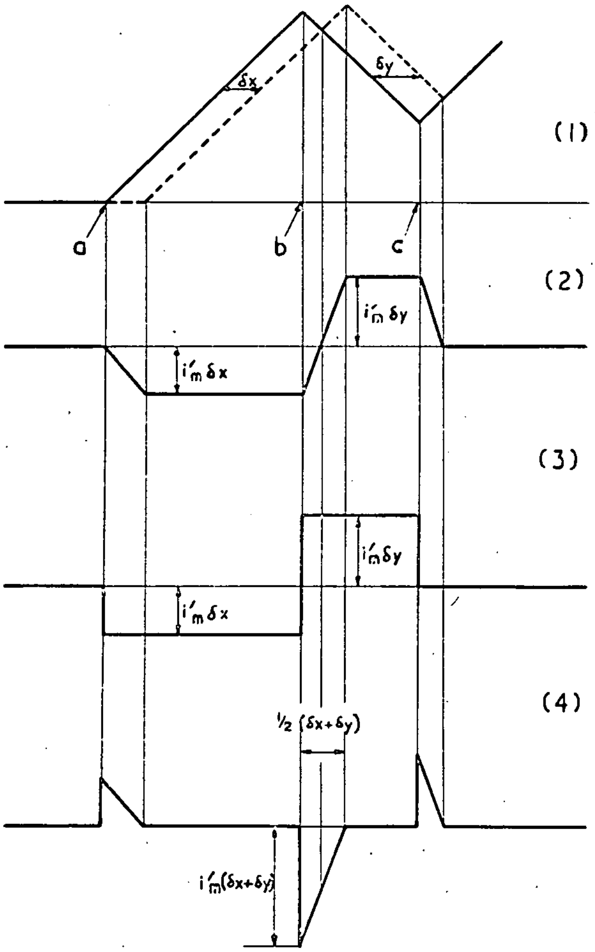
while the component in fig. 3 (4) gives higher order contributions. Necessary and sufficient conditions for an extremal $j_0(u)$ are that (8) should vanish for arbitrary δx and δy, for all zigs and zags, proving the theorem.

Consideration of the higher order contributions of fig. 3 (4) allows us to prove

Theorem 3

An extremal $j_0(u)$ gives a stationary value of y_0 which is a maximum and is the greatest attainable value.

Fig. 3



Indication of proof

As in the proof of Theorem 2, consider fig. 3 (1). We already know that the first-order contribution from the neighbourhood of the peak at b to the variation of y_0 is given by (8); now we consider the second-order contribution. The corresponding component of $\tilde{j}_0(u)-j_0(u)$ is sketched in fig. 3 (4) ; in the diagram, this component is non-zero in the following intervals:

(i) $(a, a+\delta x)$, where $\tilde{j}_0(u)-j_0(u)>0$ and $W(u)<0$.

(ii) $(b, b+\frac{1}{2}\delta x+\frac{1}{2}\delta y)$, where $\tilde{j}_0(u)-j_0(u)<0$ and $W(u)>0$.

(iii) $(c, c+\frac{1}{2}\delta y)$, where $\tilde{j}_0(u)-j_0(u)>0$ and $W(u)<0$.

Thus, the contribution to $\int_0^\infty W(u)[\tilde{j}_0(u)-j_0(u)]\,du$ is negative; other particular cases should be examined, corresponding to ditches, valleys, plateaux and zigs of zero length—but in all cases the second-order variations of $\tilde{y}_0$ from y_0 turn out to be negative.

Now, suppose that $\bar{j}_0(u)$ is an extremal, let $\tilde{j}_0(u)$ be any other $j_0(u)$, and let $(\bar{x}_1, \ldots \bar{x}_N)$, $(\tilde{x}_1, \ldots \tilde{x}_N)$ be the corresponding points of the cube (6). Join these two points by a straight line l, and let s increase from 0 to 1 as we go from $(\bar{x}_1, \ldots \bar{x}_N)$ to $(\tilde{x}_1, \ldots \tilde{x}_N)$ along l. Each point of l gives a $j_0(u)$, and so to each s in $0\leqslant s\leqslant 1$ there corresponds a value $y_0(s)$ of y_0. By definition $y_0(0)=\bar{y}_0$ and $y_0(1)=\tilde{y}_0$. Since $\bar{j}_0$ is an extremal, $dy_0/ds=0$ for $s=0$, and by the preceding paragraph, $d^2y_0/ds^2\leqslant 0$ for all s. Hence $\tilde{y}_0\leqslant\bar{y}_0$, which proves the theorem.

It follows similarly that if $\bar{x}$ and $\bar{x}'$ are two stationary points, they may be joined by a line consisting entirely of stationary points, so that y is constant. This situation actually arises, so it is not correct to suppose that there is always a unique extremal.

§ 3. Procedure for Constructing the Extremal $j_0(u)$

Having formally obtained the conditions for a stationary value in Theorem 2 and ensured that this gives the greatest attainable value of y, a procedure for constructing the extremal input will now be discussed. This procedure is essentially rather simple, but it is very tedious to describe.

First note that the step response is the integral of the weighting function, so that Theorem 2 may be restated: a necessary and sufficient condition for an extremal $j_0(u)$ is that the values of the step response at the beginning and end of each zig or zag should be equal. It will be assumed that the step response is initially positive for the physical systems of interest. It may, in fact, happen that $S(u)$ has a positive step discontinuity at the origin, but this may be neglected for the purpose of the following argument, since it may be replaced by a line of finite but sufficiently large gradient with arbitrarily small error. In practice, either $W(u)$ has only a finite number of zeros, or it may be truncated after a finite number with arbitrarily small error. Let the zeros of interest be $u_1, u_2, \ldots u_N$.

Choose an arbitrary $u=U_1$ between $u=0$ and the first maximum of $S(u)$ and let the roots of $S(u)=S(U_1)$ be $u=U_1$, U_2, ... in increasing order of magnitude. Denote by U_T either the largest of these or the first which is greater than u_N, whichever is smaller. Now construct a continuous $j(u)$ as follows :

$$\left.\begin{aligned} j(u) &= +i_m \quad \text{for} \quad 0 \leqslant u < U_1, \\ j'(u) &= -i_m' \quad \text{for} \quad U_{2n-1} < u < U_{2n} \quad (n=1,2,\ldots), \\ j'(u) &= +i_m' \quad \text{for} \quad U_{2n} < u < U_{2n+1} \quad (n=1,2,\ldots). \end{aligned}\right\} \quad . \quad . \quad (9)$$

This satisfies the stationary condition of Theorem 2 but does not necessarily satisfy the basic constraint (2), nor does it necessarily define $j(u)$ over the whole region of interest $0 \rightarrow u_N$, since U_T may be less than u_N.

Starting at $u=0$ and proceeding in the direction of increasing u, look for the first point at which (9) ceases to be an acceptable input, which we shall call the first 'difficulty'. This may occur for one of three reasons:

(i) A peak projects above $+i_m$;

(ii) A ditch projects below $-i_m$;

(iii) Neither of these occurs, but $U_T < u_N$ and $j(U_T) \neq \pm i_m$,

so that the zigzag does not terminate in an acceptable manner. In case (i) it is necessary to lower the peak until it touches $+i_m$, in case (ii) to raise the ditch until it touches $-i_m$, while in case (iii) it is necessary to adjust $j(U_T)$ until it lies on $+i_m$ if it is the end point of a zig or on $-i_m$ if the end point of a zag. In all cases the necessary adjustments involve the raising or lowering of the end points of zigs or zags, and these two processes can be quite simply carried out, as will now be shown.

It is easy to see that moving U_1 to the right lengthens all the zigs and shortens all the zags, thus raising all the peaks and ditches, while moving U_1 to the left lowers all peaks and ditches. By letting U_1 reach the first maximum of $S(u)$, the raising can be continued until the first zag vanishes, but the lowering process can be continued beyond the point at which $U_1=0$ since the first zag need not start on $+i_m$ if it starts at $u=0$. Thus with $U_1=0$ it is possible to lower the whole zigzag bodily simply by lowering the start of the first zag. The peaks and ditches are located at the roots of $S(u)=0$ throughout this operation.

Returning now to case (i), where a peak projects above $+i_m$, the procedure is to lower the peak as just described. In this way it may be possible to bring the peak down to $+i_m$ without any new difficulties being generated nearer to the origin than the one being removed, but in general one of three things may happen during the lowering process.

(*a*) One or more zigs may contract to vanishing length and one or more zags, originally of vanishing length, may expand to a finite length. This arises because of the appearance or disappearance of roots of $S(u)=S(U_1)$, as U_1 varies, but it gives rise to no further difficulties and may be neglected, apart from taking care that it does not lead to the careless placing of zigs in intervals which should be occupied by zags and vice versa.

(*b*) More roots of $S(u)=S(U_1)$ may appear in $u > U_T$, so that the zigzag extends to some new $U_{T'} > U_T$. This may remove a difficulty at the original U_T and give rise to further difficulties in $u > U_T$, but these should be temporarily ignored, concentrating attention on the first difficulty.

(*c*) One of the ditches in the region before the difficulty being removed may reach $-i_m$, when further lowering by the process described would give rise to a new difficulty, with a ditch projecting below $-i_m$. The procedure is then to move the zig at this ditch to the right, opening a valley where the ditch touched $-i_m$. All the peaks and ditches to the right of this valley are now located by the condition $S(u)=S(U_V)$ where U_V is the right-hand end point of the valley which has been opened. Moving the zig to the right to open the valley has the effect of shortening all subsequent zigs and lengthening zags, and therefore continues the lowering process.

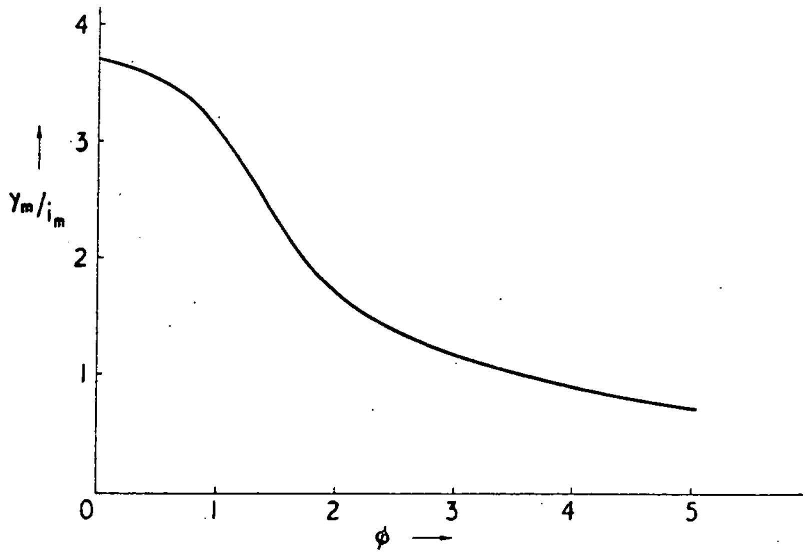
(*d*) When dealing with the first difficulty, there will not yet be any plateaux. However, in dealing with later difficulties, it may be necessary to close a plateau or valley by a process similar to that of (*c*).

By continuing in this way, opening several valleys if necessary, and possibly abolishing a few plateaux, the peak giving rise to the difficulty may be lowered till it touches $+i_m$, thus removing the difficulty. Attention is then transferred to the next difficulty in order of increasing u. Case (ii), where the first difficulty was a ditch, need not be discussed separately, as the above discussion is applicable with the words 'peak' and 'lower' replaced by 'ditch' and 'raise'. In case (iii) where the first difficulty is at U_T the same raising or lowering process is carried out in an attempt to bring $j(U_T)$ to $+i_m$ if it is the end point of a zig, or to $-i_m$ if it is the end point of a zag. In the course of this any of (*a*), (*b*) and (*c*) above may occur, and (*b*) now has the effect of removing the difficulty at U_T, but in every case where a discontinuous process of this type occurs, attention is transferred to the new difficulty with the smallest value of u, and the modification process is continued until one of two things occurs. Either a zigzag input is obtained with no difficulties up to and including the first peak or ditch beyond $u=u_N$, which provides a solution of the problem, or a zigzag terminating correctly on $\pm i_m$ at some $u<u_N$ is obtained. In the second case it is necessary to choose a starting point U_2 for a new zigzag between the end point of the original one and the next zero of $W(u)$, locating the peaks and ditches at the roots of $S(u)=S(U_2)$. This zigzag may be raised or lowered to remove difficulties as before, and the process can be continued until a completely satisfactory $j(u)$ extending beyond $u=u_N$ is obtained.

The procedure just described is formally a complete method for generating an extremal $j_0(u)$ and it could be carried out graphically for any given $W(u)$. It is also possible, in principle, to programme the procedure for automatic performance on a digital computer. In cases where $W(u)$ is poorly damped and a large number of zeros must be considered to achieve acceptable accuracy, a very large amount of computation would be involved and a very fast machine would be needed to obtain the results in a reasonable time. In the case of systems satisfying second order differential equations, however, $W(u)$ is either overdamped, in which case it has only one zero, or it is oscillatory, in which case the zeros $u_1, u_2, \dots$ are equally spaced and $W(u)$ differs in the intervals $u_n \rightarrow u_{n+1}$ and $u_m \rightarrow u_{m+1}$ only by a scale factor. For an overdamped case it is clearly practicable to consider all the different cases which may occur, since their number is not very large, and write a simple programme to identify the appropriate case and calculate the correct answer from a predetermined formula. The same is true for the oscillatory case as it is only necessary to consider separately the first interval and one other, the positioning of zigs and zags being the same for all subsequent intervals since $W(u)$ changes only by a scale factor. The total contribution to y from all intervals beyond the first then appears as the sum of a geometric series, to which the contribution from the first interval must be added. Programmes of this type have been written for an Elliott 402 computer, which accepts specifications of the step response $S(u)$ in the form:

$$\begin{aligned} S(u) &= A + B\exp(-pu) + C\exp(-qu) \quad \text{for the overdamped case} \\ &= A + B\sin\pi\nu + B\exp(-pu)\sin(qu - \pi\nu) \quad \text{for the oscillatory case} \end{aligned}$$

Fig. 4



together with a specification of $\phi(=i_m/i_m')$, and prints out the corresponding value of y_m/i_m. Figure 4 shows a typical set of results for an oscillatory step response of the above form, with $A=0{\cdot}64$, $B=1$, $p=0{\cdot}2$, $q=1$, $\nu=1{\cdot}8$. The ratio y_m/i_m is plotted as a function of ϕ from points computed along the curve, and is seen to decrease monotonically as ϕ increases, which is what would be expected of course. The programmes have been used in an investigation of level control systems, the results of which it is hoped to publish in a subsequent paper.

Acknowledgments

One of us (R. Jackson) wishes to thank the Directors of Imperial Chemical Industries Limited, Billingham Division, for permission to participate in this publication.

The Design of Control Systems with Disturbances Satisfying Certain Bounding Conditions, with Application to Simple Level Control Systems

R. JACKSON

ntroduction

'he purpose of the simplest type of automatic control system s to maintain some measured quantity $y(t)$ close to a constant lesired value Y, in spite of the effect of a disturbance $i(t)$. Although, in process control applications, it is extremely lifficult to obtain any but the most elementary properties of (t) at the design stage, the present work arose from the coniction that, faced with this situation, it should be possible to evise some design method in which the absence of desirable nformation about $i(t)$ does not lead to complete impotence, ut rather to an overdesigned system in which the degree of verdesign (or 'safety factor') is a measure of the designer's gnorance of the detailed form of $i(t)$.

The first three sections of this paper give an outline of some asic mathematical results which have been discussed elsewhere[1] in greater detail; the remainder of the paper then shows ow these results can be used to give a design method of the ype described above, using as illustration the very simple case f a level control system.

The basic mathematical problem may be stated as follows: n a stable linear filter the output $y(t)$ at time t is related to alues of the input $i(t)$ at all previous times by an equation of ne form

$$y(t) = \int_0^\infty W(u)i(t-u)\,\mathrm{d}u \tag{1}$$

here the weighting function $W(u)$ characterizes the particular ystem considered and the stability condition demands that

$$\int_0^\infty |W(u)|\,\mathrm{d}u$$

hould exist. It is assumed that $i(t)$ is uniformly bounded and iat the origin for $i(t)$ is chosen so that

$$|i(t)| < i_m \qquad \text{(all } t\text{)} \tag{2}$$

/e shall also assume that the rate of change of $i(t)$ is uniformly ounded or, slightly more generally, that

$$\frac{|i(t_1) - i(t_2)|}{|t_1 - t_2|} < i_m' \qquad \text{(all } t_1, t_2\text{)} \tag{3}$$

he problem is then to find the corresponding smallest bound $_m$ such that $|y(t)| < y_m$ for all t. The result will finally be btained in the form of a procedure for constructing that $i(t)$ hich makes y, at some specified time, as large as possible. he corresponding value of $y(=y_m)$ can then be obtained from quation 1.

The solution c. this problem can be used directly in the esign of automatic control systems when the performance is specified by giving limits $\pm y_m$ between which the controlled quantity must remain, and when the only information available about the disturbance is of the simple type specified in equations 2 and 3 above. For each trial design y_m can be calculated in terms of i_m and i_m', and a final design adopted which makes y_m satisfactorily small. The reason for choosing this method, rather than well known methods based, for example, on the mean square error, is that the information about $i(t)$ contained in equations 2 and 3 can often be estimated at the design stage from purely physical limitations in the plant and control apparatus (available pressure drops, resistances of pipes and fitments, speed of operation of valve motors, etc.), whereas more sophisticated properties of the disturbance such as its spectral density, which would be required for mean square error calculations, are not usually available at the design stage.

In spite of its apparent simplicity, the case of level control systems is important because of the high cost of vessels designed for high-pressure work or constructed from expensive materials such as stainless steel. Application of the present method enables the minimum vessel size which is compatible with a specified control performance to be determined very simply, and leads to a vessel which is certainly safe in the sense that it is large enough to achieve the specified performance with the worst possible disturbance, subject to restrictions of the type described above. In many practical applications the size obtained in this way is much smaller than the minimum necessary for the vessel to fulfil its primary purpose (e.g. minimum size for adequate disentrainment in a catchpot), in which case considerations of controllability are not a limiting factor in the design.

Conditions for an Extremal $i(t)$

Let us consider first the basic mathematical problem stated in the introduction in the particular case when there is no limitation imposed on the rate of change of $i(t)$ and equation 2 gives the only constraint on its behaviour. Then a possible input is

$$\begin{aligned} i(t-u) &= +i_m \qquad \text{when } W(u) > 0 \\ &= -i_m \qquad \text{when } W(u) < 0 \end{aligned}$$

with which, from equation 1

$$y = i_m \int_0^\infty |W(u)|\,\mathrm{d}u \tag{4}$$

and the integral exists for a stable system, as already noted. But for any $i(t)$ satisfying equation 2, it follows from equation 1 that

$$y < i_m \int_0^\infty |W(u)|\,\mathrm{d}u$$

o equation 4 gives the required y_m. The problem is therefore rivial when the additional ~~constant equation~~ 3 is not imposed.

In dealing with the general problem, the choice of the partiular value of t at which $y(t)$ is considered is clearly arbitrary, nd it is convenient to take $t = 0$ and to write $j(u)$ for $i(-u)$ nd y for $y(0)$. Equation 1 then becomes

$$y = \int_0^\infty W(u)j(u)\,\mathrm{d}u \qquad (5)$$

Consider an arbitrary $j(u)$ plotted as a function of u and let $_0(=0)$, $u_1, u_2 \ldots$ be the points at which $W(u)$ changes sign,

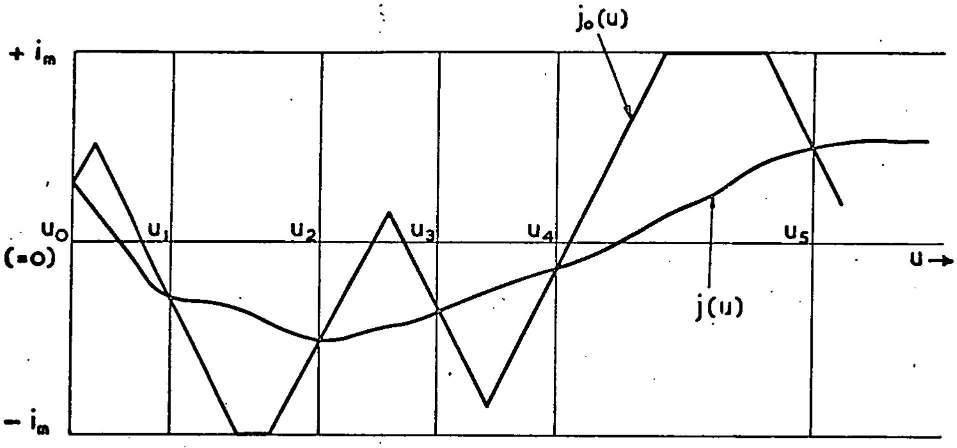


Figure 1. Construction of $j_0(u)$

rranged in increasing order of magnitude. From each $j(u)$ tisfying equations 2 and 3 construct a new function $j_0(u)$ as .lows:

(a) If $u_n \rightarrow u_{n+1}$ is an interval in which $W(u) \geqslant 0$, draw a ne of slope $+i_m'$ through $[u_n, j(u_n)]$ and a line of slope $-i_m'$ rough $[u_{n+1}, j(u_{n+1})]$. If these intersect at some $j \leqslant i_m, j_0(u)$ onsists of these two line segments in $u_n \rightarrow u_{n+1}$; if they tersect for $j > i_m$, complete $j_0(u)$ in this interval by joining em with a line segment along $j = +i_m$.

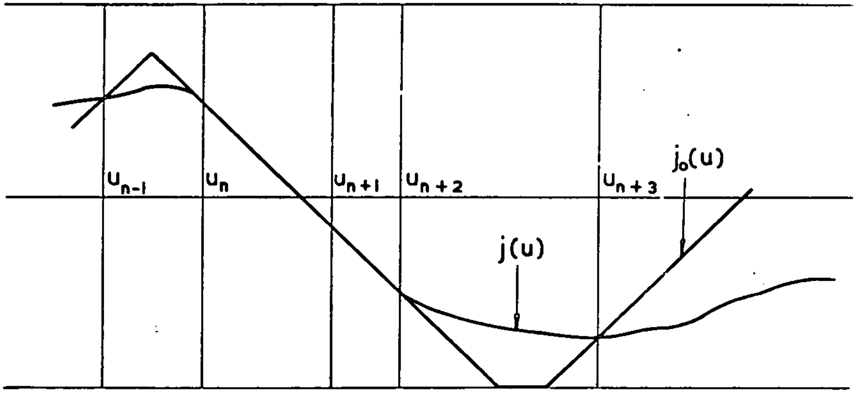


Figure 2. Occurrence of a zig of vanishing length

(b) If $u_n \rightarrow u_{n+1}$ is an interval in which $W(u) < 0$ proceed a similar way but reverse the signs of the slopes of the two nes, and join along $j = -i_m$ if they intersect for some $j < -i_m$. ne complete $j_0(u)$ obtained in this way is a sequence of straight ne segments as shown in *Figure* 1, which illustrates the conuction. Since it will be necessary to refer fairly frequently the salient features of functions of this type, it will be convenient at this stage to give them short descriptive names. Segments of slopes $\pm i_m'$ will be called 'zigs' and 'zags' respectively, while horizontal segments at $\pm i_m$ will be called 'plateaux' and 'valleys'. A zig may end at the start of a plateau or at the start of a zag, the sharp point formed in the second case being known as a 'peak', while the corresponding feature when a zag ends at the beginning of a zig will be called a 'ditch'. A sequence of zigs and zags without any intervening plateaux or valleys will be called a 'zigzag'.

Clearly each $j_0(u)$ satisfies equations 2 and 3 and the set $\{j_0(u)\}$ is a subset of the set $\{j(u)\}$ of all permissible inputs; the same $j_0(u)$ is obtained from all $j(u)$'s which are equal at the points $u = u_n$. Further, any zigzag $j(u)$ obtained by drawing straight line segments of alternate slopes $+i_m'$ and $-i_m'$, and joining them by horizontal segments at $j = \pm i_m$ if necessary, is a $j_0(u)$ provided:

(a) $j'(u_n) = i_m'$ if $W(u_n + \varepsilon) > 0$ for sufficiently small $\varepsilon > 0$

(b) $j'(u_n) = -i_m'$ if $W(u_n + \varepsilon) < 0$ for sufficiently small $\varepsilon > 0$.

(c) Between successive points u_n and u_{n+1} there is only one segment with slope $+i_m'$ and one with slope $-i_m'$. (In some cases the segment of slope $\pm i_m'$ associated with some u_n may appear to be missing; however, for formal reasons we regard it as present but of zero length. *Figure* 2 shows how this may occur in constructing the $j_0(u)$ corresponding to a $j(u)$ which maintains a constant slope $-i_m'$ over two adjacent intervals $u_n \rightarrow u_{n+1}$ and $u_{n+1} \rightarrow u_{n+2}$.)

A particular $j_0(u)$ is uniquely determined by (a), (b), (c) and ie positions of its sloping segments, specified for instance by ving their intercepts (produced if necessary) with any line.

If $j_0(u)$ is constructed from $j(u)$ as described it is seen that

$$j_0(u) \geqslant j(u) \quad \text{whenever} \quad W(u) > 0,$$
$$j_0(u) \leqslant j(u) \quad \text{whenever} \quad W(u) < 0,$$

o from equation 5 we have the result:

heorem 1

If y is the output (at t = 0) given by any input j(u) satisfying quations 2 *and* 3, *and* y_0 *the output given by the corresponding* (u), *then* $y_0 \geqslant y$.

In seeking the greatest y we may therefore restrict attention o the subset $\{y_0\}$ generated by the inputs $\{j_0(u)\}$. It has already een noted that, for stability, $\int_0^\infty |W(u)|\,du$ exists, so there xists some U such that $\int_U^\infty |W(u)|\,du < \varepsilon$, with arbitrarily nall ε. The error produced in y by truncating the integral (5) t $u = U$ will certainly not exceed εi_m for any $j(u)$ satisfying quation 2; thus, the integral may be truncated after a finite nterval, with arbitrarily small loss in accuracy. For conenience, we shall suppose from now on that this has been one, and that $W(u)$ has only finitely many zeros in the range $< u < U$, namely $u_1 \ldots u_N$.

A $j_0(u)$ is uniquely specified by the positions of all its zigs nd zags; it is convenient to define the position of the nth ig or zag (that is, the one corresponding to the zero u_n of W; s we remarked after (c), this may be of zero length) by giving s intercept x_n with the line $i = 0$, measured from u_n as origin. ince u_n must lie on this sloping segment,

$$-\phi < x_n < \phi \qquad \text{for} \qquad n = 1, 2 \ldots, N \tag{6}$$

'here $\phi = i_m/i_m'$. Since the sloping segments corresponding o u_n and u_{n+1} must meet between u_n and u_{n+1},

$$u_n < \tfrac{1}{2}(u_n + u_{n+1} + x_n + x_{n+1}) \leqslant u_{n+1} \tag{7}$$

Conversely, to any $(x_1, \ldots, x_N)$ satisfying equations 6 and 7 iere corresponds a $j_0(u)$. Accordingly, $j_0(u)$ may be represented y a point $(x_1 \ldots, x_N)$ of an N-dimensional cube of side 2ϕ, nd indeed points corresponding to j_0's fill up a convex subset f this cube (that is, if $(x_1 \ldots, x_N)$ and $(y_1 \ldots, y_N)$ both give 's, then all the points of the segment joining them give j_0's). 'he problem is therefore reduced to that of finding the greatest alue of a function of N variables whose variation is restricted y equation 6 and 7.

It is certainly possible to find an extremal $j_0(u)$, giving a tationary value of y in our truncated problem, and indeed it nay be proved that an extremal exists even if the integral (5) s left untruncated. The conditions for an extremal $j_0(u)$ are ormally very simple, as will now be seen.

'heorem 2

$j_0(u)$ *is extremal if and only if* $\int_a^b W(u)\,du = 0$ *whenever a and are the values of u at the beginning and end of any zig or zag.*

Proof—Figure 3(*a*) shows an adjacent zig and zag of some (u) meeting in a peak. The end points of the sloping segments re a, b and c as indicated. Consider a small variation to ive $\tilde{j}_0(u)$, in which the zig is displaced by δx to the right, and he zag by δy, but $j_0(u)$ and $\tilde{j}_0(u)$ are otherwise identical. $\tilde{j}_0(u)$ is indicated in *Figure* 3(*a*) by broken lines, and the difference $\tilde{j}_0(u) - j_0(u)$ is plotted in *Figure* 3(*b*). This is split into two components in *Figures* 3(*c*) and 3(*d*); it is obvious that the sum of these gives $\tilde{j}_0(u) - j_0(u)$. When $\int_0^\infty W(u)[\tilde{j}_0(u) - j_0(u)]\,du$ is formed, the component in *Figure* 3(*c*) gives an expression of first order in δx, δy:

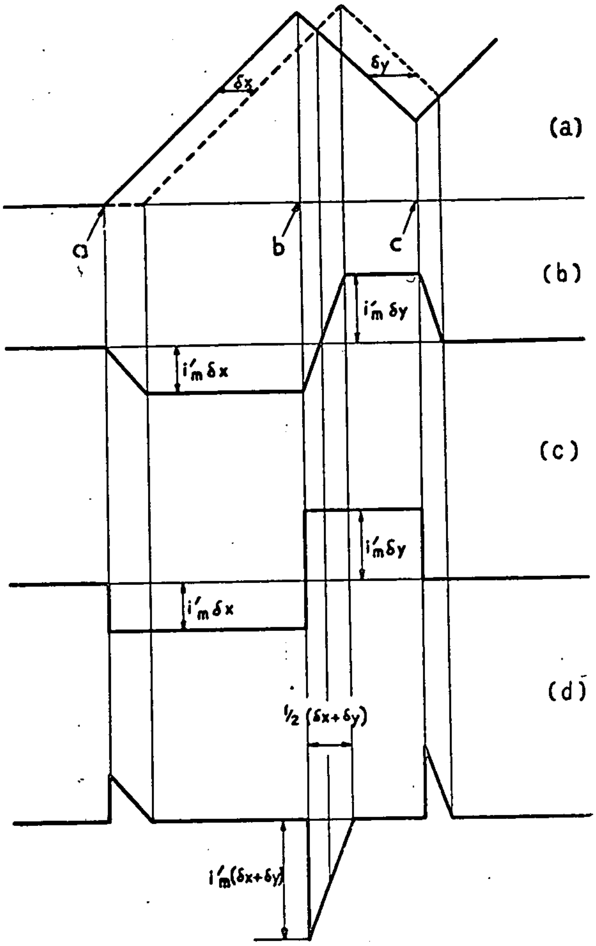


Figure 3. *Variation of* $j_0(u)$

$$-i_m'\,\delta x \int_a^b W(u)\,du + i_m'\,\delta y \int_b^c W(u)\,du \tag{8}$$

while the component in *Figure* 3(*d*) gives higher order contributions. Necessary and sufficient condition for an extremal $j_0(u)$ is that equation 8 should vanish for arbitrary δx and δy, for all zigs and zags, proving the theorem.

By going on to consider the second-order variation in y_0 corresponding to the component of $\tilde{j}_0(u) - j_0(u)$ plotted in *Figure* 3(*d*), it is possible[1] to show that the stationary value of y_0 corresponding to an extremal $j_0(u)$ is a *maximum* (rather than a minimum), and further that y_0 cannot take any value greater than this stationary value. Thus the stationary value is the required value of y_m. It is interesting to note, however, that it is false to assume that there is a unique extremal $j_0(u)$, but when there is more than one, the values of y_0 corresponding to the different extremal $j_0(u)$'s are all equal, and are greater than any other attainable value of y_0.

Procedure for Constructing the Extremal $j_0(u)$

Having formally obtained the conditions for a stationary value in Theorem 2 and ensured that this gives the greatest attainable value of y, a procedure for constructing the extremal input will be briefly sketched. Although this is essentially rather simple, it is very tedious to describe fully; a more complete discussion has been given elsewhere[1].

First note that the step response is the integral of the weighting function, so that Theorem 2 may be restated: a necessary and sufficient condition for an extremal $j_0(u)$ is that the values of the step response at the beginning and end of each zig or zag should be equal. Thus, if $S(u)$ is the step response, the successive peaks and ditches of any unbroken zigzag must occur at values of u corresponding to the intersections of some horizontal straight line S = constant with the graph of $S(u)$. An arbitrary line will, in fact, determine a zigzag with a number of peaks and ditches depending on the number of its intersections with the graph of $S(u)$, as shown in *Figure* 4. If the zigzag constructed in this way lies within the limits $\pm i_m$ for all values of u up to $u = U$, at which $S(u)$ may be truncated, it is an extremal $j_0(u)$ up to the truncation point. In general, however, this will not be the case for two reasons. Firstly, the zigzag may pass outside the limits $\pm i_m$, as in the interval between the points P and R in the example drawn in *Figure* 4, and secondly it may terminate at a point R, lying to the left of $u = U$. In the first case the procedure is to vary the position of the line S = constant until Q is lowered on to the line at $+i_m$, thus giving a complete zigzag satisfying the stationary condition and terminating at Q, while in the second case the line S = constant is again varied until the terminal point of the zigzag lies at $\pm i_m$ as appropriate. Having completed one zigzag in this way, a second must be started to the right of its terminal point, and so on until the last zigzag constructed extends beyond the point $u = U$. The set of zigzags so obtained is then the required extremal $j_0(u)$ for $u < U$, and therefore provides a solution of the problem to the accuracy required; the corresponding value of y_m can be obtained by using this $j_0(u)$ in equation 5. An attempt to carry out the procedure just outlined will quickly reveal that the process of modifying a zigzag as described may give rise to other difficulties due to peaks or ditches in other parts of the zigzag moving outside the range $\pm i_m$. This is discussed in the more complete treatment given in reference 1.

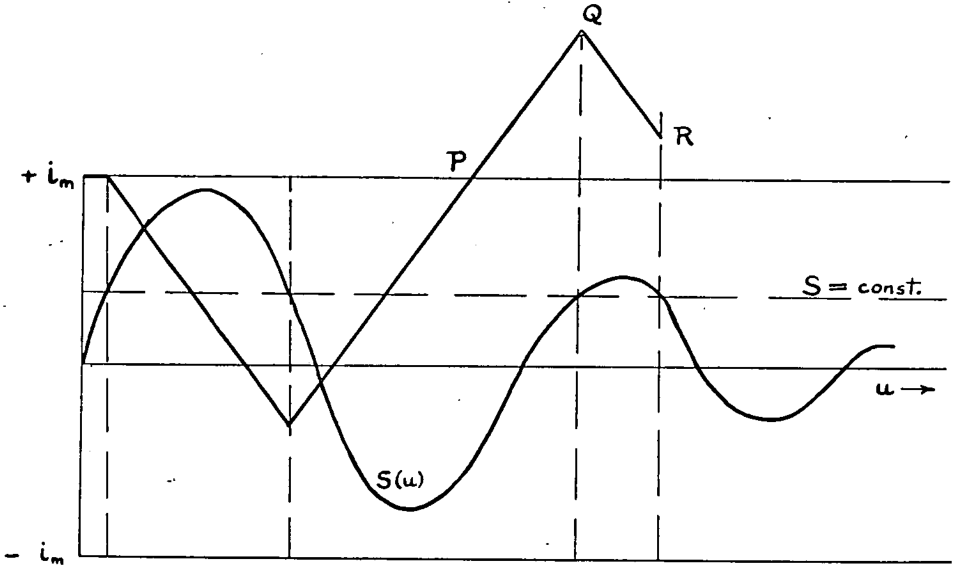


Figure 4. *Construction of an extremal* $j_0(u)$

Although the procedure briefly outlined here is formally a complete method of generating an extremal $j_0(u)$, and could be carried out graphically or programmed for a digital computer, the problem is greatly simplified in the particular case of systems described by second-order differential equations. For such a system the zeros of $W(u)$ are either finite in number or spaced at equal intervals along the u axis. This makes it possible to divide all possible cases which may occur into a fairly small number of classes, for each of which an explicit expression for the integral appearing in equation 5 may be written down. A digital computer programme can then be written which simply allocates each particular case it meets to the appropriate class and calculates y_m from the explicit expression corresponding to this class.

A programme of this type has been written for an Elliott 402 computer, which accepts specifications of the step response $S(u)$ in the form

$$S(u) = A + Be^{-pu} + Ce^{-qu}$$

for overdamped systems

$$= A + B \sin \pi v + Be^{-pu} \sin (qu - \pi v)$$

for oscillatory systems

together with a specification of $\phi(=i_m/i_m')$, and prints out the corresponding value of y_m/i_m. This programme was used in the investigation of level control systems described in the following sections and has subsequently been used in a corresponding investigation of pressure control systems.

No programme has yet been written for the general case of systems not restricted to those satisfying second-order differential equations.

The Design Problems for Level Control Systems

In this section the problems to be solved in designing simple level control systems will be described. In order to establish clearly the meaning of various symbols, *Figure* 5 shows simple level control system. An unregulated flow of liquid f_1 enters a vessel in which the level is maintained at a desired value l_d by regulating the exit flow f_2. Alternatively, the input flow may be regulated and the output flow subject to demand fluctuations, but in either case call the regulated flow f_2 and the flow subject to disturbances f_1. In order to avoid introducing the cross-sectional area of the vessel, the performance is described in terms of deviations in *volume* in the vessel from the desired volume, so that if l is the deviation in level from its desired value, and A the cross-sectional area:

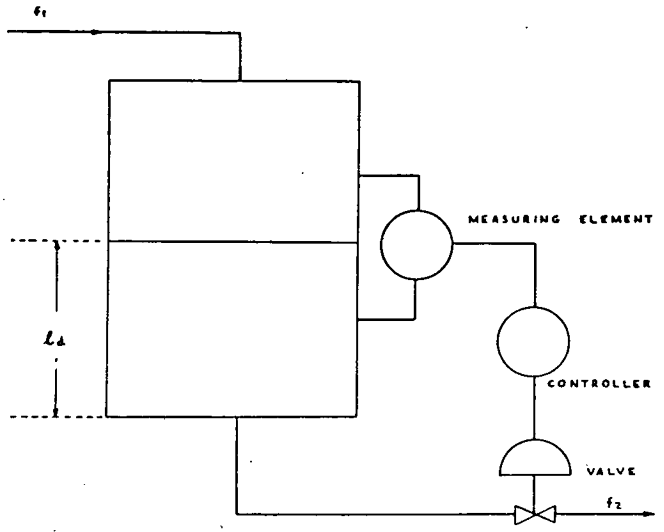


Figure 5. Simple level control system

$$V = lA$$

where V is the deviation in volume.

Level control systems fall naturally into three main classes and the design problem is different for each case. These will be described in turn.

(a) Flow smoothing system

In many systems the object is not to maintain strict control of the liquid level, but rather to make use of the available free space in the vessel to smooth out fluctuations in the flow f_1 so that they are transmitted to f_2 only in an attenuated form. The aim is to vary f_2 as slowly as possible, while still checking the variations in level before they cause the vessel to overflow or to empty. A good example of this is the level control on the liquor in the base of a distillation column, where the object is to vary the flow from the column base as slowly as possible to avoid disturbing sections of the plant further downstream, while at the same time preventing the level from rising too high and causing contamination of the liquor on the first plate by carryover, or from falling too low and exposing part of the heater.

In this case the object of the design procedure is to predict:

(1) The minimum size of vessel which can be used if the desired smoothness of the controlled flow is to be attained without any possibility of the vessel overflowing or emptying.

(2) Size and speed requirements for the control valve regulating the flow f_2.

These quantities obviously depend on the violence of the disturbances in f_1 among other things, so in order to get quantitative results it is necessary to give some method of specifying both the disturbances and the desired 'smoothness' of the controlled flow. The definitions which will be given here are one of many possibilities and are chosen to suit our particular design method. In the author's experience these use about the greatest amount of information it is normally possible to obtain at the design stage. The desired smoothness of f_2 is specified by giving a limit f_m' for its tolerable rate of change, and the violence of the disturbances is similarly specified by giving limits $f \pm F$ between which the flow f_1 must lie. It may also be possible to specify a limit F' for the rate of change of the flow f_1, which imposes a further restriction on the violence of the disturbances, but it is considered unlikely at present that any information about f_1, other than values of F and (possibly) F', will be available at the design stage. By the method described in the previous two sections it is then possible to calculate V_m, the maximum deviation in volume of liquid in the controlled vessel from its desired value, and the vessel must be sufficiently large to accommodate fluctuations of this size. The maximum variation of f_2, $\pm f_m$ about its average value, can also be calculated and this may be used in conjunction with the value of f_m' (which was the original design specification) to obtain the maximum speed of movement which will be demanded from the control valve motor. The design is then complete.

(b) Level control systems

The second, and probably smaller class of systems can more correctly be described as level control systems, since in these the object is to control the level of liquor within specified limits of the desired value. The vessel is assumed to be given and the task of the designer is merely to specify a control valve of adequate size and speed. Using the notation already discussed, the desired performance is now specified by giving limits $\pm V_m$ about the desired value, within which variations of the volume of liquid in the vessel are required to remain. (This is equivalent to specifying limits $\pm l_m$ for the level, where $l_m = V_m/A$.) Using F, and also F' if available, it is then possible to calculate the maximum correction required and the maximum speed demanded from the control valve motor, so completing the design.

(c) Given valve motor

A third case often arises, where the only requirement is that the vessel shall not fill or empty completely, but a restriction is imposed on the rate of change of output by the speed of available valve motors rather than any smoothness condition. An example would be a catchpot handling large flows. Provided that the level does not rise so high that there is appreciable carryover of liquid into the gas stream, or fall so low that gas can pass into the liquid stream, its position within the pot is not important. However, a large control valve may require a pneumatic motor with large diaphragm area to overcome off-balance forces at the plug, so it will be fairly slow in action. The limitation which this imposes on the correction rate determines a minimum size for the catchpot to ensure that it will neither completely empty nor completely fill. The size could, of course, be reduced by fitting a faster motor to the valve and it would be necessary to balance the savings resulting from a reduction in vessel size against the increased cost of a faster motor.

In practice, the majority of catchpots require a minimum size for adequate disentrainment of liquid and vapour which gives ample time even for a slow motor to apply the necessary rrection, and this type of system only presents a genuine oblem of level control in exceptional cases.

ıalysis of the Simplified Linear Model

A simple model of the system, which neglects all departures m ideal behaviour in the measurement and control equip- nt, will be analysed in this section. The theoretical and perimental justification for these approximations will be cussed in a later section.

The system to be analysed is shown in *Figure* 6 from which

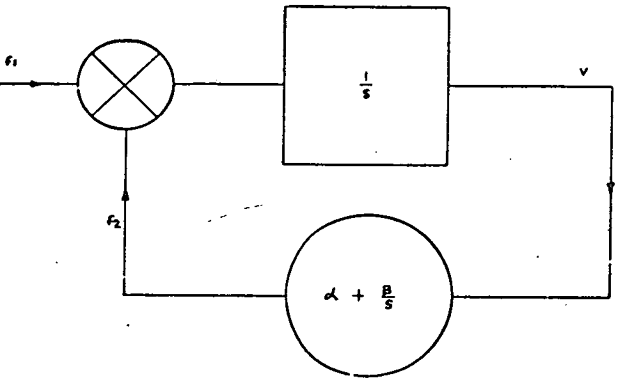


Figure 6. Block diagram of simplified model

s seen that the measuring system, controller and control ve are assumed to be described by a simple transfer function the form $\alpha + \beta/s$, corresponding to a perfect proportional s integral controller. α is then the loop gain and β/α is $_I$, where τ_I is the integral action time. Derivative action is considered, and although it is difficult to justify this orously without a complete investigation of cases in which s included, the following arguments are felt to provide ficiently good grounds for neglecting it.

Consider a system in which the object is to control the level a given vessel as closely as possible. With a transfer ction of the above form, which neglects inherent lags in pneumatic equipment, there is clearly no limit to the ntness of control attainable by increasing the gain, as the tem remains stable for all gains. Thus the addition of ivative action would not improve the attainable quality of el control, and would call for faster rates of correction. the case of a system intended to act as a flow smoothing rice, derivative action is completely unsuitable, since its rpose is to cause f_2 to respond rapidly to changes in f_1.

From *Figure* 6 it is easy to write down the transfer functions ıting V, f_2 and f_2' to f_1, and hence to deduce the corre- onding step responses. Since the system is represented by a ond-order differential equation, the step responses can be tten in a form suitable for use with the general purpose nputer described in an earlier section, and hence V_m, f_m l f_m' (defined above) can be computed. The results are ually expressed by giving values of V_m/F, f_m/F and f_m'/F as ctions of the three variables $\phi(=F/F')$, α and β. A full loration of the forms of three functions of three variables uld be very tedious, but this can be avoided by a suitable ice of time units, as will be shown.

n addition to the three functions just mentioned, it is venient to consider T_m, defined by $T_m = 2f_m/f_m'$, which is minimum time in which the control valve may be required traverse its complete range and hence determines the necessary speed of the valve motor. It is also convenient to choose ϕ, α and β/α as the independent variables rather than ϕ, α and β, since β/α is the reciprocal of the integral action time. Now consider the effect on all these quantities of multiplying the unit of measurement of time by a factor N. All the quantities depend dimensionally on time only (except f_m/F which is dimensionless) and Table 1 shows the factors by which they alter.

Table 1

Quantity	*Multiply by*	*Quantity*	*Multiply by*
V_m/F	$1/N$	ϕ	$1/N$
f_m/F	1	α	N
f_m'/F	N	$1/\tau_I$	N
T_m	$1/N$		

It is clear from this table that, by a suitable choice of the unit of time measurement, it is possible to give any pre-assigned numerical value to the design specifications, V_m/F for a level control system, f_m'/F for a flow smoothing system and T_m when the motor is given. Alternatively it is possible to give ϕ any desired numerical value. For the purpose of this general investigation the second alternative will be chosen, as it reduces the number of independent variables to two, and the quantities of interest can be exhibited graphically by plotting contour charts in the plane (α, $1/\tau_I$). The results shown in *Figures* 7–10 with $\phi = 2$ were obtained by automatically computing 'spot heights' at the points of a rectangular grid, and interpolating the contours.

One feature of the function V_m/F which does not show up in the contours of *Figure* 7 is a step discontinuity at $1/\tau_I = 0$. Thus for $\alpha = 0{\cdot}5$, $V_m/F \rightarrow 4{\cdot}0$ as $1/\tau_I \rightarrow 0$ through positive values, but $V_m/F = 2{\cdot}0$ when $1/\tau_I = 0$. The contours of equal V_m/F approach some value of α on the axis $1/\tau_I = 0$ but suddenly jump back to a point half-way between this value and the origin when they actually reach the axis. The physical reason for this behaviour is quite simple. For a very small amount of integral action, the *dynamical* behaviour of a proportional controller and of a proportional plus integral controller will be almost identical; in particular the maximum change in the controlled quantity following a step change of given magnitude in the input will be almost the same in both cases. Now consider the situation in which f_1 has been at its smallest permissible value $\bar{f}_1 - F$ for a very long time. If any integral action is present, however little, V will have balanced out at its desired value $\bar{V}$, but if there is strictly zero integral action it will settle at its smallest value $\bar{V} - V_m$. If f_1 then jumps suddenly to $\bar{f}_1 + F$, the maximum deviation of *V from its initial value* will be positive and approximately the same in both cases, but the maximum deviation from $V = \bar{V}$ will be twice as great for the system with integral action, since the starting value was $\bar{V}$ rather than $\bar{V} - V_m$. In the present case, where strict limits for the variation of f_1 are known, this feature gives a great advantage to the proportional-only controller.

It is not easy to see the behaviour of the functions in the neighbourhood of the origin from *Figures* 7–10, so enlargements of small regions near the origins have been computed. It is seen from Table 1 that an increase in the unit of time measurement increases the numerical values of α and $1/\tau_I$ and decreases

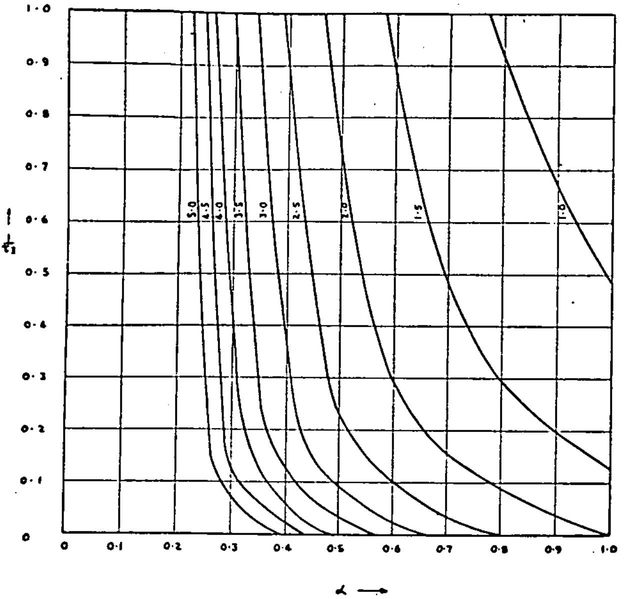


Figure 7. V_m/F *for* $\phi = 2{\cdot}0$

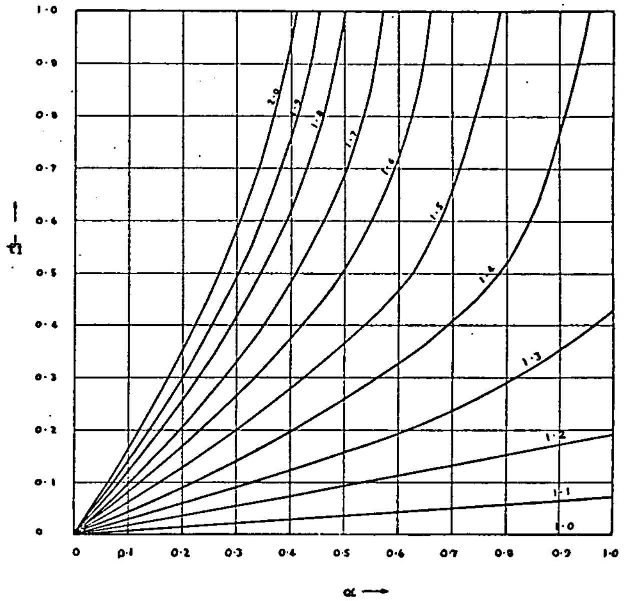


Figure 8. f_m/F *for* $\phi = 2{\cdot}0$

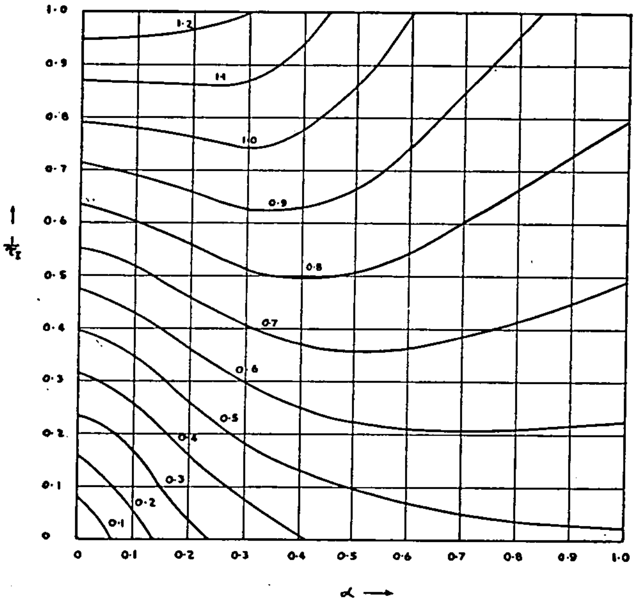


Figure 9. f_m'/F *for* $\phi = 2{\cdot}0$

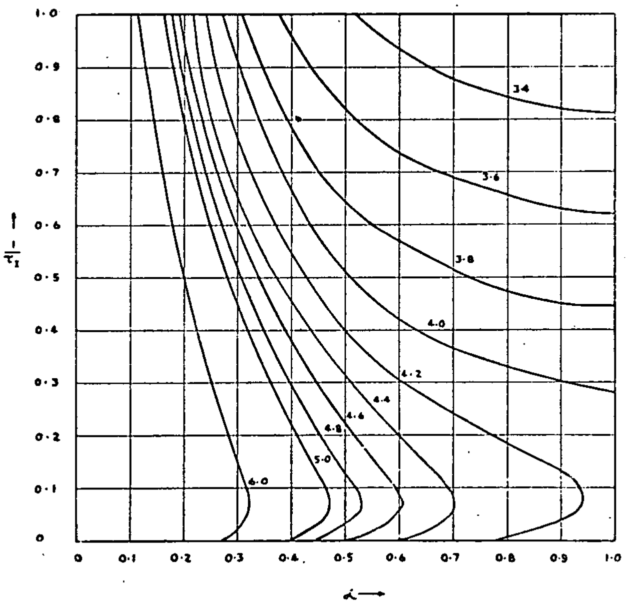


Figure 10. T_m *for* $\phi = 2{\cdot}0$

the value of ϕ in the same ratio, so the enlargements of small squares near the origin may be regarded, with equal validity, as charts drawn over the original ranges of α and $1/\tau_I$, but corresponding to a smaller value of ϕ. *Figures* 11–14 are plotted from this second point of view with $\phi = 0{\cdot}05$, but if ϕ is increased to 2·0 the values of α and $1/\tau_I$ are everywhere multiplied by 0·025, and the result is an enlargement of a small square near the origins of the original charts.

Having established the form of the functions it is now possible to see what type of system is best suited to each of the three problems discussed in the previous section.

(*a*) *Flow smoothing system*

In this case f_m'/F is specified and can be expressed in units chosen so that $\phi = 2{\cdot}0$, when it determines a contour on *Figure* 9 (it may be more convenient to make $\phi = 0{\cdot}05$ and use *Figure* 13 if f_m'/F is small). It is required to follow the variation of V_m/F, f_m/F and T_m as the representative point moves along this contour, and this can easily be done by superimposing the chart for f_m'/F on each of the other charts in turn.

For small values of f_m'/F when *Figures* 11–14 are appropriate,

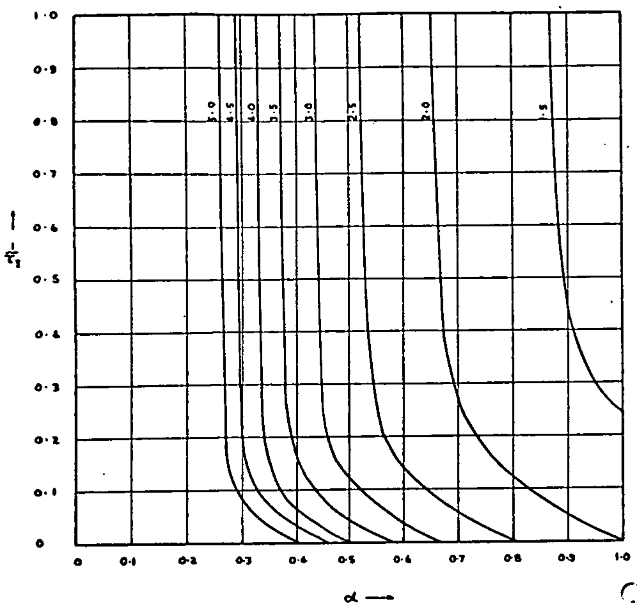


Figure 11. V_m/F *for* $\phi = 0{\cdot}05$

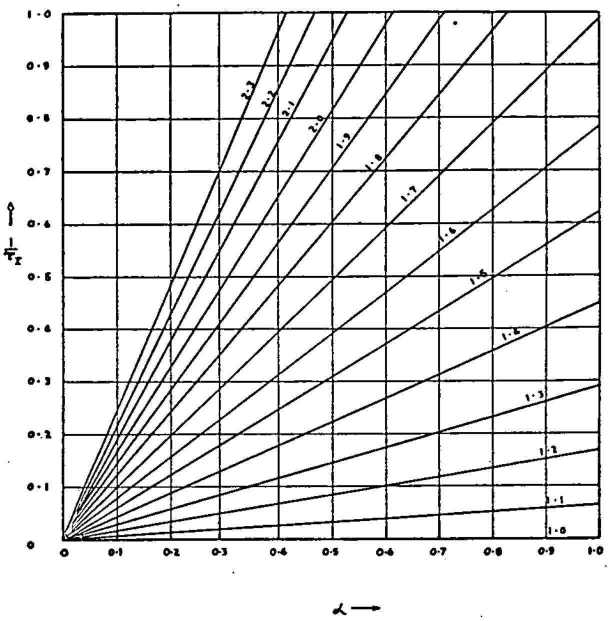


Figure 12. f_m/F *for* $\phi = 0{\cdot}05$

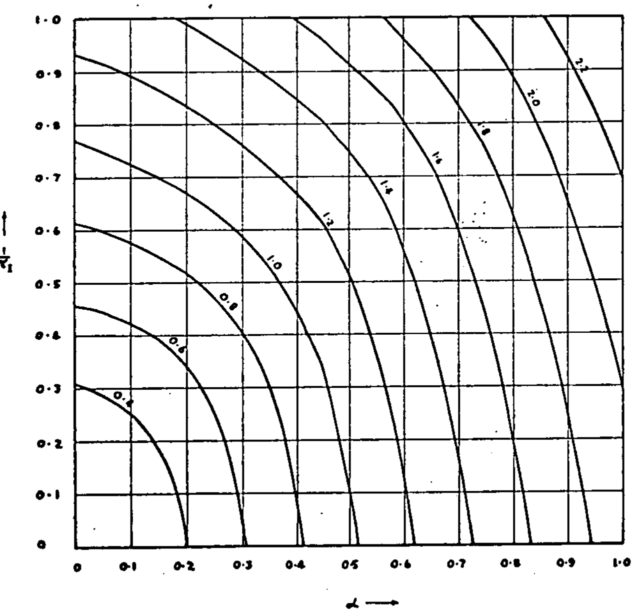


Figure 13. f_m'/F *for* $\phi = 0{\cdot}05$

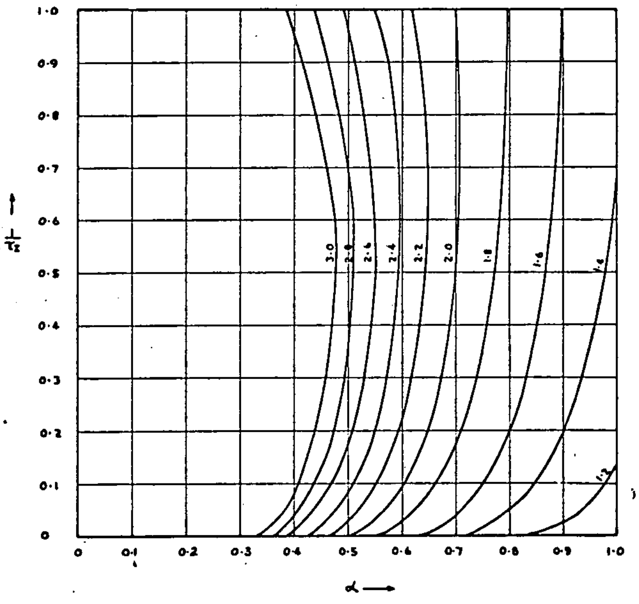


Figure 14. T_m *for* $\phi = 0{\cdot}05$

t is seen that, for finite integral action, V_m/F first decreases hen increases as $1/\tau_I$ is increased. However, even at its ninimum it never becomes as small as the value taken when $/\tau_I = 0$, so from the point of view of vessel size a proportional ontroller is best. This conclusion is unchanged for larger alues of ϕ, as can be seen from *Figures* 7 and 9. It should be emembered in considering *Figure* 9 that the contours $f_m'/F >$ $\cdot 5$ are of no interest for flow smoothing systems, since $\phi = F/F'$, o $f_m'/F > 0{\cdot}5$ implies $f_m'/F > F'/F$, i.e. the correcting flow s less smooth than the disturbing flow. References to *Figure* 9 in this section are therefore concerned only with the contours $f_m'/F < 0{\cdot}5$; it is clear in fact that the above description of the behaviour of V_m/F on moving along a contour $f_m'/F =$ constant is not valid when $f_m'/F > 0{\cdot}5$. Comparison of the f_m'/F charts with the f_m/F charts shows that f_m/F increases monotonically on moving along a contour $f_m'/F =$ constant in the direction of increasing $1/\tau_I$, so that more correction is required when integral action is used. Since f_m' is specified, this implies that slower valve motors may be used, but flow smoothing applications are not likely to be very demanding n motor speed in any case. Thus there would appear to be very advantage in using simple proportional control in flow moothing applications, and it is possible to plot a simple lesign curve to give the necessary vessel size for any specified moothing, as shown in a later section.

b) Level control system

Here V_m/F is specified and determines a contour on *Figure* 7 or *Figure* 11. The variation of T_m and f_m/F on moving along his contour can be found, as before, by superimposing the ppropriate pair of charts. For finite values of $1/\tau_I$, T_m first ncreases then decreases when the V_m/F contour is described n the direction of increasing $1/\tau_I$, but there is a discontinuous ump downwards when $1/\tau_I$ begins to increase from zero. Vhether or not the subsequent rise in T_m is greater than this nitial fall depends on the relative values of ϕ and V_m/F. Vhen $V_m/F = \phi = 2{\cdot}0$, for instance, it is seen from *Figures* 7 nd 10 that the rise is not large enough to recover the initial liscontinuous fall, but when $V_m/F = 2{\cdot}0$ and $\phi = 0{\cdot}05$, *Figures* 1 and 14 show that T_m rises to about 2·15 when $1/\tau_I \approx 0{\cdot}7$, ompared with the value 2·0 at $1/\tau_I = 0$.

In this case there is some advantage to be gained, from the oint of view of valve motor speed, by using a finite amount f integral action. However, the difference in speed is less han 10 per cent, which is about the order of accuracy to be xpected when applying the predictions to a practical system, s will be seen from the experimental work. This improvement s not sufficient to justify the extra complication introduced nto the design procedure, so by restricting attention to roportional control it is again possible to obtain a simple esign curve giving T_m.

) Given valve motor

The case in which T_m is given is very similar to that just iscussed. The specified value of T_m determines a contour on *igure* 10 or *Figure* 14 and the variations of V_m/F and f_m/F n passing along this contour can again be found by supernposing charts. When $1/\tau_I$ increases from zero there is an nitial discontinuous rise in V_m/F, and the subsequent fall and se may not lead to a minimum smaller than the initial value, epending on the relative values of ϕ and T_m. However, the est V_m is not more than 10 per cent better than the value btainable without integral action, so again it is hardly worthhile designing for anything more complicated than a simple roportional controller.

In all three cases it has been decided to design on the basis f a proportional controller, as it has been shown that the introuction of integral action does not permit any significant conomies in the design. It should be noted that this does not ecessarily mean that integral action is never useful in a level ontrol application. It has been assumed that only a very mited amount of information about the disturbance is vailable, and that it is necessary to allow for the worst possible ontingency consistent with this information when designing he system. It may turn out when the plant is in operation hat the disturbance has special features which make integral ction very useful (such as a high-frequency ripple supernposed on a much slower variation of large amplitude), but is unlikely that this information would be available at the esign stage.

esign Curves for Systems with Proportional Control

When $\beta = 0$, corresponding to no integral action, it is seen from *Figure* 6 that the transfer functions relating V, f_2 and f_2' to f_1 are:

$$\frac{V(s)}{f_1(s)} = \frac{1}{s+\alpha}; \qquad \frac{f_2(s)}{f_1(s)} = \frac{\alpha}{s+\alpha}; \qquad \frac{f_2'(s)}{f_1(s)} = \frac{s\alpha}{s+\alpha} \tag{9}$$

From which follow the weighting functions and step responses:

$$Wv = \mathrm{e}^{-\alpha t}, \qquad Sv = \frac{1}{\alpha}(1 - \mathrm{e}^{-\alpha t}) \tag{10}$$

$$W_{f_2} = \alpha\,\mathrm{e}^{-\alpha t}, \qquad S_{f_2} = (1 - \mathrm{e}^{-\alpha t}) \tag{11}$$

$$W_{f_2}' = \alpha[\delta(t) - \alpha\mathrm{e}^{-\alpha t}], \qquad S_{f_2}' = \alpha\mathrm{e}^{-\alpha t} \tag{12}$$

where W_v, etc, are weighting functions, S_v, etc. are step responses, and $\delta(t)$ is the δ-function.

The particular $f_1(t)$ in $t < 0$ which makes $V(0)$, $f_2(0)$ or $f_2'(0)$ as large as possible can be found by the method described in earlier sections. For V and f_2, where equations 10 and 11 show that the step responses are monotone increasing, the worst $f_1(t)$ is:

$$f_1 = +F \qquad \text{for all } t < 0 \tag{13}$$

and the corresponding greatest values, V_m and f_m, are given by:

$$V_m/F = 1/\alpha, \qquad f_m/F = 1 \tag{14}$$

In the case of f_2' the worst $f_1(t)$ is:

$$\begin{aligned} f_1(-u) &= F - F'u \qquad \text{for} \qquad 0 < u < 2\phi \\ &= -F \qquad\quad \text{for} \qquad 2\phi < u \end{aligned} \tag{15}$$

where we have written $u = -t$. The corresponding greatest value f_m' follows as:

$$f_m' = \int_{0-}^{\infty} W_{f_2}'(u) f_1(-u)\,\mathrm{d}u = \int_{0-}^{2\phi} \alpha[\delta(u) - \alpha\mathrm{e}^{-\alpha u}](F - F'u)\,\mathrm{d}u - \int_{2\phi}^{\infty} \alpha[\delta(u) - \alpha\mathrm{e}^{-\alpha u}]F\,\mathrm{d}u$$

or

$$f_m'/F = \frac{1}{\phi}(1 - \mathrm{e}^{-2\alpha\phi}) \tag{16}$$

Finally T_m is defined as $2f_m/f_m'$, so from equations 14 and 16:

$$T_m = \frac{2\phi}{1 - \mathrm{e}^{-2\alpha\phi}} \tag{17}$$

Equations 14, 16 and 17 contain all the required results for a proportional controller.

Consider now the three basic design problems in turn.

(a) Flow smoothing system

The performance specification is f_m'/F, with dimensions 1/time, so it is possible to choose the unit of time measurement so that $f_m'/F = 1$. Using these units equation 16 gives:

$$\frac{1}{\phi}(1 - \mathrm{e}^{-2\alpha\phi}) = 1$$

so that

$$\alpha = \frac{1}{2\phi}\log_e\left(\frac{1}{1-\phi}\right) \tag{18}$$

Substituting this into equation 14 gives:

$$V_m/F = \frac{2\phi}{\log_e\left(\dfrac{1}{1-\phi}\right)} \tag{19}$$

which determines the required vessel size. V_m/F is obtained in the time units in which $f_m'/F = 1$, of course, and it is necessary to convert back to conventional units. The function defined by equation 19 may be plotted and is shown in *Figure* 15,

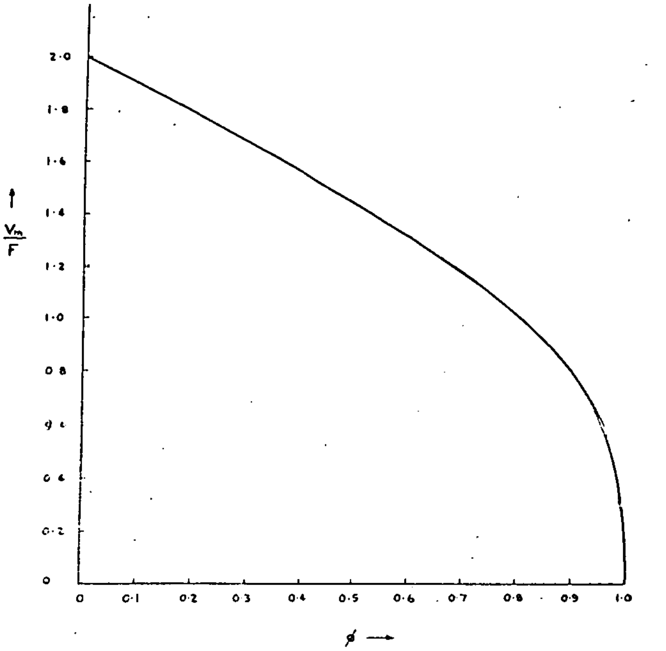


Figure 15. *Design curve for flow smoothing systems*

which can be used directly for design work. Note that if $\phi > 1$, $V_m = 0$. This is what would be expected, since in this case $F' < f_m'$ and no smoothing of the disturbance is required.

(*b*) *Level control system*

The performance specification is V_m/F, which has dimensions (time), so it is possible to choose the unit of time measurement so that $V_m/F = 1$. From equation 14 it follows that $\alpha = 1$ in these units, and substituting this into equation 17:

$$T_m = \frac{2\phi}{1 - e^{-2\phi}} \tag{20}$$

which determines the required speed of the valve motor (in the time units defined above), and can again be plotted to give a design curve as shown in *Figure* 16.

(*c*) *Given valve motor*

In this case T_m is fixed by the available valve motor. If the time unit is chosen so that $T_m = 2$, equation 17 gives:

$$1 = \frac{\phi}{1 - e^{-2\alpha\phi}}$$

which is identical with equation 18, and therefore determines the same value of V_m/F. The curve given in *Figure* 15 can therefore be used to determine V_m in this problem with the above choice of time unit.

In all three cases the value of ϕ to be used depends on the given values of F and F' of course, since $\phi = F/F'$, and if no information is available about the rate of change of the disturbance, it must be assumed that $F' = \infty$ or $\phi = 0$.

Approximations Involved in the Simplified Model

In setting up the simplified model on which the above development was based, all departures from ideal behaviour in the measurement and control system were neglected. With displacement type level measuring instruments the measurement lag is very short, so provided transmission distances are not too long the major lag will almost certainly be associated with the control valve motor driven by the controller relay. It should be remembered, however, that this time constant is effectively included in the internal feedback loop of the controller when all connections are short, so its effective value in the main control loop is equal to its actual value divided by $1 + K$, where K is the gain around the internal feedback loop of the controller. In a typical practical example the time constant of the valve in series with the controller output relay is about 15 sec, while the gain round the controller feedback loop is roughly 200, so the effective time constant for response to changes in the input to the controller is 15/200 sec, which is very small indeed.

The second important feature of the pneumatic system which has been neglected is the non-linearity of the flapper-nozzle and relay gain characteristics. The effects of saturation in these two amplifiers has been treated in detail[2] for some purely pneumatic systems, but the same type of phase-plane analysis can be applied to the present system with a proportional controller. In this way it can be shown that, when the valve motor speed is determined by the method described in the previous section, the system should always remain within its region of linear operation, which is what would be expected since the motor is chosen to be capable of the maximum speed of movement called for by an ideal controller. The phase-plane analysis can only be carried out for a proportional controller as a three-dimensional phase space would be needed to treat a system with a proportional plus integral controller.

As an experimental check of the validity of the simplifications introduced, measurements were made on a small level control installation. The vessel itself was a length of 6 in. i.d. pipe and the disturbances applied to the input flow were sufficiently large to cause the level to change an inch or two per sec if no

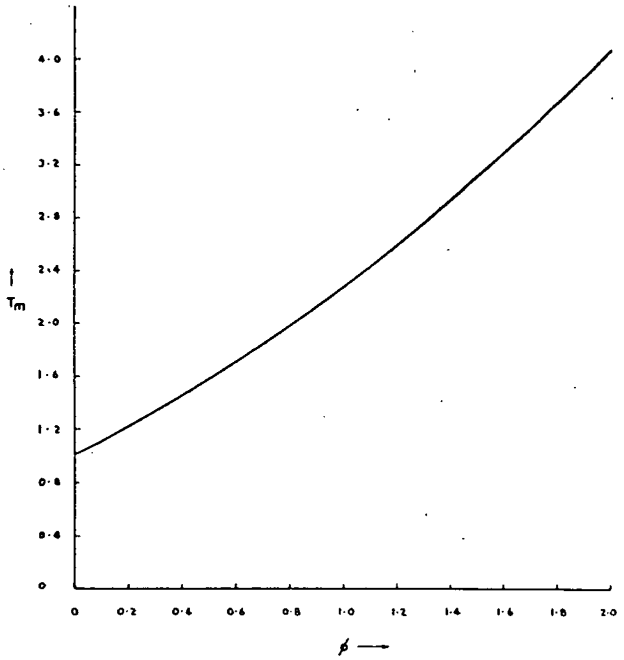


Figure 16. *Design curve for level control systems*

:orrecting action was taken. A differential pressure cell with ı range of 12 in. of water was used for measurement, so the ;peeds called for from the pneumatic system were fairly high. The changes occurring in the controlled level after step changes n the input flow were of the order of 1/2 in., so for recording ›urposes it was necessary to use a measuring system of greater ;ensitivity than the d.p. cell. For this purpose a pneumatic ;ignal was provided by a motion-balance type transmitter ›perated by a float, giving a sensitivity of 6 lb./in.^{2} per inch :hange of level. Step changes in input flow were produced by ›pening or closing a quick-acting Saunders valve in a by-pass :onnected in parallel with the inlet pipe. The size of the step :ould be altered by adjusting a second valve in this by-pass.

Measurements were made of the maximum changes, V^0 and f_2^0, in the volume of liquid in the vessel and the correcting flow espectively, following a single step change of magnitude F in he input flow. This was done for various values of the pro›ortional bandwidth and integral action time of the controller, ınd the results were compared with theoretical predictions from he simple model by plotting them as functions of $1/\tau_I$ for ›arious fixed values of the proportional bandwidth. The xperimental and theoretical values of V^0/F and f_2^0/F are :ompared in *Figures* 17 and 18 respectively. The agreement ›etween experiment and theory is very good indeed for V^0/F, vhile for f_2^0/F the greatest error occurs for the' largest values ›f $1/\tau_I$, and does not exceed about 5 per cent. It can be con:luded that the simplified model is adequate, even when the ›neumatic system is driven at an appreciable fraction of its naximum speed.

Having checked the validity of the linear model it is desirable ılso to check the predictions of V_m/F and f_m/F given in *Figures* 7, 8, 11 and 12. These follow without any mathematical ıpproximations from the linear model, and should therefore ›e correct if the simple model is a good one. When there is finite limit to the value of F', corresponding to a finite value ›f ϕ, it is not very easy to produce the worst disturbance

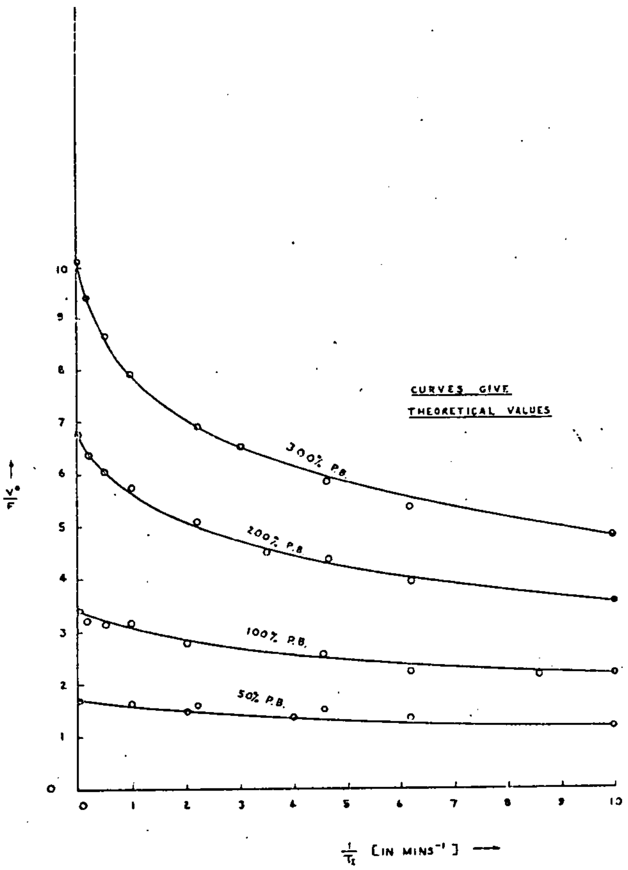


Figure 17. *Comparison of theoretical and experimental values of* V^0/F

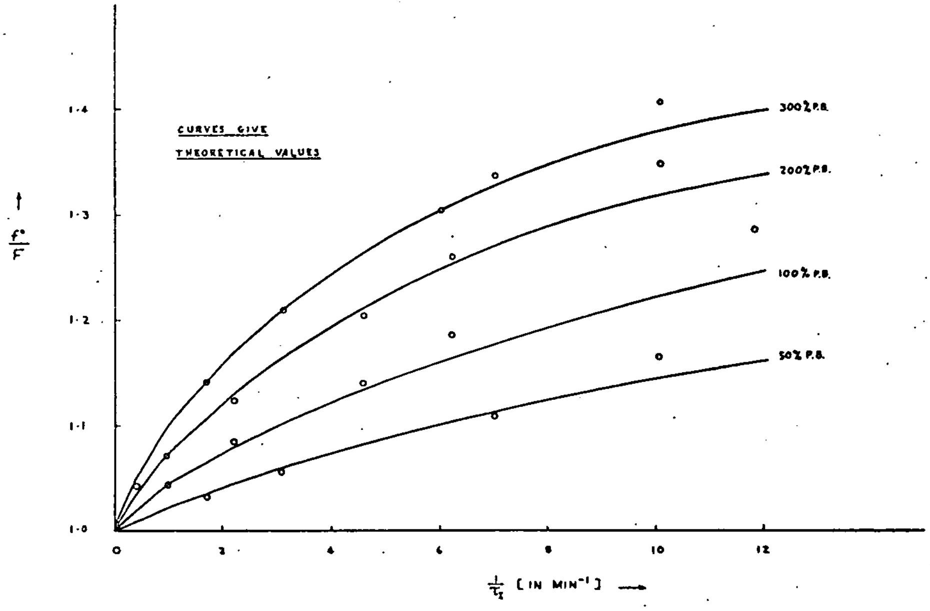


Figure 18. *Comparison of theoretical and experimental values of* f_2^0/F

experimentally, but when F' is unbounded the worst disturbance is a sequence of step changes between the limits $\pm F$, which can be produced quite easily with the arrangement already described. When the system response is overdamped the worst disturbance is a single step change of input flow, and V_m/F and f_m/F are equal to V^0/F and $f_2{}^0/F$, which have already been seen to be in good agreement with the theoretical values. For an oscillatory response, however, the worst disturbance is a sequence of equally spaced step changes, alternately positive and negative, at the resonant frequency of the system. (This is true only for a second-order system, of course, where the weighting function is periodic; for more complicated systems the worst input is not periodic.) In order to find this worst disturbance the system was excited with a square wave periodic variation in input flow, produced by switching the quick-acting Saunders valve at equal time intervals. The amplitudes of the resulting oscillations in V and f_2 were then plotted as functions of the frequency of the disturbance, and each curve exhibited a maximum which provided the required estimate of V_m or f_m. This was done for several values of the proportional bandwidth and integral action time, and the results are compared with the theoretical predictions in Table 2.

Table 2

α	β	V_m/F (obs)	V_m/F (pred)	α	β	f_m/F (obs)	f_m/F (pred)
0·3	0·12	5·4	4·3	0·3	0·16	2·0	2·1
0·6	0·24	2·8	2·2	0·3	0·072	1·7	1·6
0·3	0·031	4·6	4·7	0·15	0·082	2·8	2·7
0·6	0·046	2·4	2·7	0·6	0·33	2·0	1·7

The agreement of theoretical and experimental results is very good considering the rather crude experimental procedure.

It can be concluded from the experiments described that the design method described in the previous section should give predictions which can be relied on in practice to be in error by not more than about 10 per cent, provided the control valve motor is not pushed to the extreme limits of its speed of response.

Conclusions

Mathematical methods recently developed have been applied to the simplest type of automatic control system and lead to a design method with the two properties essential for day-to-day application:

(a) It is sufficiently simple for routine use by personnel with limited mathematical background and limited time available for design work.

(b) It demands only the simplest information about the disturbances affecting the system, and produces the most economical conventional system which is *safe*, in the sense that the desired performance will be attained even with the worst disturbance which can occur.

Although the original mathematical investigation was quite extensive, it was found that no system was very much better than a simple proportional controller. Thus the final design curves could be calculated very simply, though it would not have been possible to establish these results without the initial investigation, since it is by no means clear *a priori* that the addition of integral action does not give any significant advantage.

The same methods have also been applied to pressure control systems and lead to simple formulae for calculating the minimum sizes of pressure vessels compatible with given control specifications.

The author wishes to thank Mr. E. N. Martin for frequent useful discussions in the course of this work, and Mr. T. F. Farr, who undertook the experimental work on the pilot system. The work described is part of the research programme of Imperial Chemical Industries Limited, Billingham Division.

References

BIRCH B. J. and JACKSON, R. The behaviour of linear systems with inputs satisfying certain bounding conditions. *J. Electronics and Control* 6, No. 4 (1959) 366

JACKSON, R. A non-linear theory of the dynamical behaviour of pneumatic devices. *Trans. Soc. Instrum. Tech.* 10 (1958) 161

Summary

DL

Stable linear filters with bounded inputs give outputs which are also bounded. In the first three sections of the paper a mathematical method is outlined for obtaining the least upper bound of the output in the case where bounds are specified both for the magnitude of the input and its rate of change. The result has immediate applications in the design of regulators with disturbances which satisfy bounding conditions of this type, and in the remainder of the paper the method is illustrated by applying it to the simple case of a level control system with proportional plus integral control.

A5

THE INSTITUTION OF
ELECTRICAL ENGINEERS
FOUNDED 1871: INCORPORATED BY ROYAL CHARTER 1921
SAVOY PLACE, LONDON, W.C.2

OPTIMUM SAMPLED-DATA CONTROL

By

R. JACKSON, M.A.

MONOGRAPH No. 426 M
January 1961

To be republished in
PART C OF THE PROCEEDINGS OF THE INSTITUTION

The Institution is not, as a body, responsible for the opinions expressed by individual authors

621—52

The Institution of Electrical Engineers
Monograph No. 426 M
Jan. 1961
©

OPTIMUM SAMPLED-DATA CONTROL

By R. JACKSON, M.A.

(The paper was first received 23*rd July, and in revised form* 6*th October,* 1960. *It was published as an* INSTITUTION MONOGRAPH *in January,* 1961.)

SUMMARY

A method of rendering feedback control systems amenable to treatment by the Wiener theory is applied to the case in which the controller operates on a sampled measurement. An explicit expression is obtained for the minimum attainable mean-square error for certain classes of system transfer functions and disturbance power spectra, and the form of the optimum controller is derived. The results show the inherent limitations in controllability imposed by the structure of the controlled system and by the sampling process.

(1) INTRODUCTION

The Wiener optimum filter theory[2] cannot be applied directly to the problem of the optimum feedback regulator because of difficulties in imposing the condition of physical realizability on the control mechanism to be placed in the feedback loop. However, the closed-loop configuration can always be formally reduced to an equivalent open-loop configuration, and it was recently shown by Price[1] that this can always be done in such a way that the realizability condition takes a simple form in the open-loop case.

Price used this method to investigate the inherent limitations on the attainable control quality imposed by the structure of the controlled system when the controller is allowed to be any linear, continuous device, but a slight modification of the method allows it to be applied to the case in which the controller operates on samples of the measured variable taken at equal intervals of time. In this way it is possible to calculate the best control quality attainable with a linear sampled-data controller of given sampling interval and to derive the form of the optimum controller.

In this paper, the main interest is in the calculation of the best attainable control quality (measured by the mean-square error) and in comparing this with the best control quality obtainable with a continuous controller. This gives a direct measure of the reduction in controllability which must necessarily accompany the loss of information involved in the sampling process. The results have proved useful in estimating the frequency with which automatic batch-analytical instruments on a chemical plant must operate if their signals are to be useful for automatic control.

The type of system to be treated is shown in Fig. 3. A disturbance $d(t)$ causes the controlled quantity $e(t)$ to deviate from its desired value, which is taken as zero for convenience, and the controller operates on samples of $e(t)$ taken at intervals T. Given the fixed element P, the object is to find that physically realizable, linear operation C on the samples $e(rT)$ which will minimize $e^2(t)$, averaged in the manner discussed below over a statistical assembly of disturbances, and to calculate the minimum value of this quantity.

(2) STOCHASTIC SIGNALS IN SAMPLED SYSTEMS

The Wiener theory, as developed by solution of the Wiener–Hopf integral equation, leads to a solution of problems of the above type which minimizes the time average $\langle e^2(t)\rangle$. However, in practice, the ability of the system to reduce this time average for one particular disturbance is of less interest than its ability to keep $e^2(t)$ small, on the average, for all disturbances belonging to some statistically defined assembly. With the usual assumptions that the assembly in question, $\{d(t)\}$, is stationary and ergodic, it follows for continuous systems that the assembly average $\overline{e^2(t)}$ is independent of time t and is equal to the time average $\langle e^2(t)\rangle$ for any member function of the assembly (with the possible exception of a subset of measure zero).

These hypotheses of stationary and ergodic signals throughout the system are not tenable in sampled-data systems, since the result of sampling a stationary, ergodic signal is not stationary and ergodic; in fact, its statistical properties vary periodically with period equal to the sampling interval. There are two different but closely related methods of dealing with this situation. In the first, which is adopted by Ragazzini and Franklin,[3] the assembly considered is enlarged by considering an assembly of systems (as well as disturbances) which are physically identical but have their set of sampling instants displaced in a random manner relative to each other. The complete assembly of outputs, generated by all the signals of the assembly of disturbances applied to all these systems of identical structure, is stationary and ergodic if the assembly of disturbances was stationary and ergodic. Alternatively, it is not difficult to show that, if the assembly of functions $\{y(t)\}$ is generated from the stationary ergodic assembly $\{x(t)\}$ by any linear sampled-data filter, then

$$\langle y^2(t)\rangle = \frac{1}{T}\int_{-T/2}^{+T/2}\overline{y^2(t)}dt \quad . \quad . \quad . \quad . \quad . \quad (1)$$

Thus the Wiener theory, which minimizes $\langle y^2(t)\rangle$, will lead to a system which minimizes the assembly average $\overline{y^2(t)}$, further averaged with respect to time over a sampling interval, and it is in this sense that the control systems discussed here are optimum systems.

(3) REPRESENTATION OF LINEAR SAMPLED-DATA FILTERS

A linear sampled-data filter is a device which linearly relates an output function of time, $y(t)$, to values of an input function, $x(t)$, at the sampling instants $t = rT$. Thus

$$y(t) = \sum_{r=-\infty}^{+\infty} h(t - rT)x(rT) \quad . \quad . \quad . \quad . \quad (2)$$

where $h(u)$ is a function characterizing the particular filter considered. In a physically realizable system, $h(u) = 0$ for $u < 0$.

Filters of this type may be represented by shaded blocks [Fig. 1(*a*)] to distinguish them from continuous filters, which are normally represented by unshaded blocks. An alternative representation may be obtained by considering a continuous filter with weighting function $k(u) \equiv h(u)$ and input consisting of the following sequences of delta functions:

$$x^*(t) = \sum_{r=-\infty}^{+\infty} x(rT)\delta(t - rT) \quad . \quad . \quad . \quad . \quad (3)$$

Correspondence on Monographs is invited for consideration with a view to publication.

Mr. Jackson is with Imperial Chemical Industries, Ltd., Billingham Division.

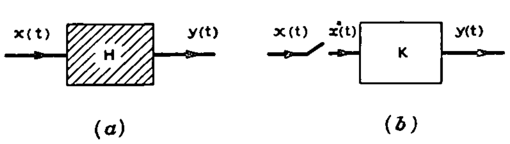


Fig. 1.—Sampled-data filters.

Then

$$y(t) = \int_0^{\infty} k(u)x^*(t-u)du = \sum_{r=-\infty}^{+\infty} x(rT)\int_0^{\infty} k(u)\delta(t-u-rT)du$$

or

$$y(t) = \sum_{r=-\infty}^{+\infty} k(t-rT)x(rT) \quad . \quad . \quad . \quad . \quad (4)$$

which is identical with eqn. (2) since $k(t - rT) = h(t - rT)$. This arrangement may be represented in a block diagram as shown in Fig. 1(*b*). It is clear from this discussion that every sampled-data filter may be represented in this form and it is often convenient to do so; nevertheless some caution must be used in discussing the behaviour of this representation. In general, the division into a sampler and a continuous linear filter does not correspond to any physical division in the actual filter, and, in particular, no attempt must be made to discuss the response of K to a continuous input at the point x^*. Even if the sampled-data filter may be physically divided into a sampler and a subsequent filter, this filter need not be identical with the continuous filter K.

As an example, consider a system with the structure shown in Fig. 2(*b*). The relation between $x(t)$ and $y(t)$ is clearly linear if the filters X and Y are linear, and the presence of the sampler S_1 ensures that $y(t)$ can depend only on the values of $x(t)$ at the sampling instants. Thus the system is a linear sampled-data filter according to the definition given at the beginning of this Section and, as shown in eqn. (4), it is certainly possible to find a system of the form shown in Fig. 2(*c*) which will give an equivalent relation between $x(t)$ and $y(t)$, where K is a suitably chosen continuous linear filter. Although the relation between $x(t)$ and $y(t)$ is unaltered by replacing the contents of the dotted boundary in Fig. 2(*b*) by the continuous filter K of Fig. 2(*c*), this does not, of course, mean that K and the contents of the dotted boundary have identical dynamical properties. They are only known to have the same effect on the special class of inputs which consist of a sequence of delta-function impulses synchronized with the sampling instants of the samplers S_1 and S_2.

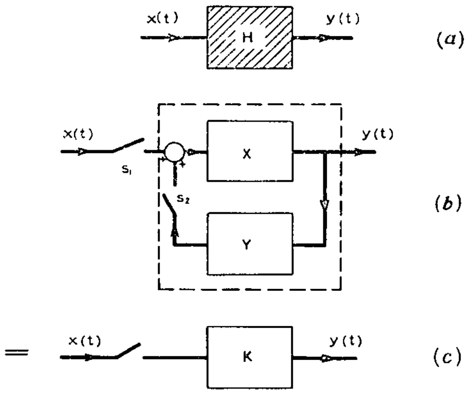


Fig. 2.—Filters with equivalent input-output relations.

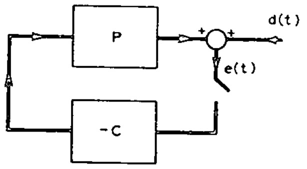


Fig. 3.—Sampled-data controller.

(4) CORRELATION FUNCTIONS AND SPECTRA IN SAMPLED-DATA SYSTEMS

Various auto- and cross-correlation functions and the corresponding spectral densities will be required for the sampled-data filter shown in Fig. 1(*b*). These are given by Ragazzini and Franklin[3] and are listed below, with a complete derivation in one case to illustrate the method.

The cross-correlation function, Φ_{ab}, of two functions $a(t)$ and $b(t)$ will be defined by

$$\Phi_{ab}(u) = \lim_{T_0 \to \infty} \frac{1}{2T_0}\int_{-T_0}^{+T_0} a(t)b(t+u)dt = \langle a(t)b(t+u)\rangle \quad . \quad (5)$$

and since the functions are assumed to be members of a stationary ergodic random process, the time average could be replaced by an assembly average if desired. When $b(t) \equiv a(t)$, the function $\Phi_{aa}(u)$ is known as the auto-correlation function of $a(t)$. The cross spectral density, $S_{ab}(s)$, of $a(t)$ and $b(t)$ is defined as the Fourier transform of $\Phi_{ab}(u)$, i.e.

$$\left.\begin{aligned} S_{ab}(s) &= \int_{-\infty}^{+\infty} \Phi_{ab}(u)\varepsilon^{-j\omega u}du \qquad \text{for } s = j\omega \\ &= \text{analytic continuation for other values of } s \end{aligned}\right\} \quad . \quad (6)$$

and when $b(t) \equiv a(t)$, the corresponding function $S_{aa}(s)$ is called the power spectrum of $a(t)$. Corresponding to eqn. (6), of course, is the inverse transform

$$\Phi_{ab}(u) = \frac{1}{2\pi j}\int_{-j\infty}^{+j\infty} S_{ab}(s)\varepsilon^{su}ds$$

The required relations for the system of Fig. 1(*b*) will now be dealt with in turn; the complete derivation given of result (iv) typifies the methods used in handling sampled time series.

(i) For the system in Fig. 1(*b*),

$$y(t) = \int_{-\infty}^{+\infty} k(u)x^*(t-u)du \quad [k(u) = 0 \text{ for } u < 0] \quad . \quad (7)$$

$$x^*(t) = \sum_{r=-\infty}^{+\infty} x(rT)\delta(t-rT) \quad . \quad . \quad . \quad . \quad (8)$$

(ii) Since K is a continuous filter, the output spectral density and the cross spectral density of input and output are given by the well-known relations[4]

$$S_{yy}(s) = K(s)K(-s)S_{x^*x^*}(s) \quad . \quad . \quad . \quad . \quad (9)$$

$$S_{x^*y}(s) = K(s)S_{x^*x^*}(s) = S_{yx^*}(-s) \quad . \quad . \quad (10)$$

(iii) As shown by Ragazzini and Franklin,[3] the auto-correlation function and corresponding spectral density of $x^*(t)$ are given by

$$\Phi_{x^*x^*}(u) = \frac{1}{T}\sum_{r=-\infty}^{+\infty} \Phi_{xx}(rT)\delta(u-rT) \quad . \quad . \quad (11)$$

$$S_{x^*x^*}(j\omega) = \frac{1}{T}\sum_{r=-\infty}^{+\infty} \Phi_{xx}(rT)\varepsilon^{-j\omega rT} \quad . \quad . \quad . \quad (12)$$

(iv) Finally, the cross-correlation function and cross spectral density for $x(t)$ and $y(t)$ will be derived as an illustration of the methods used.

From the definition given in eqn. (5), by splitting the range of integration into segments of length T, we obtain

$$\Phi_{xy}(u) = \lim_{N\to\infty} \frac{1}{(2N+1)T} \sum_{n=-N}^{+N} \int_{(n-\frac{1}{2})T}^{(n+\frac{1}{2})T} x(t)y(t+u)dt$$

$$= \lim_{N\to\infty} \frac{1}{(2N+1)T} \sum_{n=-N}^{+N} \int_{(n-\frac{1}{2})T}^{(n+\frac{1}{2})T} x(t-u)y(t)dt$$

Substituting for $y(t)$ from eqn. (2) gives

$$\Phi_{xy}(u) = \lim_{N\to\infty} \frac{1}{(2N+1)T} \times \sum_{n=-N}^{+N} \int_{(n-\frac{1}{2})T}^{(n+\frac{1}{2})T} x(t-u) \sum_{r=-\infty}^{+\infty} k(t-rT)x(rT)dt$$

or

$$\Phi_{xy}(u) = \lim_{N\to\infty} \frac{1}{(2N+1)T} \times \sum_{n=-N}^{+N} \int_{(n-\frac{1}{2})T}^{(n+\frac{1}{2})T} \sum_{r=-\infty}^{+\infty} x(t-u)x(rT+nT)k(t-rT-nT)dt$$

Now put $v = t - nT$, which reduces the above to

$$\Phi_{xy}(u) = \lim_{N\to\infty} \frac{1}{(2N+1)T} \times \sum_{n=-N}^{+N} \int_{-\frac{1}{2}T}^{+\frac{1}{2}T} \sum_{r=-\infty}^{+\infty} k(v-rT)x(rT+nT)x(v-u+nT)dv$$

$$= \frac{1}{T} \sum_{r=-\infty}^{+\infty} \int_{-\frac{1}{2}T}^{+\frac{1}{2}T} k(v-rT) \times \left[\lim_{N\to\infty} \frac{1}{(2N+1)} \sum_{n=-N}^{+N} x(rT+nT)x(v-u+nT) \right] dv$$

and it may be proved that, when $x(t)$ is stationary and ergodic,

$$\lim_{N\to\infty} \frac{1}{(2N+1)} \sum_{n=-N}^{+N} x(nT)x(nT+t) = \Phi_{xx}(t) \text{ (exactly)}$$

Using this result, the expression for $\Phi_{xy}(u)$ becomes

$$\Phi_{xy}(u) = \frac{1}{T} \sum_{r=-\infty}^{+\infty} \int_{-\frac{1}{2}T}^{+\frac{1}{2}T} k(v-rT)\Phi_{xx}(v-u-rT)dv$$

$$= \frac{1}{T} \int_{-\infty}^{+\infty} k(w)\Phi_{xx}(w-u)dw \quad \text{(say)}$$

Since Φ_{xx} is an even function of its argument, this reduces finally to

$$\Phi_{xy}(u) = \frac{1}{T} \int_{-\infty}^{+\infty} k(w)\Phi_{xx}(u-w)dw \quad . \quad . \quad . \quad (13)$$

and correspondingly,

$$S_{xy}(s) = \frac{1}{T} K(s)S_{xx}(s) = S_{yx}(-s) \quad . \quad . \quad . \quad (14)$$

This completes the set of spectrum relations which will be needed. but before leaving this topic it is worth defining two operations on spectra which will be required in the discussion of the Wiener optimum controller.

A power spectrum, $S_{aa}(s)$, is said to be *Wiener-factorizable* if it is possible to write

$$S_{aa}(s) = S^1_{aa}(s)S^2_{aa}(s) \quad . \quad . \quad . \quad . \quad . \quad (15)$$

where $S^1_{aa}(s)$ has all the zeros and singularities of $S_{aa}(s)$ in the left half-plane and is free from zeros and singularities in the right half-plane, while $S^2_{aa}(s)$ has all the zeros and singularities of $S_{aa}(s)$ in the right half-plane and is free from them in the left half-plane.

The decomposition of a function, F(s), of the complex variable s given by

$$F(s) = [F(s)]_+ + [F(s)]_- \quad . \quad . \quad . \quad . \quad (16)$$

will also be important, where

$$\left.\begin{aligned} [F(s)]_+ &= \frac{1}{2\pi j}\int_0^{\infty} \varepsilon^{-j\omega t}\left[\int_{-\infty}^{+\infty} F(ju)\varepsilon^{jut}du\right]dt \quad \text{for } s = j\omega \\ &= \text{analytic continuation} \qquad\qquad \text{for } s \neq j\omega \end{aligned}\right\} \quad . \quad (17)$$

$[F(s)]_+$ is then analytic and bounded in the right half-plane, while $[F(s)]_-$ is analytic and bounded in the left half-plane.

(5) A REARRANGEMENT ANALOGOUS TO PRICE'S METHOD FOR CONTINUOUS SYSTEMS

Returning now to the basic sampled-data regulator of Fig. 3, the block diagrams shown in Fig. 4(a) represent systems which give the same relation between $e(t)$ and $d(t)$. C_1 is physically realizable because it is constructed by physical interconnection of the two physically realizable blocks P and C, and so to every system of type A [Fig. 4(a)] there corresponds a physically realizable system of type B. The truth of the converse follows in the same way from the second sequence of equivalent systems given in Fig. 4(b), so arrangements A and B are completely equivalent so far as the relation between $e(t)$ and $d(t)$ is concerned.

From Fig. 4(b) it might appear that even the best C_1 would only correspond to the best C of a particular class of sampled-data filters with a sampler in the feedback loop. However, in view of the remarks in Section 3, this is not the case, and the equivalent C gives the optimum controller for the conventional arrangement represented by A in the class of all linear sampled-data filters.

Having established the equivalence of arrangements A and B, it is now possible to proceed to find the optimum linear C_1 which minimizes $\langle e^2 \rangle$. This can quite easily be done after re-drawing B in the equivalent form shown in Fig. 5. It is permissible to invert the order of the blocks P and C_1, as shown, since each is a continuous linear filter.

(6) OPTIMUM C_1 AND MINIMUM MEAN-SQUARE ERROR

The C_1 which minimizes $\langle e^2 \rangle$ follows immediately from the arrangement shown in Fig. 5 using the conventional Wiener theory:

$$C_1(s) = \frac{1}{S^1_{gg}(s)}\left[\frac{S_{gd}(s)}{S^2_{gg}(s)}\right]_+ \quad . \quad . \quad . \quad . \quad (18)$$

where $S^1_{gg}(s)$ and $S^2_{gg}(s)$ are the Wiener factors of $S_{gg}(s)$ (assuming this is factorizable); the notation $[F(s)]_+$ has been explained in Section 4.

The various terms in eqn. (18) will now be evaluated for the system shown in Fig. 5. The following relations arise from the results given in Section 4:

$$S_{gd}(s) = \frac{1}{T}P(-s)S_{dd}(s) \quad . \quad . \quad . \quad . \quad (19)$$

$$S_{gg}(s) = P(s)P(-s)S_{d^*d^*}(s) . \quad . \quad . \quad . \quad (20)$$

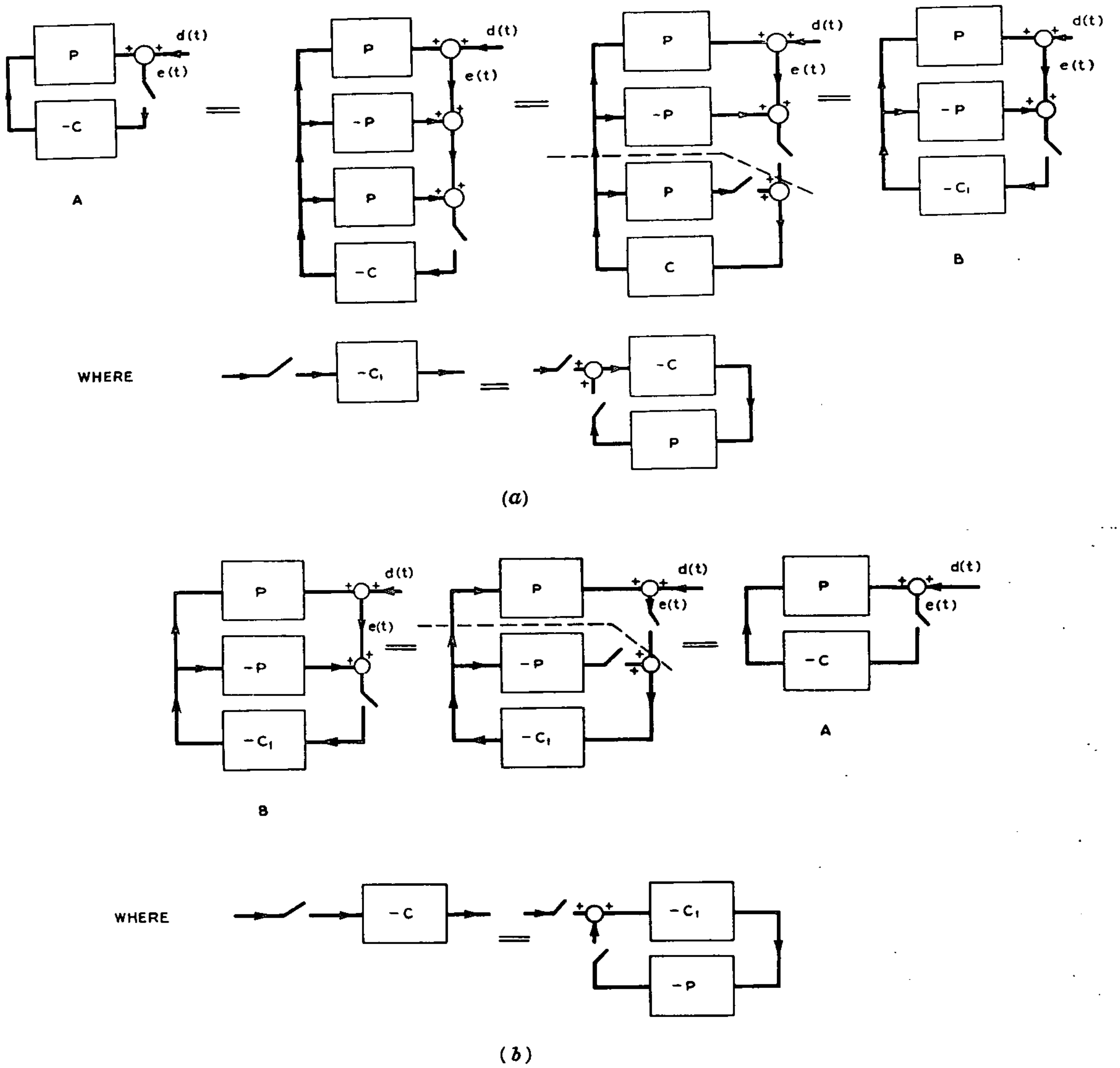


Fig. 4.—Block diagrams of equivalent arrangements.

(*a*) Relation of C_1 to C.
(*b*) Relation of C to C_1.

Attention will be restricted to stable transfer functions P(s) of the form

$$\mathrm{P}(s) = \mathrm{P}^1(s)\mathrm{P}^2(s)\varepsilon^{-s\tau} \quad . \quad . \quad . \quad . \quad . \quad (21)$$

where P(s)$\varepsilon^{s\tau}$ is a rational function, and $\mathrm{P}^1(s)$ and $\mathrm{P}^2(s)$ have the properties of Wiener factors. Then

$$\mathrm{S}_{gd}(s) = \frac{1}{T}\mathrm{P}^1(-s)\mathrm{P}^2(-s)\varepsilon^{s\tau}\mathrm{S}_{dd}(s) \quad . \quad . \quad . \quad (22)$$

$$\mathrm{S}^1_{gg}(s) = \mathrm{P}^1(s)\mathrm{P}^2(-s)\mathrm{S}^1_{d^\circ d^\circ}(s) \quad . \quad . \quad . \quad . \quad (23)$$

$$\mathrm{S}^2_{gg}(s) = \mathrm{P}^1(-s)\mathrm{P}^2(s)\mathrm{S}^2_{d^\circ d^\circ}(s) \quad . \quad . \quad . \quad . \quad (24)$$

Substituting these in eqn. (18) gives the optimum $\mathrm{C}_1(s)$ in the form

$$\mathrm{C}_1(s) = \frac{1}{\mathrm{P}^1(s)\mathrm{P}^2(-s)\mathrm{S}^1_{d^\circ d^\circ}(s)}\left[\frac{\varepsilon^{s\tau}}{T}\,\frac{\mathrm{P}^2(-s)}{\mathrm{P}^2(s)}\,\frac{\mathrm{S}_{dd}(s)}{\mathrm{S}^2_{d^\circ d^\circ}(s)}\right]_+ \quad . \quad (25)$$

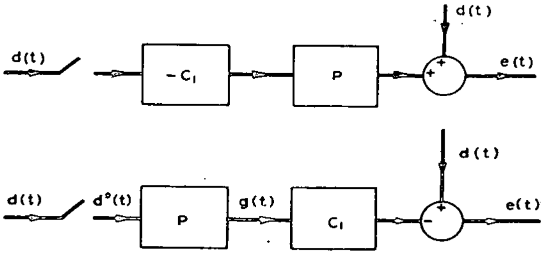


Fig. 5.—Open-loop configuration for control system.

All the quantities appearing in this are known, so it gives the desired solution of the problem. It will be convenient also to have the result in the slightly rearranged form

$$\mathrm{C}_1(s) = \frac{1}{\mathrm{P}^1(s)\mathrm{P}^2(-s)\mathrm{S}^1_{dd}(s)} \times \frac{1}{\mathrm{S}^1_{d^\circ d^\circ}(s)} \times \mathrm{S}^1_{dd}(s)$$
$$\times\left[\frac{\varepsilon^{s\tau}}{\mathrm{S}^1_{dd}(s)}\,\frac{\mathrm{P}^2(-s)\mathrm{S}^1_{dd}(s)}{\mathrm{P}^2(s)}\,\frac{1}{T}\,\frac{\mathrm{S}_{dd}(s)}{\mathrm{S}^2_{d^\circ d^\circ}(s)}\right]_+ \quad . \quad (26)$$

Notice that, when $P(s) \equiv 1$, $C_1(s)$ becomes simply the optimum filter for reconstructing $d(t)$ from the sampled signal $d^*(t)$. Denoting this by $C_{1r}(s)$, eqn. (25) gives

$$C_{1r}(s) = \frac{1}{S^1_{d^*d^*}(s)}\left[\frac{1}{T}\frac{S_{dd}(s)}{S^2_{d^*d^*}(s)}\right]_+ \quad . \quad . \quad . \quad (27)$$

These results should be compared with Price's optimum $C_1(s)$ for the continuous case, which will be denoted by $C_{1c}(s)$:

$$C_{1c}(s) = \frac{1}{P^1(s)P^2(-s)S^1_{dd}(s)}\left[\frac{P^2(-s)S^1_{dd}(s)\varepsilon^{s\tau}}{P^2(s)}\right]_+ \quad . \quad (28)$$

Comparing eqns. (27) and (28) with eqn. (26) it is seen that the optimum $C_1(s)$ for the sampled-data case is Price's optimum continuous $C_{1c}(s)$ in series with the optimum data reconstruction filter if, and only if,

$$\left[\frac{1}{S^1_{dd}(s)}\frac{P^2(-s)S^1_{dd}(s)\varepsilon^{s\tau}}{P^2(s)}\frac{1}{T}\frac{S_{dd}(s)}{S^2_{d^*d^*}(s)}\right]_+$$

$$\equiv \frac{1}{S^1_{dd}(s)} \times \left[\frac{P^2(-s)S^1_{dd}(s)\varepsilon^{s\tau}}{P^2(s)}\right]_+ \times \left[\frac{1}{T}\frac{S_{dd}(s)}{S^2_{d^*d^*}(s)}\right]_+ \quad . \quad (29)$$

One obvious case in which this factorization is valid arises when $P(s)$ is minimum phase, in which case $\tau = 0$, $P^1(s) \equiv P(s)$ and $P^2(s) \equiv 1$; other cases will be discussed later.

Having obtained the optimum $C_1(s)$, it is of direct interest to calculate the corresponding value of the mean-square error. The power spectrum of $e(t)$ is

$$S_{ee}(s) = S_{dd}(s) - \frac{1}{T}P(s)C_1(s)S_{dd}(s) - \frac{1}{T}P(-s)C_1(-s)S_{dd}(s)$$
$$+ C_1(s)C_1(-s)P(s)P(-s)S_{d^*d^*}(s) \quad . \quad (30)$$

from which $\langle e^2 \rangle$ can be obtained using

$$\langle e^2 \rangle = \frac{1}{2\pi j}\int_{-j\infty}^{+j\infty} S_{ee}(s)ds \quad . \quad . \quad . \quad . \quad . \quad (31)$$

Although the method used here gives the form of the optimum filter $C_1(s)$ directly, the form of the optimum $C(s)$ is of greater interest for the purpose of synthesizing an approximate optimum controller. From consideration of the block diagram in Fig. 4(*b*) showing the relation between C and C_1, it follows in a straightforward manner that

$$C(s) = \frac{C_1(s)}{[1 - PC_1(z)]_{z=\varepsilon^{sT}}} \quad . \quad . \quad . \quad . \quad (32)$$

where $PC_1(z)$ is the *z*-transform[3] corresponding to the Laplace transform $P(s)C_1(s)$. The explicit form of $C(s)$ for a particular simple system is given in Section 9.

(7) OPTIMUM CONTROLLER FOR A CLASS OF DISTURBANCE SPECTRA

$C_1(s)$ and $\langle e^2 \rangle$ will be evaluated for a particular but very extensive class of disturbance spectra.

Laning and Battin[4] show that any bounded auto-correlation function, the square of whose magnitude is integrable over the infinite interval, can be approximated in the mean by a sequence of terms of the form $A_k\varepsilon^{-c_k|u|}$ with $c_k > 0$. Thus any autocorrelation function likely to be of interest can be approximated arbitrarily closely, for the purpose of computing mean-square errors, by a sum of the form

$$\Phi_{dd}(u) = \sum_{k=1}^{n} A_k\varepsilon^{-c_k|u|} \quad . \quad . \quad . \quad . \quad . \quad (33)$$

If $S_{dd}(s)$ is the power spectrum corresponding to $\Phi_{dd}(u)$, and $S_{d^*d^*}(s)$ the spectral density of the corresponding sampled signal, using eqns. (6), (11) and (12) it is not difficult to show that

$$S_{dd}(s) = \sum_{k=1}^{n}\frac{2A_kc_k}{c_k^2 - s^2} \quad . \quad . \quad . \quad . \quad (34)$$

and that

$$S_{d^*d^*}(s) = \frac{1}{T}\sum_{k=1}^{n}\frac{A_k(1 - \varepsilon^{-2c_kT})}{(1 - \varepsilon^{-c_kT}\varepsilon^{-sT})(1 - \varepsilon^{-c_kT}\varepsilon^{sT})} \quad . \quad (35)$$

Eqn. (25) for $C_1(s)$ will now be evaluated for spectra of this particular form. Paying attention first to the square bracket on the right-hand side of eqn. (25), and denoting its contents by $F(s)$, it follows from eqns. (16) and (17) that it is possible to write

$$[F(s)]_+ = \int_0^\infty f(t)\varepsilon^{-st}dt \text{ with } s = j\omega$$

where
$$f(t) = \frac{1}{2\pi j}\int_{-j\infty}^{+j\infty}F(s)\varepsilon^{st}ds$$

When $S_{dd}(s)$ and $S_{d^*d^*}(s)$ are given by eqns. (34) and (35), inspection of the form taken by $F(s)$ shows that, in evaluating $f(t)$, the integration contour may be closed by a large semicircle in the left half-plane when $t > 0$, so that

$$f(t) = \sum \text{res}\,[F(s)\varepsilon^{st}] \quad (\text{for } t > 0) \quad . \quad . \quad . \quad (36)$$

the sum of the residues being taken over all poles of $F(s)\varepsilon^{st}$ in the left half-plane. The factors $P^2(s)$ and $S^2_{d^*d^*}(s)$ which appear in the denominator of $F(s)$ have, by definition, no zeros in the left half-plane, while the possibility of $P^2(-s)$ having singularities in the left half-plane is excluded by the fact that attention is limited to stable transfer functions $P(s)$. Thus, the only singularities of $F(s)\varepsilon^{st}$ in the left half-plane are the simple poles of $S_{dd}(s)$ at $s = -c_k$, and eqn. (36) may be evaluated immediately:

$$f(t) = \sum_{k=1}^{n}\frac{\varepsilon^{-c_k\tau}}{T}\frac{P^2(c_k)}{P^2(-c_k)}\frac{\varepsilon^{-c_kt}}{S^2_{d^*d^*}(-c_k)}\lim_{s\to -c_k}[(s + c_k)S_{dd}(s)]$$

or
$$f(t) = \frac{1}{T}\sum_{k=1}^{n}\varepsilon^{-c_k\tau}\frac{P^2(c_k)}{P^2(-c_k)}\frac{\varepsilon^{-c_kt}}{S^2_{d^*d^*}(-c_k)}A_k \quad (\text{for } t > 0)$$

whence

$$[F(s)]_+ = \frac{1}{T}\sum_{k=1}^{n}\frac{P^2(c_k)}{P^2(-c_k)}\frac{A_k\varepsilon^{-c_k\tau}}{S^2_{d^*d^*}(-c_k)}\frac{1}{s + c_k}$$

and the optimum $C_1(s)$ then follows from eqn. (25) as

$$C_1(s) = \frac{1}{P^1(s)P^2(-s)S^1_{d^*d^*}(s)}\frac{1}{T}\sum_{k=1}^{n}\frac{P^2(c_k)}{P^2(-c_k)}\frac{A_k\varepsilon^{-c_k\tau}}{S^2_{d^*d^*}(-c_k)}\frac{1}{s + c_k}$$
$$. \quad . \quad . \quad . \quad (37)$$

Since the following combination of factors will occur frequently from now on, it will be convenient to define

$$Q(A_k, c_k) = Q_k = \frac{1}{T}\frac{P^2(c_k)}{P^2(-c_k)}\frac{A_k\varepsilon^{-c_k\tau}}{S^2_{d^*d^*}(-c_k)} \quad . \quad (38)$$

when eqn. (37) takes the form

$$C_1(s) = \frac{1}{P^1(s)P^2(-s)S^1_{d^*d^*}(s)}\sum_{k=1}^{n}\frac{Q_k}{s + c_k} \quad . \quad . \quad (39)$$

The main difficulty in handling this when $n > 1$ is the cumbersome algebraic form of the factors $S^1_{d^*d^*}(s)$ and $S^2_{d^*d^*}(s)$ when written out explicitly.

Before going on to deal with the mean-square error, however, it is interesting to consider some special results which hold for $n = 1$. In this particular case, eqn. (37) becomes

$$\mathrm{C}_1(s) = \frac{1}{\mathrm{P}^1(s)\mathrm{P}^2(-s)\mathrm{S}^1_{d^*d^*}(s)}\,\frac{1}{T}\,\frac{\mathrm{P}^2(c)}{\mathrm{P}^2(-c)}\,\frac{A\varepsilon^{-c\tau}}{\mathrm{S}^2_{d^*d^*}(-c)}\,\frac{1}{s+c}\quad . \quad (40)$$

(where $c = c_1$) while, from eqn. (27), the optimum data-reconstruction filter is

$$\mathrm{C}_{1r} = \frac{1}{\mathrm{S}^1_{d^*d^*}(s)}\,\frac{1}{T}\,\frac{A}{\mathrm{S}^2_{d^*d^*}(-c)}\,\frac{1}{s+c}\quad . \quad . \quad . \quad (41)$$

Since $\mathrm{S}_{dd}(s) = 2Ac/(c^2 - s^2)$ in this case, the right-hand side of eqn. (28) can easily be evaluated, and the optimum continuous controller is

$$\mathrm{C}_{1c} = \frac{1}{\mathrm{P}^1(s)\mathrm{P}^2(-s)}\,\frac{\mathrm{P}^2(c)}{\mathrm{P}^2(-c)}\varepsilon^{-c\tau}\quad . \quad . \quad . \quad (42)$$

Comparison of eqns. (41) and (42) with eqn. (40) shows immediately that

$$\mathrm{C}_1(s) = \mathrm{C}_{1r}(s)\mathrm{C}_{1c}(s)$$

so again we have the result that the optimum sampled-data C_1 is equivalent to the optimum continuous C_{1c} in series with the optimum data reconstruction filter C_{1r}. This may also be proved by checking directly that the factorization condition, eqn. (29), is satisfied. It should be noted that, although this attractively simple result is valid for this very popular spectrum, it does not appear to be generally true when $n > 1$.

(8) MINIMUM MEAN-SQUARE ERROR

Returning now to the general case, $n > 1$, the minimum attainable value of $\langle e^2 \rangle$ is calculated by using eqns. (30) and (31) with $\mathrm{C}_1(s)$ given by eqn. (39). The terms in eqn. (30) can be integrated separately, the first giving immediately

$$\frac{1}{2\pi j}\int_{-j\infty}^{+j\infty}\mathrm{S}_{dd}(s)ds = \sum_{k=1}^{n} A_k \quad . \quad . \quad . \quad . \quad (43)$$

Considering the second term of eqn. (30), it is necessary to evaluate the contribution to eqn. (31) from

$$\frac{1}{T}\mathrm{P}(s)\mathrm{C}_1(s)\mathrm{S}_{dd}(s) = \frac{1}{T}\,\frac{\mathrm{P}^1(s)\mathrm{P}^2(s)\varepsilon^{-s\tau}}{\mathrm{P}^1(s)\mathrm{P}^2(-s)\mathrm{S}^1_{d^*d^*}(s)} \times \sum_{k=1}^{n}\frac{Q_k}{s+c_k}$$
$$\times \sum_{l=1}^{n} A_l\left(\frac{1}{c_l+s}+\frac{1}{c_l-s}\right) \quad . \quad (44)$$

Now $\mathrm{S}^1_{d^\circ d^\circ}$ has the form

$$\frac{\mathrm{N}(s)}{\prod_{k=1}^{n}(1-\varepsilon^{c_kT}\varepsilon^{-sT})}$$

where $\mathrm{N}(s)$ is a polynominal in ε^{-sT} of degree not exceeding $n-1$. Thus $1/\mathrm{S}^1_{d^\circ d^\circ}(s)$ remains bounded when $|s| \to \infty$ provided that $\mathscr{R}(s) > 0$, and it is seen that every term of eqn. (44) tends to zero faster than $1/|s|$ when $|s| \to \infty$ with $\mathscr{R}(s) > 0$. The contour of integration in eqn. (31) may therefore be closed by a large semicircle in the right half-plane; it is then necessary to remember that the closed contour is described in a negative sense. The only poles of the integrand in eqn. (44) in the right half-plane are at $s = c_l$, and they are simple, so the residues can be obtained by multiplying through by $s - c_l$ and letting $s \to c_l$, with the result

$$\frac{1}{2\pi jT}\int_{-j\infty}^{+j\infty}\mathrm{P}(s)\mathrm{C}_1(s)\mathrm{S}_{dd}(s)ds$$
$$= \frac{1}{T}\sum_{l=1}^{n}\frac{A_l}{\mathrm{S}^1_{d^*d^*}(c_l)}\,\frac{\mathrm{P}^2(c_l)}{\mathrm{P}^2(-c_l)}\varepsilon^{-c_l\tau}\sum_{k=1}^{n}\frac{Q_k}{c_k+c_l}\quad . \quad (45)$$

In a similar way, by closing the contour with a semicircle in the left half-plane, it can be shown that the same contribution is obtained from the third term of eqn. (30).

When written out fully the contribution from the last term of eqn. (30) arises from the integrand

$$\frac{1}{\mathrm{S}^1_{d^*d^*}(s)\mathrm{S}^1_{d^*d^*}(-s)}\sum_{k=1}^{n}\frac{Q_k}{c_k+s}\sum_{l=1}^{n}\frac{Q_l}{c_l-s}\mathrm{S}_{d^*d^*}(s)\quad . \quad (46)$$

Now $\qquad \mathrm{S}_{d^*d^*}(s) = \mathrm{S}^1_{d^*d^*}(s)\mathrm{S}^2_{d^*d^*}(s)$

while $\qquad \mathrm{S}_{d^*d^*}(-s) = \mathrm{S}^1_{d^*d^*}(-s)\mathrm{S}^2_{d^*d^*}(-s) = \mathrm{S}_{d^*d^*}(s)$

since $\mathrm{S}_{d^*d^*}(s)$ is even. Further, $\mathrm{S}^1_{d^*d^*}(-s)$ has all its poles and zeros in the right half-plane, while $\mathrm{S}^2_{d^*d^*}(-s)$ has all its poles and zeros in the left half-plane. $\mathrm{S}^1_{d^*d^*}(-s)$ and $\mathrm{S}^2_{d^*d^*}(-s)$ can therefore differ only by constant factors from $\mathrm{S}^2_{d^*d^*}(s)$ and $\mathrm{S}^1_{d^*d^*}(s)$ respectively, and the factorization can be carried out in such a way that $\mathrm{S}^1_{d^*d^*}(-s) = \mathrm{S}^2_{d^*d^*}(s)$. Introducing this into integrand (46), the first and last factors cancel. The contribution to eqn. (31) may then be evaluated by closing the contour with a semicircle in either half-plane. Choosing a semicircle in the left half-plane, the contour is described in a positive sense and it is only necessary to sum the residues at the poles at $s = -c_k$, with the result

$$\sum_{k=1}^{n}\sum_{l=1}^{n}\frac{Q_kQ_l}{c_k+c_l}\quad . \quad . \quad . \quad . \quad . \quad . \quad (47)$$

which is seen to be identical with eqn. (45), remembering that $\mathrm{S}^2_{d^*d^*}(-c_l) = \mathrm{S}^1_{d^*d^*}(c_l)$ with the present choice of factorization.

Collecting together the terms contributing to eqn. (31) gives

$$\langle e^2 \rangle = \sum_{k=1}^{n} A_k - \sum_{k=1}^{n}\sum_{l=1}^{n}\frac{Q_kQ_l}{c_k+c_l}\quad . \quad . \quad . \quad (48)$$

When the system is without control, $\langle e^2 \rangle = \langle e^2 \rangle_0 = \langle d^2 \rangle = \sum_{k=1}^{n} A_k$, so that the mimimum attainable value of $\langle e^2 \rangle/\langle e^2 \rangle_0$ is given by

$$\xi = \min\frac{\langle e^2 \rangle}{\langle e^2 \rangle_0} = 1 - \frac{1}{T^2\sum_{k=1}^{n}A_k}\sum_{k=1}^{n}\sum_{l=1}^{n}\frac{A_kA_l}{\mathrm{S}^1_{d^*d^*}(c_k)\mathrm{S}^1_{d^*d^*}(c_l)}$$
$$\times\frac{P^2(c_k)}{P^2(-c_k)}\,\frac{\mathrm{P}^2(c_l)}{\mathrm{P}^2(-c_l)}\times\frac{\varepsilon^{-(c_k+c_l)\tau}}{c_k+c_l}\quad . \quad (49)$$

where the explicit forms of Q_k and Q_l have been re-introduced, and it is assumed that the factorization is carried out in such a way that $\mathrm{S}^1_{d^*d^*}(s) = \mathrm{S}^2_{d^*d^*}(-s)$.

As in the case of $\mathrm{C}_1(s)$, the greatest difficulty in handling this is the clumsy algebraic form of the factors $\mathrm{S}^1_{d^\circ d^\circ}(s)$ when $n > 1$.

(9) EVALUATION OF THE RESULTS FOR SIMPLE EXAMPLES

A programme to evaluate the right-hand side of eqn. (49) has been written for a digital computer. This deals with the case where $n = 2$ [i.e. two terms in expression (35) for $\mathrm{S}_{d^*d^*}(s)$] and $\mathrm{P}^2(s)$ is a polynomial of degree not greater than two. It was used initially to calculate ξ for the system shown in Fig. 6, for which the continuous case has been treated by Price.[1]

The plant consists of a transfer lag with unit time-constant,

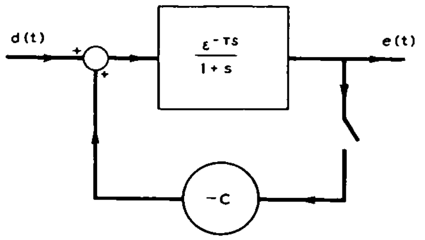


Fig. 6.—System used in illustrative examples.

together with a distance-velocity lag, τ, while the disturbance, with spectrum $D(s) = 2c/(c^2 - s^2)$, enters through the plant by the same path as the correction. For the corresponding continuous system, Price obtains the result

$$\left.\begin{aligned}\xi &= 1-(1+2\tau+2\tau^2)\varepsilon^{-2\tau} && \text{when } c=1\\ &= 1-\frac{2c(1+c)}{(1-c)^2}[\phi(2)-2\phi(1+c)+\phi(2c)] && \\ & && \text{when } c\neq 1\end{aligned}\right\} \quad . \quad (50)$$

where $\phi(x) = \varepsilon^{-\tau x}/x$.

Using the programme described above, corresponding calculations have been carried out for sampled systems with various values of c. The results for $c = 0{\cdot}05$ and for $c = 100$ are recorded in Figs. 7 and 8. Note that:

(i) For fixed values of c and τ, ξ increases with T. This increase is much more rapid for large values of c than for small values.

(ii) For fixed values of c and T, ξ increases with τ and approaches unity as $\tau \to \infty$.

These properties are all as expected intuitively. The curves for $T = 0$ are drawn using Price's formulae above; the convergence of the computed curves to these when T becomes small provides a check on the theory and the programme for the sampled case.

The form of C(s) will also be obtained and compared with the optimum continuous controller for the system shown in Fig. 6, but to avoid unnecessary algebraic complication the disturbance with the spectrum considered above will be replaced by white noise. This affects the measured variable only after passing through the block representing the controlled plant, so

$$S_{dd}(s) = \frac{2}{1-s^2} \quad \text{and} \quad P(s) = \frac{\varepsilon^{-s\tau}}{1+s}$$

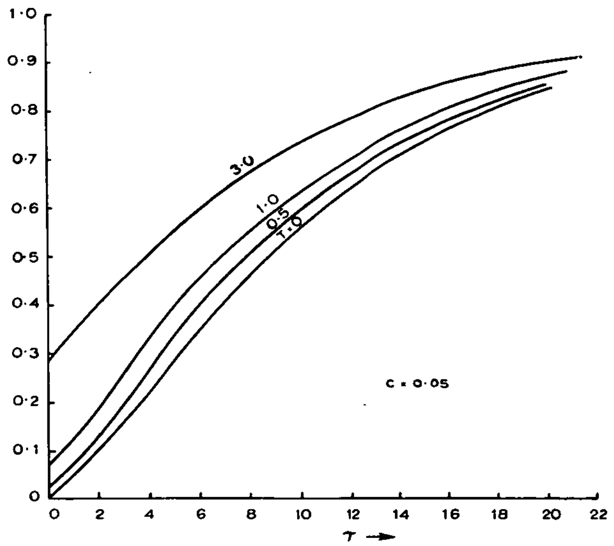


Fig. 7.—ξ as a function of τ for $c = 0{\cdot}05$.

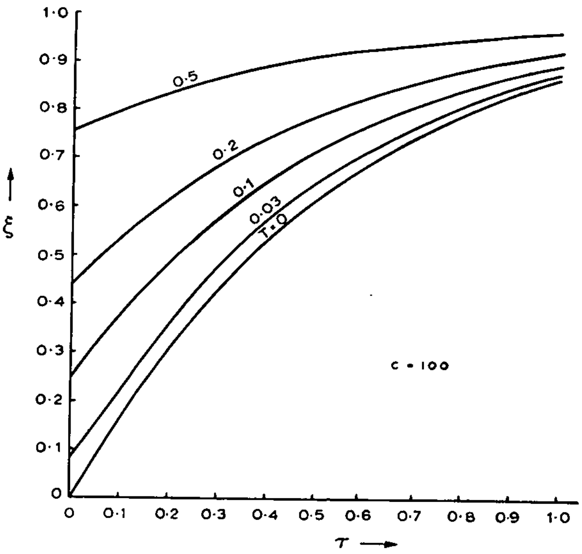


Fig. 8.—ξ as a function of τ for c = 100.

and correspondingly,

$$P^1(s) = \frac{1}{1+s} \quad \text{and} \quad P^2(s) = 1$$

The sampled spectrum $S_{d^*d^*}(s)$ corresponding to the above form of $S_{dd}(s)$ is given by eqn. (35):

$$S_{d^*d^*}(s) = \frac{1}{T}\,\frac{(1-\varepsilon^{-2T})}{(1-\varepsilon^{-T}\varepsilon^{-sT})(1-\varepsilon^{-T}\varepsilon^{sT})}$$

From eqn. (40) it now follows that

$$C_1(s) = \varepsilon^{-\tau}(1-\varepsilon^{-T}\varepsilon^{-sT}) \quad . \quad . \quad . \quad . \quad (51)$$

and hence that

$$P(s)C_1(s) = \frac{\varepsilon^{-\tau}(1-\varepsilon^{-T}\varepsilon^{-sT})\varepsilon^{-s\tau}}{1+s} \quad . \quad . \quad (52)$$

If τ is written in the form $\tau = (p - \lambda)T$, where p is an integer and $0 \leqslant \lambda < 1$, the z-transform corresponding to eqn. (52) is

$$PC_1(z) = \varepsilon^{-pT}z^{-p} \quad . \quad . \quad . \quad . \quad . \quad (53)$$

C(s) for the optimum controller may now be obtained by substituting from eqns. (51) and (53) into eqn. (32):

$$C(s) = \frac{\varepsilon^{-\tau}[1-\varepsilon^{-(1+s)T}]}{1-\varepsilon^{-p(1+s)T}} \quad . \quad . \quad . \quad . \quad (54)$$

The transfer function of the optimum continuous controller can be obtained by substituting the above form of P(s) in eqn. (42) to give $C_{1c}(s)$. Then if $C_c(s)$ is the corresponding filter for use in a simple feedback loop,

$$C_c(s) = \frac{C_{1c}(s)}{1-P(s)C_{1c}(s)} \quad . \quad . \quad . \quad . \quad (55)$$

which corresponds to eqn. (32) for the sampled case. In the present example,

$$C_c(s) = \frac{(1+s)\varepsilon^{-\tau}}{1-\varepsilon^{-\tau(1+s)}} \quad . \quad . \quad . \quad . \quad . \quad (56)$$

Difficulty is experienced in comparing $C_c(s)$ with the limiting behaviour of C(s) when $T \to 0$, because C(s) is assumed to operate on a sequence of delta functions. This difficulty can be avoided by representing C(s) as a zero-order hold, with transfer function $(1 - \varepsilon^{-sT})/s$, in series with a filter C′(s). We must then have

$$C'(s) = \frac{sC(s)}{1 - \varepsilon^{-sT}} = \frac{s}{1 - \varepsilon^{-sT}} \frac{\varepsilon^{-\tau}[1 - \varepsilon^{-(1+s)T}]}{1 - \varepsilon^{-p(1+s)T}} \quad . \quad (57)$$

The zero-order hold converts the sequence of delta functions leaving the sampler into a 'staircase' function, which approximates to the continuous function at the sampler input more and more closely as $T \rightarrow 0$. Correspondingly, C′(s) would be expected to approximate in the limit to the transfer function $C_c(s)$ of the continuous controller.

In examining the behaviour of eqn. (57) for small values of T, it must be remembered that $p \rightarrow \infty$ as $T \rightarrow 0$ in such a way that $pT \rightarrow \tau$; thus, $\varepsilon^{-p(1+s)T} \simeq \varepsilon^{-\tau(1+s)}$ when $T \rightarrow 0$. The remaining exponentials in eqn. (57) may be expanded in their exponents, neglecting powers beyond the first, when it follows that $C'(s) \rightarrow C_c(s)$ when $T \rightarrow 0$ for each value of s. The convergence of C′(s) to $C_c(s)$ is not uniform in s, but we may say that the behaviour of the optimum sampled-data controller with very short sampling interval approximates to that of the optimum continuous controller, except at very high frequencies. The frequency range over which the approximation is good may be extended as far as we please by taking a sufficiently short sampling interval.

(10) ACKNOWLEDGMENTS

The author wishes to thank Dr. P. C. Price for making his results available prior to publication and for much useful discussion. Valuable suggestions were also made by Mr. G. Hext. The programme described in Section 9 was written by Mr. D. B. Swinson during his tenure of a Vacation Scholarship at Billingham in 1959.

The work described forms part of the research programme of Imperial Chemical Industries Limited, Billingham Division.

(11) REFERENCES

(1) PRICE, P. C.: 'An Analytical Treatment of Process Controllability' (to be published).

(2) WIENER, N.: 'The Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications' (Wiley, 1948).

(3) RAGAZZINI, J. R., and FRANKLIN, G. F.: 'Sampled-Data Control Systems' (McGraw-Hill, 1958).

(4) LANING, J. H., and BATTIN, R. H.: 'Random Processes in Automatic Control' (McGraw-Hill, 1956).

Printed in Great Britain by
UNWIN BROTHERS LIMITED, WOKING AND LONDON

Group B

Chemical plants normally consist of a number of more or less distinguishable units such as reactors, distillation columns, mixers and evaporators, between which streams of material are transported. The flow rates and compositions of these streams depend on the design and operating conditions of the units, so the complete plant consists of a number of interconnected and interacting units. In the simplest case the connection may be a sequential one, in which the product from one unit forms the feed for the next, but frequently this pattern is complicated by the presence of bypasses, recycle loops or cross feeds between two chains of sequentially connected units. The problem of determining the most economic design and operating conditions in these circumstances is a problem associated with the plant as a whole, whose complex interconnected form usually leads to an exceedingly cumbersome mathematical formulation. The question therefore arises whether it is possible to break down the connecting structure to some extent and to define separate optimization problems for each of the units from which the plant is composed, such that solutions of each of these separate, and simpler problems may be adjoined to synthesise a solution of the optimization problem for the complete plant.

In the early 1960's it appeared that two mathematical techniques might be adapted to accomplish this, namely the algorithm of Dynamic Programming, and the Maximum Principle of Pontryagin. In their original forms, neither of the techniques was entirely suitable. Dynamic programming was indeed a method of decomposing optimization problems ininterconnected structures but, as originally developed, it required the connection to be sequential. The Maximum Principle, on the other hand, did not deal with interconnected discrete units at all, but was a result in the Calculus of Variations. However, an analogous/

analogous statement could clearly be formulated for sequential structures composed of discrete units, and it was expected that this would be generalisable to more complex structures.

The earliest attempts to generalise Dynamic Programming to structures with recycle loops met with immediate success, but publication B1 showed, by means of a simple counter example, that they led to results which were incorrect,, and traced the fallacy in the reasoning leading to the proposed generalisation of the sequential algorithm. Since the date of this publication valid extensions of the Dynamic Programming algorithm to complex structures have been developed by Nemhauser, Wilde and others.

The second approach, namely the development of a Discrete Maximum Principle analogous to Pontryagin's Principle in the variational calculus, had reached a much more advanced stage by 1964, at which time it formed the subject of a large number of papers and a textbook. However, in a seminar at Imperial College, London in 1964, the present writer suggested that the basis of this work, namely the Discrete Maximum Principle itself, was fallacious and supported this contention with the first example quoted in publication B2. The seminar was attended by Dr. F. Horn who devised a much more conclusive counter example and joined with the present writer in publication B2, in which a valid but weaker form of the Discrete Maximum Principle was suggested. In publication B3 we went on to show that this weak form is nothing more than a simple rearrangement of a formula of elementary differential calculus.

Publication B4 explores the relation between the generally valid weakened form of the Maximum Principle and the original strong form, which is true only in very special classes of problems. Some of these classes are identified explicitly. Since the date of these publications the relation between the weak and/

and strong Discrete Maximum Principles has been thoroughly investigated by pure mathematicians, notably Halkin at the University of California, who has identified classes of problems for which the strong form is valid, other than those enumerated in publication B4.

In publications B5 and B6 the valid form of the Maximum Principle is developed for interconnected structures of arbitrary complexity and it is shown that it may be a useful tool for practical computations. Finally publication B7 extends this work to situations in which the state of the plant is time-dependent.

B1

Reprinted from

CHEMICAL ENGINEERING SCIENCE

Le Journal International de Génie Chimique

PERGAMON PRESS

OXFORD · LONDON · NEW YORK · PARIS

Chemical Engineering Science, 1963, Vol. 18, pp. 215–217. Pergamon Press Ltd., Oxford. Printed in Great Britain.

Letters to the Editors

Comments on the paper Optimum cross-current extraction with product recycle

D. F. RUDD and E. D. BLUM

(*Received* 13 *November* 1962)

IN A recent paper[1] RUDD and BLUM have proposed a simple extension of the method of dynamic programming to deal with stagewise processes with product recycle. In view of the widespread occurrence of these processes, such a generalization is of great potential importance. However, it is the purpose of the present note to show that the proposed method is in fact fallacious, leading to the correct answer only in certain very special cases.

We first recapitulate the method briefly using the notation of RUDD and BLUM's paper. Referring to Fig. 1 of their paper, the object is to choose the operating conditions $\mathbf{W}_1, \ldots \mathbf{W}_N$ so as to maximize an objective function Q of the form

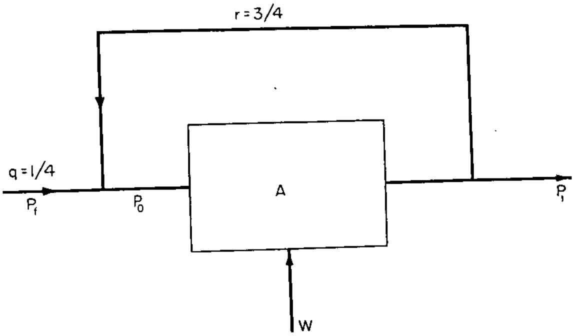


FIG. 1. Simple recycle system.

$$Q = R[(\mathbf{P}_f - \mathbf{P}_N)q] - \sum_{i=1}^{N} C(\mathbf{W}_i) \tag{1}$$

where the given functions R and C represent the increase in value of the process stream and the cost of applying the operating policy respectively. The quality $\mathbf{P}_0$ of the stream entering the first stage is related to $\mathbf{P}_f$ and $\mathbf{P}_N$ by the mixing condition

$$\mathbf{P}_0 = (q\mathbf{P}_f + r\mathbf{P}_N)/(q + r) \tag{2}$$

$\mathbf{P}_f$ may then be eliminated from equations (1) and (2), giving

$$Q = R[(\mathbf{P}_0 - \mathbf{P}_N)(q + r)] - \sum_{i=1}^{N} C(\mathbf{W}_i) \tag{3}$$

Let $\mathbf{W}_i{}^m(\mathbf{P}_0)$ represent the set of operating conditions which maximize Q for a given $\mathbf{P}_0$. They may be found by conventional dynamic programming computations, and in turn they determine a value of $\mathbf{P}_N$.

$$\mathbf{P}_N = \mathbf{P}_N[\mathbf{W}_i{}^m(\mathbf{P}_0), \mathbf{P}_0] \tag{4}$$

The pair of equations (2) and (4) may then be solved simultaneously for $\mathbf{P}_0$ and $\mathbf{P}_N$, which suffice to determine the optimum operating conditions and the maximum value of Q.

The right-hand side of equation (4) must, in general, be computed numerically through the dynamic programming tables, so equations (2) and (4) must be solved numerically, and the authors propose a particular iterative method for doing this. A given approximation to $\mathbf{P}_0$ determines a value of $\mathbf{P}_N$ through equation (4), and this in turn is used in equation (2) to determine the next approximation to $\mathbf{P}_0$. The value $\mathbf{P}_0 = \mathbf{P}_f$ is suggested as a suitable initial approximation.

The fallacy of the above procedure can best be exposed by considering a simple example. Fig. 1 shows a single stage process with recycle, in which P_f, P_0 and P_1 each represent a single quantity and there is one adjustable operating condition W. Block A is such that P_1 is related to $\mathbf{P}_0$ by

$$P_1 = 2P_0 - W^2 \tag{5}$$

while the mixing condition is

$$P_0 = \tfrac{1}{4}P_f + \tfrac{3}{4}P_1 \tag{6}$$

and W is to be chosen to maximize

$$Q = P_1 - P_f \tag{7}$$

for given P_f. This is a special case of the problem treated by RUDD and BLUM and is so simple that it can be solved directly without resort to their procedure. The relation between P_1 and P_f for any value of W can be found by eliminating $\mathbf{P}_0$ between equations (5) and (6), giving

$$P_1 = 2W^2 - P_f$$

whence

$$Q = P_1 - P_f = 2(W^2 - P_f) \tag{8}$$

and this takes its smallest value $Q = -2P_f$ at $W = 0$.

The procedure used by RUDD and BLUM is as follows: From equations (6) and (7)

$$Q = 4(P_1 - P_0) \tag{9}$$

corresponding to equation (3) of the general case. Using equation (5) for P_1, we see that Q has a unique stationary maximum at $W = W^m = 0$ for all values of P_0, and with this value of W

$$P_1 = P_1[W^m(P_0), P_0] = 2P_0 \qquad (10)$$

corresponding to equation (4) of the general case. Equations (6) and (10) are then solved simultaneously, giving

$$P_0 = -\tfrac{1}{2}P_f \text{ and } P_1 = -P_f \text{ with } W = W^m = 0,$$

and correspondingly

$$Q = -2P_f$$

However, direct solution has already shown us that this is the smallest value taken by Q for any choice of W!

This example is so simple that an iterative solution of equations (6) and (10) is unnecessary, but if we ignored this and applied the iterative procedure proposed by RUDD and BLUM we should find that the iterations diverged. We should therefore be prevented from actually arriving at the false conclusicn.

It is interesting to give a somewhat more general analysis of the RUDD and BLUM procedure. For simplicity attention will be restricted to the single stage system shown in Fig. 1, but we shall take a general functional relation

$$P_1 = f(P_0, W) \qquad (11)$$

to represent block A, a mixing condition of the form

$$P_0 = \alpha P_f + \beta P_1 \qquad (\alpha + \beta = 1) \qquad (12)$$

and an objective function

$$Q = P_1 - P_f - C(W) \qquad (13)$$

Elimination of P_f between equations (12) and (13) gives

$$Q = (P_1 - P_0)/\alpha - C(W) \qquad (14)$$

and it will be assumed that the form of $f(P_0, W)$ is such that Q has a single stationary maximum with respect to W at constant P_0. The procedure of RUDD and BLUM can now be followed without difficulty. W^m corresponds to the stationary value of Q at constant P_0 and therefore satisfies

$$\frac{1}{\alpha}\left(\frac{\partial f}{\partial W}\right)_{P_0} - \frac{dC}{dW} = 0 \qquad (15)$$

This value of W is then substituted into equations (11) and (12), which are solved for P_0 and P_1 and these in turn determine W through equation (15). In other words, the values of W, P_0 and P_1 taken to represent optimum operation satisfy equations (11), (12) and (15).

The problem may also be approached directly by eliminating P_0 from equations (11) and (12) to give an implicit equation for P_1 in terms of P_f and W. It is then required to find the value of W which maximizes Q, as given by equation (13) with this value for P_1. First consider the condition for Q to take a stationary value with respect to W at constant P_f. We have

$$\left(\frac{\partial Q}{\partial W}\right)_{P_f} = \left(\frac{\partial P_1}{\partial W}\right)_{P_f} - \frac{dC}{dW} \qquad (16)$$

while differentiation of equation (11) gives

$$dP_1 = \left(\frac{\partial f}{\partial P_0}\right)_W dP_0 + \left(\frac{\partial f}{\partial W}\right)_{P_0} dW$$

$$= \beta\left(\frac{\partial f}{\partial P_0}\right)_W dP_1 + \left(\frac{\partial f}{\partial W}\right)_{P_0} dW \quad \text{(using equation (12) with constant } P_f)$$

It follows that

$$\left(\frac{\partial P_1}{\partial W}\right)_{P_f} = \frac{(\partial f/\partial W)P_0}{1 - \beta(\partial f/\partial P_0)_W}$$

and substituting this into equation (16) we see that the condition for Q to take a stationary value is

$$\frac{(\partial f/\partial)W_{P_0}}{1 - \beta(\partial f/\partial P_0)_W} - \frac{dC}{dw} = 0 \qquad (17)$$

and the corresponding values of P_0, P_1, and W are obtained by solution of equations (11), (12) and (17).

In general equations (15) and (17) are not identical, so RUDD and BLUM's solution does not even correspond to a stationary value of Q with respect to W at constant P_f. Thus there are values of W in the neighbourhood of the value they determine which give larger values of Q.

However, in the particular case when $C(W) \equiv 0$, equation (15) reduces to

$$\frac{1}{\alpha}(\partial f/\partial W)_{P_0} = 0 \qquad (18)$$

and equation (17) to

$$\frac{(\partial f/\partial W)_{P_0}}{1 - \beta(\partial f/\partial P_0)_W} = 0 \qquad (19)$$

Satisfaction of (18) is now sufficient to ensure that (19) is also satisfied (unless $1 - \beta(\partial f/\partial P_0)_W = 0$ "accidentally"), so RUDD and BLUM's solution corresponds to a stationary value of Q when $C \equiv 0$. Nevertheless, as shown by our example, it would be fallacious to assume that this stationary value is necessarily a maximum.

To investigate its nature, we write down an expression for the second derivative

$$\left(\frac{\partial^2 Q}{\partial W^2}\right)_{P_f}$$

at a stationary point where $(\partial Q/\partial W)P_f = 0$. When $C \equiv 0$ this is

$$\left(\frac{\partial^2 Q}{\partial W^2}\right)_{P_f} = \frac{(\partial^2 f/\partial W^2)_{P_0}}{1 - \beta(\partial f/\partial P_0)_W} \qquad (20)$$

The solution of equations (11), (12) and (18) corresponds to a stationary maximum value of f at constant P_0, so $(\partial^2 f/\partial W^2)P_0 < 0$, but $(\partial^2 Q/\partial W^2)_{P_f}$ may have either sign depending on the sign of the denominator on the right-hand side of equation (20). Thus the solution of RUDD and BLUM may give either a maximum or a minimum value of Q with respect to W at constant P_f: in our example it gave a minimum.

It is not difficult to show that the necessary and sufficient condition for convergence of RUDD and BLUM's iterative method of solution of equations (11), (12) and (18) is

$$\left|\beta\left(\frac{\partial f}{\partial P_0}\right)_W\right| < 1 \qquad (21)$$

and this is also sufficient condition for the denominator of the right-hand side of equation (20) to be positive. Thus, when the iterative procedure proposed by RUDD and BLUM converges, the solution obtained corresponds to a stationary maximum value of Q with respect to W at constant P_f. (Provided $C \equiv 0$, of course).

We may summarize the general results obtained for the single stage process in which Q has a unique stationary maximum with respect to variations in W at constant P_0.

(i) In the general case with $C(W) \not\equiv 0$, the solution of Rudd and Blum does not correspond to a stationary value of Q, so there are values of W in the neighbourhood of their solution which give larger values of Q.

(ii) When $C(W) \equiv 0$, a solution of their equations by some suitably convergent iterative method always corresponds to a stationary value of Q, which may be either a maximum or a minimum.

(iii) When $C(W) \equiv 0$ and the particular iterative procedure proposed by Rudd and Blum is used to solve the equations, the solution corresponds to a stationary maximum value of Q whenever the iterations converge.

Case (ii) has already been illustrated by the example worked above. If the objective function in this example is modified to include a term $C(W)$, taking

$$Q = P_1 - P_f - (W - 2)^2 \tag{22}$$

in place of equation (7), the procedure of Rudd and Blum leads to the value $W = W^m = 2/5$. But with this value of W $(\partial Q/\partial W)_{P_f} = 4{\cdot}8$, showing that the solution does not represent a stationary value of Q and illustrating case (i). In case (iii) it might be thought that Rudd and Blum's procedure would lead to the correct answer, but a simple counterexample shows that this is erroneous. We consider a single stage system of the type shown in Fig. 1 with

$$P_1 = P_0^2 - (W - P_0)^2 \tag{23}$$

a mixing condition

$$P_0 = \tfrac{1}{2}P_f + \tfrac{1}{2}P_1 \tag{24}$$

and an objective function

$$Q = P_1 - P_f \tag{25}$$

In this case $W^m = P_0$ and equations (23) and (24) give

$$P_1 = P_0^2 \text{ and } P_0 = \tfrac{1}{2}P_1 \tag{26}$$

where we have confined attention to finding the solution for $P_f = 0$ to simplify the algebra. Equations (26) can be solved either directly or by the iterative method of Rudd and Blum, which converges in one step to the solution

$$P_0 = P_1 = 0 \text{ with } W^m = P_0 = 0 \tag{27}$$

These values might therefore be expected to correspond to the largest value of Q for $P_f = 0$. However, direct solution of equations (23), (24) and (25) for Q as a function of W shows that

$$Q = \frac{W^2}{W - 1} \text{ when } P_f = 0 \tag{28}$$

which has the form sketched in Fig. 2. The solution (27) corresponds to a stationary maximum of Q, as expected, but all values of $W > 1$ give larger values of Q, and indeed the largest value is obtained by allowing W to approach unity from above.

It must be concluded that the procedure proposed by Rudd and Blum can in no circumstances be counted upon to lead to the best operating policy, though it is doubtless possible to invent special examples in which it is successful.† (In particular the authors' worked example of cross current extraction with recycle has an objective function with finite terms $C(W)$, so it is unlikely that the result they give represents the true solution.) However, the principal purpose of this note is to draw attention to the fact that the method is based on a fallacious principle, namely the assumption that the process of determining optimum conditions in the recycle system is mathematically equivalent to sub-optimizing in the forward loop and balancing conditions at the recycle point. These two processes are not the same, and the assumption that they are can lead to completely misleading results even in cases which are completely "well behaved" mathematically.

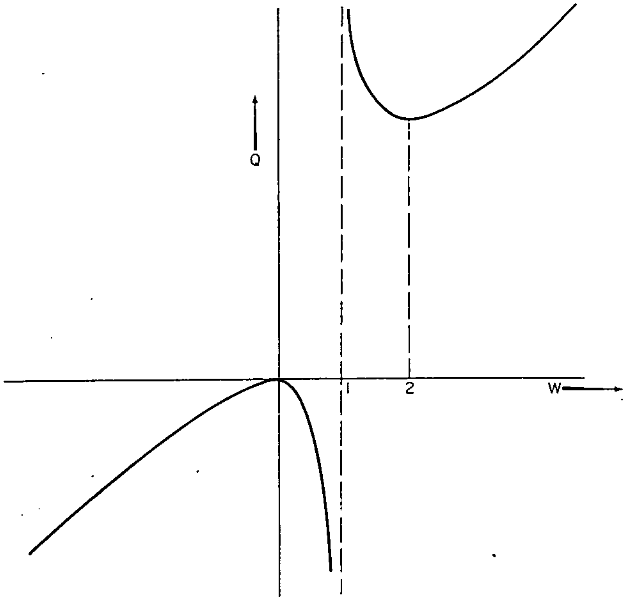


Fig. 2. Objective function for final counter example.

R. Jackson
University of Edinburgh
and Heriot-Watt College

Notation

C	Cost of applying an operating policy.
$\mathbf{P}_N$	Quality of product stream after N stages.
$\mathbf{P}_f$	Quality of feed stream.
$\mathbf{P}_0$	Quality of stream obtained by mixing feed and recycle streams.
q	Flow of feed and product streams.
Q	Objective function.
r	Flow of recycle stream.
R	Inciease in value of process stream.
W	Operating condition in a single stage process.
$\mathbf{W}_1 \ldots \mathbf{W}_N$	Operating conditions in a multi-stage process.
α	Equal to $q/(q + r)$.
β	Equal to $r/(q + r)$.

Reference

[1] Rudd, D. F. and Blum, E. D., *Chem. Engng. Sci.* 1962 **17** 277.

† The present discussion has been limited to the simplest case in which the P's are single numbers, i.e. one dimensional vectors. It has been pointed out to the writer by Dr. F. Horn that Rudd and Blum's procedure does not lead to a stationary value of the objective function, even when $C \equiv O$, if the P vectors have more than one component.

Chemical Engineering Science, 1963, Vol. 18, p. 218. Pergamon Press Ltd., Oxford. Printed in Great Britain.

Book Reviews

F. G. HELFFERICH, **Ion Exchange.** McGraw-Hill, New York 1962. 624 pp. $16.00.

A RECENT addition to the McGraw-Hill series in Advanced Chemistry, "Ion Exchange" by Dr. FRIEDRICH G. HELFFERICH belongs without doubt in the library of every serious student of this most interesting subject. Dr. HELFFERICH has written a comprehensive and penetrating treatise in which the emphasis has been placed on the fundamental physical chemical aspects of the subject. He has dealt essentially with the nature of ion exchangers and their behaviour and only incidentally with the techniques of their application. Thus, while this book may not necessarily answer the questions of those concerned with operating apparatus or processes, it does probe the underlying frame work on which such practical considerations ultimately rest. Working from a background of extensive research experience, Dr. HELFFERICH has written a book with a point of view. He has critically surveyed the rather staggering literature in this area and has not hesitated to point out those approaches which he considers most fruitful in explaining the nature of ion exchangers. The main emphasis in the text is on models which explain and account for the observed characteristics of ion exchange resins. Old and new approaches are considered and considerable qualitative discussion is presented, along with sufficient mathematical development to enable interested readers to follow the more detailed structure and consequences of such models. In addition to these theoretical considerations, the text also contains a brief but useful summary of the main methods used in experimental investigations of ion exchange and related phenomena.

The book is divided into twelve chapters each of which contains a generous number of references and a complete summary of the chapter contents. After a brief introduction there is a discussion of the structure and properties of ion exchangers, which is in turn followed by a consideration of the chemistry of their preparation. This largely qualitative section is followed in the succeeding chapters by extensive descriptive and quantitative discussion of the properties and behaviour of ion exchange materials. Chapters are devoted to exchanger capacity, equilibrium in ion exchange reactions, solvent sorption and resin swelling, kinetics of exchange and sorption, electrochemical properties of exchangers, and the theory and properties of exchanger membranes. In particular, the chapters on equilibria and membrane processes are quite comprehensive, and that on kinetics is perhaps the most complete and rigorous discussion of this subject available at present. There is a chapter devoted to ion exchange column behaviour which contains the usual quantitative relations as well as some illuminating general discussion. In addition, there are also interesting chapters devoted to the subject of ion exchange in non-aqueous and mixed solvents, catalysis by exchange resins and electron and redox exchangers and their properties. The text concludes with a useful appendix which contains a detailed table of nomenclature, a listing of all the commercial available ion exchangers and several tables of mathematical functions pertinent to solutions presented in the theoretical developments.

In the opinion of the reviewer, this volume will be of value to anyone interested in a fundamental understanding of the properties and behaviour of ion exchange materials. It will certainly serve as a most useful reference to the expert in this field, while selected readings from the various chapters will give the more general reader an excellent picture of ion exchange and its many ramifications.

J. S. DRANOFF

~~A pulsed column crystallizer was developed for *p*-xylene purification and showed appreciably higher capacity and efficiency when compared to conventional units (*6*). In addition the crystallizer was fairly simple and said to be easy to operate. This unit had solid and liquid phases moving countercurrently and had only a single stage.~~

~~The application of cycling to fluid bed reactions was reported to have two beneficial effects (*5*). The first was that the particle size range that could be handled was extended from a ratio of about 5 to 1 to the handling of everything from fine powders to coarse metal chips from machining operations. In addition, flow rates up to 10 times those available in steady flow systems were achieved. The pulses were timed so that the bed settled between pulses to the position it would assume during steady-state flow. A recent report presents laboratory data for the production of butadiene by dehydrogenation (*8*). By operating the reactor with reactant butenes pulsed into a diluent stream, conversion to butadiene far in excess of equilibrium can be achieved. Furthermore, acceptable conversions were achieved at temperatures where no measurable quantity of butadiene would be produced under conventional steady flow conditions.~~

~~The application of cycling to heat transfer was made to welding operations (*1*) and to film boiling (*3*). Welding is a particularly complicated heat transfer operation and as such makes an interesting study. The cycling nature of the operation is somewhat different from that described previously. Two power sources were used; one provided background current at a low enough level not to fuse the metal while a second pulsed unit provided current in shots with which the weld was made. The operation is especially suited to joining thin sheets of metal where the problems of maintaining an arc and not burning the metal at the same time are eliminated. Up to 100% increases in heat transfer rate were observed when cycling was applied to stable film boiling.~~

~~There is some information available which indicates that fuel cell performance can be improved by cycling (*4*). The improvement is primarily an increase in lifetime of the electrodes at high current densities.~~

~~Nomenclature~~

~~H = holdup~~
~~n = stage number~~
~~t = time~~
~~V = vapor flow rate~~
~~X = mole fraction in liquid~~
~~Y = mole fraction in vapor~~

~~Literature Cited~~

~~(1) British Welding Research Association, *Steel* **153,** No. 8, 52 (1964).~~
~~(2) Cannon, M. R., *Ind. Eng. Chem.* **53,** 629 (1961).~~
~~(3) DiCicco, D. A., Schoenhels, R. J., *J. Heat Transfer* **86,** 457 (1964).~~
~~(4) Interagency Advanced Power Group, Project Brief, Contract NAS3-2752, August 1963.~~
~~(5) Levey, R. P., Heidt, H. M., Hamrin, C. M., Jr., 53rd National Meeting, A.I.Ch.E., Pittsburgh, Pa., May 1964.~~
~~(6) Marwil, S. J., Kolner, S. J., *Chem. Eng. Progr.* **59,** No. 2, 60 (1963).~~
~~(7) Robertson, D. C., M. S. thesis, The Pennsylvania State University, University Park, Pa., 1957.~~
~~(8) Semenenko, E. I., Rojinskii, S. Z., *Kinetika i Katiliz* **5,** No. 3, 490 (1964).~~
~~(9) Szabo, T. T., Lloyd, W. A., Cannon, M. R., Speaker, S. S., *Chem. Eng. Progr.* **60,** No. 1, 66 (1964).~~
~~(10) Ziolkewski, Z., Filip, S., *Intern. Chem. Eng.* **3,** 433 (1963).~~

~~VERLE N. SCHRODT~~

~~*Monsanto Co.*~~
~~*St. Louis, Mo.*~~

~~RECEIVED for review September 23, 1964~~
~~ACCEPTED December 3, 1964~~

B2

CORRESPONDENCE

DISCRETE MAXIMUM PRINCIPLE

SIR: The maximum principle of Pontryagin (*11*) is now a well known method of dealing with a wide class of extremal problems associated with the solution of ordinary differential equations with given initial conditions. In a particularly lucid exposition of this principle, Rozonoer (*12*) has pointed out that, although one might hypothesize a discrete analog of the maximum principle for difference equations rather than differential equations, such a result is invalid except in certain very special conditions which render it almost trivial. Nevertheless, in three recent papers (*8,–10*), Katz has presented a proof of a discrete maximum principle around which a significant amount of work—some already published (*1,–3, 13, 14*) and some still in press—is beginning to build up. However, the purpose of this note is to reaffirm, largely by means of simple counterexamples, Rozonoer's original statement that the discrete maximum principle is invalid, and to show that the "proof" given of it is fallacious.

Firstly, we will briefly recapitulate Katz's results. He considers a system of difference equations of the form

$$x_i^n = F_i^n(x_k^{n-1}, \theta^n); \; (i = 1, 2, \ldots S); \; (n = 1, 2, \ldots N) \qquad (1)$$

with given initial conditions

$$x_i^0 = a_i; \; (i = 1, 2, \ldots S) \qquad (2)$$

It is customary and convenient to regard each Equation 1 as describing the relation between outputs and inputs for some physical unit, so that the complete set of equations describes the behavior of a sequential chain of units as shown in Figure 1. It is then required to find those values of the variables θ^1, θ^2, . . θ^N which minimize (maximize) the value of x_1^N.

The proposed solution makes use of the solutions z_i^n of a set of difference equations adjoined to Equations 1—namely

$$z_i^{n-1} = \sum_{j=1}^{S} \frac{\partial F_j^n}{\partial x_i^{n-1}} \cdot z_j^n; \; (i = 1, 2, \ldots S); \; (n = 1, 2, \ldots N) \qquad (3)$$

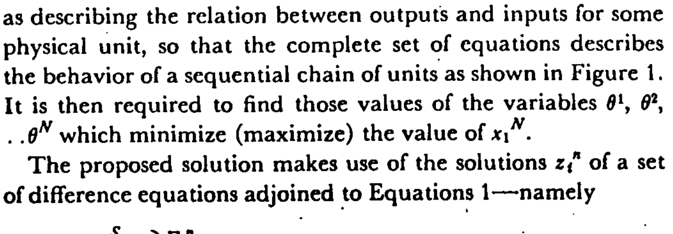


Figure 1. Sequential chain

with the terminal conditions

$$\left.\begin{aligned} z_i^N &= 1; \text{ for } i = 1 \\ &= 0 \text{ otherwise} \end{aligned}\right\} \tag{4}$$

It is then asserted that each θ^n should be chosen to minimize (maximize) the corresponding quantity

$$H^n = \sum_{j=1}^{S} z_j^n F_j^n \tag{5}$$

with the z_j^n regarded as constants from the point of view of the minimization (maximization) process.

In Katz's formulation of the problem, the functions F_j^n are assumed to be the same for each value of n and are written F_j. However, this restriction is not vital to the argument, and Fan and Wang (*2*, *3*) derive the same result without any such assumption. Indeed it has no bearing on the validity or otherwise of the result, as we shall show.

Some doubt may be thrown on the correctness of the above algorithm by the very simple example shown in Figure 2, where the objective is to maximize x^2. Direct elimination of x^1 shows that

$$x^2 = A - (x^0 + \theta^1)^2 - (\theta^2)^2$$

with a stationary maximum value at $\theta^1 = -x^0$, $\theta^2 = 0$, which also gives the largest value for any choice of θ^1 and θ^2.

However, F^1 is linear in the adjustable variable θ^1, so it is not possible to determine a value for θ^1 by the condition that

$$H^1 = \text{const. } x^1$$

should be maximized with respect to θ^1, as would be required by the computational procedure suggested by Katz (*8*).

Although this may cast some doubt on the result, it is easy to see that $z^1 = 0$ for the optimal policy, so that H^1 is actually independent of θ^1 and is, in a sense, maximized for any value of θ^1. To obtain a completely unambiguous counterexample, therefore, one must take $S = 2$, corresponding to a two-dimensional system. Consider, for example, the system shown in Figure 3, where the problem is to minimize x_1^2 with respect to θ^1 and θ^2. By direct elimination of x_1^1 and x_2^1, it is easily shown that

$$x_1^2 = 2 + {}^1/_2\,(\theta^1)^2 + (\theta^2)^2 \tag{6}$$

with a unique stationary minimum at $\theta^1 = \theta^2 = 0$, which also gives the smallest value of x_1^2. This is, therefore, the solution of the problem. We now pursue Katz's proposed procedure, solving the adjoined equations backward along the chain, starting from the boundary conditions.

$$z_1^2 = 1,\ z_2^2 = 0$$

According to Equation 3, we then have

$$z_1^1 = \frac{\partial F_1^2}{\partial x_1^1} \cdot z_1^2 + \frac{\partial F_2^2}{\partial x_1^1} \cdot z_2^2 = 1$$

and

$$z_2^1 = \frac{\partial F_1^2}{\partial x_2^1} \cdot z_1^2 + \frac{\partial F_2^2}{\partial x_2^1} \cdot z_2^2 = 2x_2^1$$

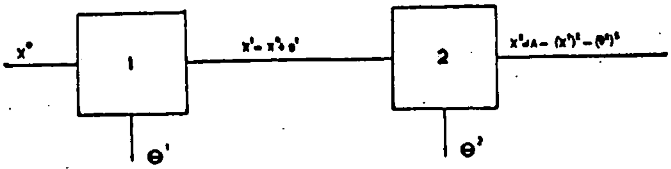


Figure 2. One-dimensional example

H^1 can then be written down by substituting into Equation 5, giving

$$H^1 = z_1^1\left[1 - 2\theta^1 - \frac{1}{2}\,(\theta^1)^2\right] + z_2^1(1 + \theta^1)$$

where

$$\frac{\partial H^1}{\partial \theta^1} = (z_2^1 - 2z_1^1) - z_1^1\theta^1 \text{ and } \frac{\partial^2 H^1}{\partial(\theta^1)^2} = -z_1^1 = -1$$

using the value of z_1^1 found above. It is seen that $\partial^2 H^1/\partial(\theta^1)^2$ is negative for all values of θ^1, so it follows that H^1 can never be minimized with respect to θ^1, as would be required by Katz's principle expressed in Equation 5. Indeed the values $\theta^1 = \theta^2 = 0$ and the corresponding solutions $x_1^1 = x_2^1 = 1$, $z_1^1 = 1$, $z_2^1 = 2$, which have been shown by direct calculation (Equation 6) to minimize x_1^2, actually maximize H^1 in direct contradiction of Katz's algorithm.

In this simple counterexample the functions F_i^n are different for different values of n. It remains to demonstrate the truth of the assertion made above that Katz's result remains invalid even if all the functions F_i^n are the same, as in his original derivation. We shall do this by showing that, from any S dimensional counterexample, it is possible to generate an $(S + 1)$ dimensional counterexample in which all the functions F_i^n are the same.

Consider then an example in S dimensions with state variables

$$x_i^n\ (i = 1, 2, \ldots S);\ (n = 1\ 2\ \ldots N)$$

and recurrence relations $x_i^n = F_i^n(x_k^{n-1}, \theta^n)$, with boundary conditions x_i^0 given, and suppose that this contradicts Katz's result in the same way as the example just quoted, and is therefore a counterexample. Let us call it Example I.

Now introduce a second example (Example II) with $S + 1$ dimensions and recurrence relations

$$x_i^n = G_i(x_k^{n-1}, \theta^n) \tag{7}$$

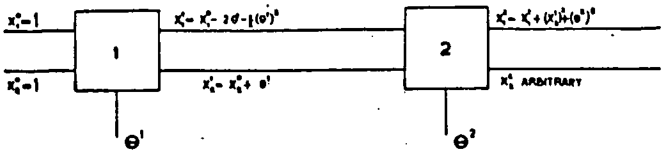
with the functions G_i independent of n. The functions G_i of any $S + 1$ variables, $\xi_1, \xi_2, \ldots \xi_{S+1}$ are defined by the following relations:

$$\left.\begin{aligned} G_i(\xi_1 \ldots \xi_{S+1}) &= \sum_{m=1}^{N} \phi_m(\xi_{S+1}) F_i^m(\xi_1, \ldots \xi_S, \theta);\ (i = 1 \ldots S) \\ G_{S+1}(\xi_1 \ldots \xi_{S+1}) &= \xi_{S+1} + 1 \end{aligned}\right\} \tag{8}$$

where the functions ϕ_m have the following properties

$$\left.\begin{array}{lll} \text{(i)} & \phi_m(x) = 1 & (x = \text{integer} = m) \\ \text{(ii)} & \phi_m(x) = 0 & (x = \text{integer} \neq m \text{ and } 1 \leqslant x \leqslant N) \\ \text{(iii)} & \phi_m(x) \text{ arbitrary for other values of } x. & \end{array}\right\} \tag{9}$$

There are many such sets of functions—for example,



$x_1^0 = 1$
$x_2^0 = 1$
1
$x_1^1 = x_1^0 - 2\theta - \frac{1}{2}(\theta^1)^2$
$x_2^1 = x_2^0 + \theta^1$
2
$x_1^2 = x_1^1 + (x_2^1)^2 + (\theta^2)^2$
x_2^2 ARBITRARY
θ^1
θ^2

Figure 3. Two-dimensional example

$$\phi_m(x) = \frac{(x-1)(x-2)\ldots(x-m+1)(x-m-1)\ldots(x-N)}{(m-1)(m-2)\ldots(m-m+1)(m-m-1)\ldots(m-N)}$$

The boundary conditions for the recurrence relations (Equation 7) are x_i^0, the same as those given for Example I

$$\left.\begin{array}{l}(i = 1, 2, \ldots S) \\ \text{and} \quad x_{S+1}^0 = 1\end{array}\right\} \qquad (10)$$

Putting $\xi_1 \ldots \xi_{S+1}$ equal to $x_1^{n-1}, \ldots x_{S+1}^{n-1}$ in Equations 8 and substituting into Equation 7, it follows on taking account of Equation 9 that only one term in the sum over m retains a nonzero value—namely, the term $m = n$—so we have

$$\left.\begin{array}{l}x_i^n \;\;= F_i^n(x_k^{n-1}, \theta^n); \;\; (i = 1, 2, \ldots S) \\ x_{S+1}^n = x_{S+1}^{n-1} + 1\end{array}\right\} \qquad (11)$$

Thus the variables $x_i^n (i = 1, \ldots S)$ take the same values in Example II as in Example I. It follows that identical values of $\theta^1, \theta^2, \ldots \theta^N$ in the two examples give identical values of x_1^N, so the same set of values of the θ's minimizes (maximizes) x_1^N in both cases. Similarly it can be shown that the adjoined variables z_i^n $(i = 1, 2, \ldots S)$ are identical in Examples I and II, so the function

$$H_{\text{II}}^n = \sum_{j=1}^{S+1} z_j^n G_j = \sum_{j=1}^{S} z_j^n F_j^n + z_{S+1}^n G_{S+1}$$

for Example II differs from the function

$$H_{\text{I}}^n = \sum_{j=1}^{S} z_j^n F_j^n$$

for Example I only by the term $z_{S+1}^n G_{S+1}$, which is independent of the adjustable variables $\theta^1, \theta^2, \ldots \theta^N$. Thus, if a set of values of $\theta^1, \ldots \theta^N$ which minimizes x_1^N also maximizes some H^n in Example I, thus contradicting Katz's result, the same will be true in Example II. Example II is therefore a counterexample if Example I is, and furthermore Example II has the same recurrence relations at all stages, thus proving our assertion.

Though Katz's discrete maximum principle is false, as has now been demonstrated, what he refers to as his "weaker algorithm" is true. This states merely that x_1^N takes a stationary value with respect to variations in $\theta^1, \theta^2, \ldots \theta^N$ if and only if each of the functions H^n takes a stationary value, and makes no comment on the relation between the natures of these stationary values. This result was earlier derived and used by the first of the present writers (*4*). The extension of Katz's results by Fan and Wang (*2, 3*) to systems topologically more complex than a simple sequential chain is also false, but once again a weaker algorithm relating stationary values is true, and was published by the second of the present writers (*6, 7*).

The fallacy in the proofs given by Katz and by Fan and Wang lies in the attempt to deduce the natures of stationary values from a consideration of first-order variations only, and the results obtained are simply consequences of a confusion in orders of magnitude of small quantities. The natures of stationary values of the objective function and of the functions H^n are determined, of course, by Hessian matrices of second derivatives with respect to the variables $\theta^1, \theta^2, \ldots \theta^N$. Indeed, there is no difficulty in writing down the Hessian for variations of x_1^N and hence deducing correctly the nature of a stationary value of x_1^N, as is shown in more detail in another publication (*5*), in which we also investigate certain special circumstances in which a stronger result is true. Very briefly, we may enumerate these cases here.

In order that x_1^N should take a stationary minimum (maximum) value with respect to variations in $\theta^1, \ldots \theta^N$, it is necessary and sufficient that each function H^n should take a stationary minimum (maximum) value with respect to the same variables in the following circumstances.

A. When the functions $F_i^n(x_k^{n-1}, \theta^n)$ take the special form

$$F_i^n(x_k^{n-1}, \theta^n) = \sum_k \alpha_{ik}^n x_k^{n-1} + f_i^n(\theta^n)$$

where the α_{ik}^n are constants. This is the case quoted by Rozonoer (*12*).

B. When $S = 1$, in other words when there is only one x-variable at each stage [though there are exceptions in this case, as is discussed elsewhere (*5*)]. The condition is also necessary, but not sufficient, if the functions F_i^n are linear in the variables x_k^{n-1}, but the coefficients may depend on the θ's.

These results refer to relations between local minima (maxima) of x_1^N and local minima (maxima) of the functions H^n. Pontryagin's principle is a more powerful result relating the absolute minimum (maximum) of x_1^N with absolute minima (maxima) of the H^n, and this is valid only in the case A above, as was asserted by Rozonoer.

Literature Cited

(1) Fan, L. T., "Optimization of Multistage Heat Exchanger System by the Discrete Maximum Principle," Kansas State Univ., Eng. Expt. Sta., Spec. Rept., 1964.
(2) Fan, L. T., Wang, C. S., *J. Electron. Control* **16**, 441 (1964).
(3) Fan, L. T., Wang, C. S., *J. Soc. Ind. Appl. Math.* **12**, 226 (1964).
(4) Horn, F., *Chem. Eng. Sci.* **15**, 176 (1961).
(5) Horn, F., Jackson, R., *J. Electron. Control* (in press).
(6) Jackson, R., *Chem. Eng. Sci.* **19**, 19 (1964).
(7) *Ibid.*, p. 253.
(8) Katz, S., IND. ENG. CHEM. FUNDAMENTALS **1**, 226 (1962).
(9) Katz, S., *J. Electron. Control* **13**, 179 (1962).
(10) *Ibid.*, **16**, 189 (1964).
(11) Pontryagin, L. S., *Uspekhi. Mat. Nauk* **14**, 3 (1959).
(12) Rozonoer, L. I., *Automation & Remote Control* **20**, 1288, 1405, 1517 (1959).
(13) Wang, C. S., Fan, L. T., IND. ENG. CHEM. FUNDAMENTALS **3**, 38 (1964).
(14) Zahradnik, R. L., Archer, D. H., *Ibid.*, **2**, 238 (1963).

F. Horn

Imperial College of Science & Technology
Prince Consort Rd.
South Kensington, London, England

R. Jackson

University of Edinburgh and Heriot-Watt College
Chambers St.
Edinburgh, Scotland

~~# CORRESPONDENCE~~

~~## A NORMALIZATION FOR THE THIELE MODULUS~~

~~SIR: Bischoff, who has proposed an identical normalization (1), has kindly pointed out an algebraic error in the above communication [IND. ENG. CHEM. FUND. 4, 227 (1965)]. It may be corrected by substituting the following for the first two and the seventh lines of the second column of page 228:~~

~~lines 1 and 2, $E = 1 - {}^2/_3\rho(1 - \zeta) + \ldots$~~

~~$= 1 - {}^1/_3\rho Q^2(1)\,\Lambda^2 + O(\Lambda^4)$~~

~~line 7, $\Lambda < \dfrac{1}{10\,Q(1)}\left\{\dfrac{3}{|\rho|}\right\}^{1/2}$~~

~~It may also be worth nothing that this inequality can be written as~~

~~$d < \{3D(1)/100|R'(1)|\}^{1/2}$~~

~~thus putting a bound on the pellet size in terms of the conditions at the surface.~~

~~### Literature Cited~~

~~(1) Bischoff, K. B., *A.I.Ch.E.J.* **2,** 351 (1965).~~

~~*University of Minnesota*~~
~~*Minneapolis, Minn.*~~

~~*Rutherford Aris*~~

B3

CORRESPONDENCE

DISCRETE MAXIMUM PRINCIPLE

SIR: We are grateful to Denn (2) for drawing attention to some other papers which may mislead and feel that his remarks on the Lagrange multiplier method are timely, since he draws attention to a simple error which is frequently seen in print.

It seems likely, as Denn suggests, that Rozonoer's remark has been widely misinterpreted by practitioners of engineering mathematics in both the Soviet and western literature. Nevertheless, we would find it difficult to believe that the mathematical originators of the maximum principle were not fully aware of the true situation. In this connection it is interesting to speculate why mathematicians have written so little about the discrete case, while the continuous maximum principle has received so much attention.

The reason, we think, is simple. The continuous maximum principle is a result of considerable intrinsic importance, relating absolute rather than local maxima in a way which goes beyond the earlier theorems of the calculus of variations; the discrete principle on the other hand, though useful, is a result of no mathematical interest whatever.

The essentially trivial nature of the discrete case tends to be hidden by some of the methods which have been used to treat it. Thus, Katz and others (4, 6, 7) have started from first principles and used a method analogous to that employed in treating the continuous case, while Denn himself (3) has presented an elegant method based on Green's identity. In fact, however, the discrete result is no more than a trivial rewriting of a well known formula of elementary differential calculus, as we will show briefly.

Given a set of sequential functional relations

$$x_i^n = F_i^n(F_k^{n-1}, \theta_r^n)\ (i = 1, 2, \ldots S), (n = 1, 2, \ldots N) \quad (1)$$

The problem is to choose the variables θ_r^n so that an objective function

$$P = c_i x_i^N$$

is maximized, with given values of the variables x_i^1. (In the above expression summation is implied over the repeated suffix, and this convention will be adhered to throughout.) Expressions for the partial derivatives $\partial P/\partial\theta_r^n$ can then be written down immediately using the chain rule of differentiation, which can be found in any textbook of elementary calculus (1).

$$\partial P/\partial\theta_r^n = c_i\partial x_i^N/\partial\theta_r^n = \left(c_i \frac{\partial F_i^N}{\partial x_j^{N-1}} \frac{\partial F_j^{N-1}}{\partial x_m^{N-2}} \cdots \frac{\partial F_l^{n+1}}{\partial x_k^n}\right) \frac{\partial F_k^n}{\partial\theta_r^n} \quad (3)$$

If we now define

$$z_k^n = \left(c_i \times \frac{\partial F_i^N}{\partial x_j^{N-1}} \times \frac{\partial F_j^{N-1}}{\partial x_m^{N-2}} \cdots \frac{\partial F_l^{n+1}}{\partial n_k^n}\right) \quad (4)$$

then clearly

$$z_i^{n-1} = \frac{\partial F_k^n}{\partial x_i^{n-1}} \times z_k^n \text{ with } z_i^N = c_i \quad (5)$$

Equation 3 can then be written

$$\partial P/\partial\theta_r^n = z_k^n\,\partial F_k^n/\partial\theta_r^n \quad (6)$$

and the condition that P should be stationary with respect to the θ_r^n is

$$z_k^n\,\partial F_k^n/\partial\theta_r^n = 0 \quad \text{(all } n, r) \quad (7)$$

We now see that the variables z_i^n are just the adjoint variables introduced in the "discrete maximum principle" and Equations 5 are the corresponding adjoint equations and boundary conditions, while Equation 7 is the condition that the functions $P^n = z_k{}^n F_k{}^n$ should be stationary if P is to be stationary. The change from the elementary formula (Equation 3) to the "discrete maximum principle" embodied in Equations 5 and 7 is purely notational and introduces nothing new. Even the algorithm for sequential computation of the derivatives suggested by Equation 6 is equally clearly indicated by the original formula (*4*).

A derivation similar to the above, and revealing the elementary nature of the result in the same way, has been given for systems of arbitrarily complex topology by one of the present writers (*5*).

Literature Cited

(1) Courant, R., "Differential and Integral Calculus," Vol. I, Blackie & Son, London, 1937.
(2) Denn, M. M., Ind. Eng. Chem. Fundamentals **4**, 240 (1965).
(3) Denn, M. M., Aris, R., *Ibid.*, p. 7
(4) Fan, L. T., Wang, C. S., *J. Electron. Control* **16**, 441 (1964).
(5) Jackson, R., *Chem. Eng. Sci.* **19**, 19 (1964).
(6) Katz, S., Ind. Eng. Chem. Fundamentals **1**, 226 (1962).
(7) Katz, S., *J. Electron. Control* **13**, 179 (1962).

University of Edinburgh and Heriot-Watt College **R. Jackson**
Edinburgh, Scotland

Rice University **F. Horn**
Houston, Tex.

CORRESPONDENCE

DYNAMIC PROGRAMMING AND LAGRANGIAN MULTIPLIERS

Sir: Bellman *et al.* (*1, 2*) have shown the usefulness of the *k*th best policy search technique. Roberts (*5*) has proposed modifying this procedure by incorporating a parameter to generalize this search. When the parameter is zero, Roberts' procedure becomes identical to Bellman's method.

This paper gives geometric interpretations of this procedure over discrete as well as continuous distributions. While the discrete distribution, is considered, failure of the Lagrangian multiplier technique observed by Roberts is explained.

Discrete Distribution

Consider a situation where all possible returns for the permissible values of the policy function are represented by a set of discrete point values. This set may be convex, nonconvex, or a combination of both. Roberts' problem (*5*) gives a discrete set which is partly convex. This characteristic makes it an interesting study.

Problem Statement. Extremize

$$X_1 = y_1^2 + y_2^2 + y_3^2 - 2y_3 - y_2 + 300 - 3S \tag{1}$$

Subject to

$$y_1 + y_2 + y_3 - 18 = 0 \tag{2}$$

$$y_1 = 2, 4, 6, 8, 10 \tag{3a}$$

$$y_2 = 2, 4, 6, 8, 10 \tag{3b}$$

$$y_3 = 2, 4, 6 \tag{3c}$$

A Geometric Analysis. Let Equations 1 and 2 be mapped from a subset of three-dimensional y_i space to a part of two-dimensional x_i space. Let $S = 0$.

Then

$$X_1 = y_1^2 + y_2^2 + y_3^2 - 2y_3 - y_2 + 300 \tag{4}$$

$$X_2 = y_1 + y_2 + y_3 - 18 \tag{5}$$

The discrete set obtained by using all the permissible values of y_1, y_2, and y_3 can be represented as shown in Figure 1.

It can now be shown that the *k*th best return search is tantamount to a sweep of the set by a straight line with slope λ.

Nonconvex Region. The search for the maximum is over a nonconvex region of the set. For the problem, as stated in Equations 4 and 5, the *k*th best return values reported by Roberts are obtained by the downward sweep of the set by straight lines with slopes λ = 0 and λ = 11, approaching the set from the top. Each encounter of the sweeping line with a point belonging to the set constitutes a best return with progressive rank (Figure 1).

Now, it can be shown that Roberts' criterion (*5*) for the choice of λ using the tabular values of returns for various λ as "a sieve to isolate the best λ to be employed in the *k*th best return calculation" has no general basis.

Consider two modifications of the problem where the constraint Equation 5 is modified as in Equation 6 or 7.

$$X_2 = y_1 + y_2 + y_3 - 24 \tag{6}$$

$$X_2 = y_1 + y_2 + y_3 - 8 \tag{7}$$

Then the values in Table I are obtained from Figures 1 to 3. These values clearly show that λ chosen by Roberts' criterion fares no better and even worse than some other λ. It is evident, then, that the most advantageous value of λ depends upon the configuration and the density of population of the set.

Convex Region. Now consider the minimization problem. This search is in a convex region and constitutes an upward sweep of the set.

There always will be one particular slope of the sweep which will give the optimal return with minimum rank, which is one, irrespective of the density and configuration of the set. This, however, does not apply to the ranks of the suboptimal returns.

Table I. *k*th Best Return

$$f_N(S) = \underset{y_N}{\text{Max}} \left[y_N^2 - S + 100 - \lambda y_N + f_{N-1}(S + y_N)\right]$$

Constraint	λ	$f_3(0)$	X_1	*Rank of Optimal Return*
Equation 5	0	430	430	15
	11ᵃ	232	430	5
Equation 6	0	498	498	2
	6ᵃ	354	498	2
Equation 7	12ᵃ	222	318	3
	13	214	318	2
	20	158	318	2

ᵃ *By Roberts' criterion.*

B4

Reprinted from
INTERNATIONAL JOURNAL OF CONTROL, Vol. 1, No. 4, p. 389, April 1965

On Discrete Analogues of Pontryagin's Maximum Principle†

By R. JACKSON
University of Edinburgh and Heriot-Watt College

and F. HORN
Imperial College of Science and Technology

[Received July 29, 1964]

ABSTRACT

A discrete form of Pontryagin's Maximum Principle recently proposed by a number of authors is shown to be fallacious and a corresponding correct but weaker result is derived. Certain classes of problem are identified for which the original stronger result is valid.

IN the last year or two a number of papers have appeared (Fan and Wang 1964 a, b, Katz 1962 a, b, 1964) in which discrete analogues of Pontryagin's Maximum Principle, applicable to difference equations rather than differential equations, have been derived. Briefly, the result obtained may be described as follows. Given a set of difference equations of the form

$$x_i^n = F_i^n(x_k^{n-1}, \theta^n) \quad (i=1,2,\dots S); \quad (n=1,2,\dots N) \tag{1}$$

with initial conditions $x_i^0 = a_i$, it is required to find the values of $\theta^1, \theta^2, \dots \theta^N$ which minimize x_1^N. To solve this problem one introduces a set of difference equations adjoint to (1) in the new variables z_i^n:

$$z^{n-1} = \sum_{j=1}^{S} \frac{\partial F_j^n}{\partial x_i^{n-1}} z_j^n \quad (i=1,2,\dots S); \quad (n=1,2,\dots N) \tag{2}$$

and imposes the terminal conditions:

$$\left.\begin{aligned} z_i^N &= 1 \text{ for } i=1 \\ &= 0 \text{ otherwise.} \end{aligned}\right\} \tag{3}$$

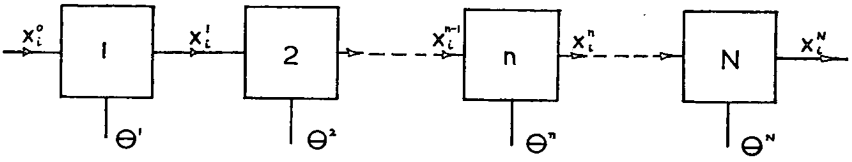
The solution is then said to be obtained by choosing each θ^n to minimize the corresponding quantity:

$$H^n = \sum_{j=1}^{S} z_j^n F_j^n \tag{4}$$

† Communicated by the Authors.

with the z_j^n regarded as constants in the minimization process. Equations (1) may conveniently be regarded as describing the behaviour of a sequential chain of physical units such as that shown in fig. 1, and corresponding results have also been given for systems of more complex topology (Fan and Wang 1964 a, b).

Fig. 1



Sequential chain.

Now the present writers have shown elsewhere (Horn and Jackson 1965) by means of counter-examples that, except in certain very special cases, the above result is fallacious. A weaker result is true (Horn 1961, Jackson 1964 a, b), namely that z_1^N has a stationary value with respect to variations in the θ's if and only if each function H^n is stationary with respect to variations in the corresponding θ^n, but in general the natures of the stationary values of x_1^N and the H^n are unrelated; in other words it is not generally true that H^n must be *minimized* in order to *minimize* x_1^N.

The fallacy in the proof referred to above is of a curiously elementary nature and arises from a confusion in orders of magnitude of small quantities, which apparently permits conclusions to be drawn about the nature of stationary values without considering terms beyond the first order in small variations. The nature of a stationary value is, of course, dependent on second-order terms and is determined by the Hessian matrix of second derivatives. This same fallacious proof has now been published at least four times (Fan and Wang 1964 a, b, Katz 1962 a, b) in different journals to the present writers' knowledge, so it is felt desirable to give a discussion of the problem properly based on second-order variations. Accordingly we will show that there is no difficulty in writing down the Hessian of x_1^N with respect to variations in the θ's and, at the same time, demonstrate that the solution of the adjoint difference eqns. (2) has a simple mathematical significance and can be written down as easily as the equations themselves.

Suppose $\theta^1, \theta^2, \ldots \theta^N$ are changed by increments $d\theta^1, d\theta^2, \ldots d\theta^N$. Then the corresponding variation dx_1^N may be expanded as a Taylor series in these increments. In general, each θ^n may represent a vector of adjustable parameters, and if we introduce suffixes r, s to distinguish components of these vectors, we may write:

$$\left.\begin{aligned} dx_1^N = \sum_{n=1}^{N} R_r^n \, d\theta_r^n + \tfrac{1}{2} \sum_n \sum_m P_{rs}^{nm} \, d\theta_r^n \, d\theta_s^m \\ + \text{terms of the third order,} \end{aligned}\right\} \qquad (5)$$

where, for brevity, the summation convention is assumed to apply to all repeated suffixes r, s, and $P_{rs}{}^{nm} = P_{sr}{}^{mn}$. The numbers $P_{rs}{}^{nm}$ are the elements of the Hessian matrix which determines the nature of a stationary value of $x_1{}^N$, while the numbers $R_r{}^n$ are the components of the gradient of $x_1{}^N$ in the space of the variables $\theta_r{}^n$.

Considering each eqn. (1) separately it is similarly possible to write down the following Taylor series expansions:

$$\left.\begin{aligned} dx_i{}^N &= (F_x{}^N)_j{}^i\,dx_j{}^{N-1} + (F_\theta{}^N)_r{}^i\,d\theta_r{}^N \\ &\quad + \tfrac{1}{2}[(F_{xx}{}^N)_{jk}{}^i\,dx_j{}^{N-1}\,dx_k{}^{N-1} + 2(F_{\theta x}{}^N)_{rj}{}^i\,d\theta_r{}^N\,dx_j{}^{N-1} \\ &\quad + (F_{\theta\theta}{}^N)_{rs}{}^i\,d\theta_r{}^N\,d\theta_s{}^N] + \text{terms of third order}, \\ dx_i{}^{N-1} &= (F_x{}^{N-1})_j{}^i\,dx_j{}^{N-2} + (F_\theta{}^{N-1})_r{}^i\,d\theta_r{}^{N-1} \\ &\quad + \tfrac{1}{2}[(F_{xx}{}^{N-1})_{jk}{}^i\,dx_j{}^{N-2}\,dx_k{}^{N-2} + 2(F_{\theta x}{}^{N-1})_{rj}{}^i\,d\theta_r{}^{N-1}\,dx_j{}^{N-2} \\ &\quad + (F_{\theta\theta}{}^{N-1})_{rs}{}^i\,d\theta_r{}^{N-1}\,d\theta_s{}^{N-1}] + \text{terms of third order}, \end{aligned}\right\} \quad (6)$$

where

$$\left.\begin{aligned} &(F_x{}^n)_j{}^i = \frac{\partial F_i{}^n}{\partial x_j{}^{n-1}}\ ;\ (F_\theta{}^n)_r{}^i = \frac{\partial F_i{}^n}{\partial \theta_r{}^n}\ ;\ (F_{xx}{}^n)_{jk}{}^i = \frac{\partial^2 F_i{}^n}{\partial x_j{}^{n-1}\,\partial x_k{}^{n-1}}, \\ &(F_{\theta x}{}^n)_{rj}{}^i = \frac{\partial^2 F_i{}^n}{\partial \theta_r{}^n\,\partial x_j{}^{n-1}}\ ;\ (F_{\theta\theta}{}^n)_{rs}{}^i = \frac{\partial^2 F_i{}^n}{\partial \theta_r{}^n\,\partial \theta_s{}^n} \end{aligned}\right\} \quad (7)$$

and the summation convention is implied with respect to repeated suffixes j, k, r and s.

By successive elimination of $dx_i{}^{N-1}$, $dx_i{}^{N-2}$, ... from eqns. (6) it is a straightforward matter to express $dx_1{}^N$ in terms of the variations $d\theta^n$ only. Comparison of terms of the first and second orders in the $d\theta$'s obtained in this way with the corresponding terms in eqn. (5) then gives expressions for the components of the gradient and the Hessian, namely:

$$R_r{}^n = (F_x{}^N)_j{}^1(F_x{}^{N-1})_k{}^j \ldots (F_x{}^{n+1})_m{}^l(F_\theta{}^n)_r{}^m \quad (8)$$

and

$$\begin{aligned} P_{rs}{}^{nn} &= (F_x{}^N)_j{}^1(F_x{}^{N-1})_k{}^j \ldots (F_x{}^{n+1})_m{}^l(F_{\theta\theta}{}^n)_{rs}{}^m \\ &\quad + (F_x{}^N)_j{}^1 \ldots (F_x{}^{n+2})_l{}^k(F_{xx}{}^{n+1})_{ab}{}^l(F_\theta{}^n)_r{}^a(F_\theta{}^n)_s{}^b \\ &\quad + (F_x{}^N)_j{}^1 \ldots (F_x{}^{n+3})_l{}^k(F_{xx}{}^{n+2})_{ab}{}^l(F_x{}^{n+1})_c{}^a(F_\theta{}^n)_r{}^c(F_x{}^{n+1})_d{}^b \\ &\quad \times (F_\theta{}^n)_s{}^d + \ldots + (F_{xx}{}^N)_{jk}{}^1(F_x{}^{N-1})_h{}^j \ldots (F_x{}^{n+1})_b{}^a(F_\theta{}^n)_r{}^b \\ &\quad \times (F_x{}^{N-1})_l{}^k \ldots (F_x{}^{n+1})_d{}^c(F_\theta{}^n)_s{}^d; \end{aligned} \quad (9)$$

while

$$
\begin{aligned}
&P_{rs}{}^{mn}=(F_x{}^N)_j{}^1\dots(F_x{}^{m+1})_l{}^k(F_{\theta x}{}^m)_{rh}{}^l(F_x{}^{m-1})_a{}^h\dots\\
&(n>m)\\
&\qquad\times(F_x{}^{n+1})_c{}^b(F_\theta{}^n)_s{}^c+(F_x{}^N)_j{}^1\dots(F_x{}^{m+2})_l{}^k\\
&\qquad\times(F_{xx}{}^{m+1})_{ab}{}^l(F_\theta{}^m)_r{}^a(F_x{}^m)_c{}^b\dots(F_x{}^{n+1})_e{}^d(F_\theta{}^n)_s\\
&\qquad+(F_x{}^N)_j{}^1\dots(F_x{}^{m+3})_l{}^k(F_{xx}{}^{m+2})_{ab}{}^l(F_x{}^{m+1})_e{}^a\\
&\qquad\times(F_\theta{}^m)_r{}^e(F_x{}^{m+1})_f{}^b\dots(F_x{}^{n+1})_h{}^g(F_\theta{}^n)_s{}^h\\
&\qquad+\dots+\\
&\qquad+(F_{xx}{}^N)_{jk}{}^1(F_x{}^{N-1})_h{}^j\dots(F_x{}^{m+1})_f{}^g(F_\theta{}^m)_r{}^f(F_x{}^{N-1})_a{}^k\dots\\
&\qquad\times(F_x{}^{n+1})_c{}^b(F_\theta{}^n)_s{}^c. \qquad (10)
\end{aligned}
$$

In the right-hand sides of these equations, as before, summation is implied over all repeated suffixes except n.

Now eqn. (8) describes the propagation of first derivatives through the sequential chain of units, while from eqns. (2) and (3) it is not difficult to see that

$$z_i{}^n=(F_x{}^N)_a{}^1(F_x{}^{N-1})_b{}^a\dots(F_x{}^{n+1})_i{}^k, \qquad (11)$$

so comparing eqns. (8) and (11):

$$R_r{}^n=z_i{}^n(F_\theta{}^n)_r{}^i, \qquad (12)$$

where suffix i is summed over, according to our convention. If $x_1{}^N$ is to be stationary, each of the first derivatives $R_r{}^n$ must vanish, so from (12):

$$\sum_i z_i{}^n\frac{\partial F_i{}^n}{\partial\theta_r{}^n}=0 \quad (n=1,2,\dots N) \qquad (13)$$

where we have re-introduced the summation sign explicitly to facilitate comparison with eqn. (4). Now (13) is simply the condition that each of the functions H^n introduced in eqn. (4) should be stationary with respect to θ^n, so we have proved the weaker result relating only to stationary values which was stated above. The adjoint variables $z_i{}^n$ are seen to be very simply related to the first derivatives of the quantity to be maximized with respect to the adjustable variables.

The nature of a stationary value of $x_1{}^N$ is determined by the matrix $P_{rs}{}^{mn}$ of second derivatives given by eqns. (9) and (10) and it is not, in general, related to the natures of the stationary values of the functions H^n, which are determined by matrices:

$$Q_{rs}{}^n=\sum_j z_j{}^n(F_{\theta\theta}{}^n)_{rs}{}^j. \qquad (14)$$

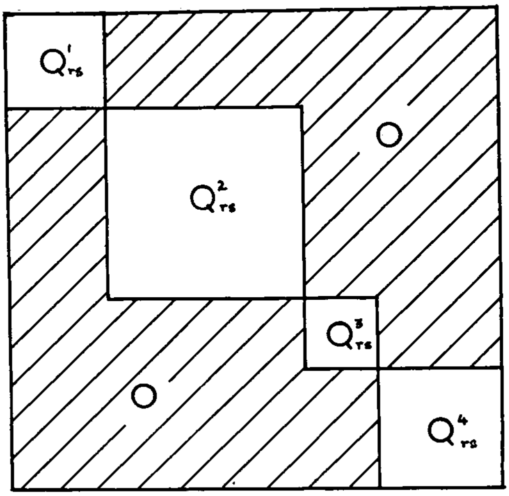
Using eqn. (11), the matrices (14) may alternatively be written:

$$Q_{rs}{}^n=(F_x{}^N)_j{}^1(F_x{}^{N-1})_k{}^j\dots(F_x{}^{n+1})_m{}^l(F_{\theta\theta}{}^n)_{rs}{}^m \qquad (15)$$

where the summation convention is once again introduced for repeated suffixes.

It is now possible to discern the circumstances under which the 'strong' result of Katz will be true, since the first term on the right of eqn. (9) is identical with the right-hand side of eqn. (15). Thus, if all other terms on the right-hand sides of eqns. (9) and (10) vanished, the matrix P_{rs}^{mn} would reduce to the block diagonal form indicated in fig. 2 with the matrices Q_{rs}^{n} arranged along its principal diagonal and all other elements vanishing. In these circumstances the conditions on P_{rs}^{mn} if x_1^N is to be a minimum are satisfied if and only if each of the matrices Q_{rs}^{n} satisfies the corresponding conditions for H^n to be a minimum. Thus x_1^N is minimized if and only if each function H^n is minimized with respect to the θ's.

Fig. 2



Block diagonal form of P_{rs}^{nm}.

We have therefore simply to identify any general situations in which P_{rs}^{mn} reduces to the block diagonal form of fig. 2, and these will be the situations in which Katz's strong result is true.

Two important cases can easily be recognized. Firstly, there is the situation correctly identified by Rozonoer in which the recurrence relations take the form:

$$x_i^n = \sum_k \alpha_{ik}^n x_k^{n-1} + f_i^n(\theta^n), \tag{16}$$

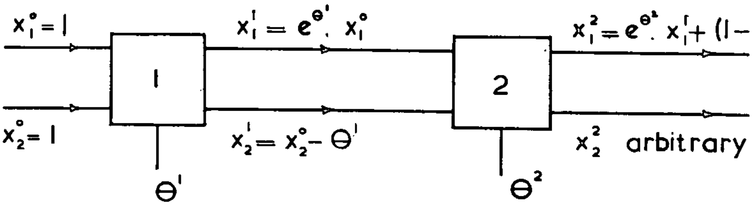
where the α_{ik}^n are constants. All second derivatives with respect to x-variables or an x-variable and a θ-variable then vanish and it is immediately seen from eqns. (9) and (10) that the Hessian reduces identically to the desired form.

Secondly, the strong result is true, with certain exceptions, when $S=1$ and only a single x-variable is associated with each unit in fig. 1. Each quantity F_x^n appearing on the right-hand side of eqn. (8) is then a single number rather than a matrix and, provided none of these numbers vanishes, R_r^n will vanish if and only if each derivative F_θ^n vanishes. Thus all derivatives F_θ^n vanish at a stationary value of x_1^N, and since the first term on the right-hand side of eqn. (9) is the only term in P_{rs}^{mn} which does not contain one of these derivatives as a factor, it is the only surviving non-zero term, and the matrix reduces to the form shown in fig. 2. In this case, however, the matrix reduces to this form only at stationary points and not identically as in Rozonoer's example.

Exceptions may arise in this one dimensional case, since one of the derivatives R_r^n may vanish as a result of a factor F_x^m on the right-hand side of eqn. (8) vanishing. It then no longer follows that F_θ^n must vanish and consequently the Hessian need no longer reduce to the form shown in fig. 2. An example with this property has been given elsewhere (Horn and Jackson 1965).

In each of the two situations just described, minimization of each H^n is both necessary and sufficient condition for x_1^N to be a minimum, but we should finally mention a case in which the condition is necessary but not sufficient. If the quantities α_{ik}^n appearing on the right-hand side of eqn. (16) are functions of θ^n rather than constants, the second derivatives with respect to x-variables still vanish, but the mixed derivatives with respect to an x-variable and a θ-variable do not. The elements P_{rs}^{nn} are then seen from eqn. (9) to reduce to the form (15), but the elements P_{rs}^{mn} $(m \neq n)$ no longer vanish. Miminization of the functions H^n is then necessary if P_{rs}^{mn} is to satisfy the appropriate conditions for x_1^N to be a minimum but a simple example, such as that given in fig. 3, suffices to show that this condition is no longer sufficient. In this example H^1 and H^2 both have minima at $\theta^1 = \theta^2 = 0$, but x_1^2 has a saddle point.

Fig. 3


$x_1^0 = 1$
$x_2^0 = 1$
1
Θ^1
$x_1^1 = e^{\Theta^1} . x_1^0$
$x_2^1 = x_2^0 - \Theta^1$
2
Θ^2
$x_1^2 = e^{\Theta^2} . x_1^1 + (1 - \Theta^2) x_2^1$
x_2^2 arbitrary

Example in which Katz's conditions are necessary but not efficient.

Although the introduction of the adjoint variables and the functions H^n gives an elegant formulation of the solution and a useful iterative algorithm for computations, the treatment above shows that one has actually accomplished no more than can be obtained using elementary calculus and straightforward elimination of unwanted variables. Thus eqn. (8) contains the same information as the adjoint equations, and indeed gives their solution. Using the elementary approach of elimination of variables we have extended the discussion to second derivatives, and it is interesting to see that these could also be developed through the adjoint equations.

We have shown (eqn. (12)) that

$$R_r^n = \frac{\partial x_1^N}{\partial \theta_r^n} = z_j^n \frac{\partial F_j^n}{\partial \theta_r^n} \tag{17}$$

with summation implied over the lower suffixes. Since this equation is true for any values of the θ's, we may differentiate both sides with respect to θ_s^m. If $m < n$, F_j^n depends on θ_s^m through its dependence on x_k^{n-1} but if $m \geqslant n$ we may write:

$$\frac{\partial^2 x_1^N}{\partial \theta_r^n \partial \theta_s^m} = \frac{\partial z_j^n}{\partial \theta_s^m} \cdot \frac{\partial F_j^n}{\partial \theta_r^n} + z_j^n \frac{\partial^2 F_j^n}{\partial \theta_r^n \partial \theta_s^m}. \tag{18}$$

The second term is, of course, zero if $m \neq n$, while the first term is neglected in Katz's treatment. In order to calculate these derivatives, we need to be able to compute derivatives such as $\partial z_j^n / \partial \theta_s^m$ which appear on the right-hand sides of eqns. (18). A recurrence relation for these derivatives can be obtained by differentiating the adjoint equation:

$$z_i^{l-1} = \frac{\partial F_j^l}{\partial x_i^{l-1}} . z_j^l$$

with respect to θ_s^m, giving:

$$\frac{\partial z_i^{l-1}}{\partial \theta_s^m} = \frac{\partial F_j^l}{\partial x_i^{l-1}} \cdot \frac{\partial z_j^l}{\partial \theta_s^m} + \left(\frac{\partial^2 F_j^l}{\partial x_i^{l-1} \partial x_k^{l-1}} \cdot \frac{\partial x_k^{l-1}}{\partial \theta_s^m} + \frac{\partial^2 F_j^l}{\partial x_i^{l-1} \partial \theta_s^m} \right) z_j^l. \tag{19}$$

The derivatives $\partial x_k^{l-1} / \partial \theta_s^m$ appearing on the right-hand side of (19) may be obtained from equations analogous to (17), and (19) then permits the derivatives to be computed successively for decreasing values of 1. Equations (17), (18) and (19) lead to the results given earlier in eqns. (9) and (10).

References

Fan, L. T., and Wang, C. S., 1964 a, *J. Electron. Contr.*, **16,** 441; 1964 b, *J. Soc. indust. appl. Math.*, **12,** 226.

Horn, F., 1961, *Chem. Engng Sci.*, **15,** 176.

Horn, F., and Jackson, R., 1965, *Ind. Eng. Chem.* (*Fundamentals*) (to be published).

Jackson, R., 1964 a, *Chem. Engng Sci.*, **19,** 19 ; 1964 b, *Ibid.*, **19,** 253.

Katz, S., 1962 a, *Ind. Eng. Chem.* (*Fundamentals*), **1,** 226 ; 1962 b, *J. Electron. Contr.*, **13,** 179; 1964, *Ibid.*, **16,** 189.

Rozonoer, L. I., 1959, *Automation and Remote Control,* **20,** 1288, 1405, 1517.

Chemical Engineering Science, 1964, Vol. 19, pp. 19-31. Pergamon Press Ltd., Oxford. Printed in Great Britain.

Some algebraic properties of optimization problems in complex chemical plants

R. JACKSON

University of Edinburgh and Heriot-Watt College

(*Received* 3 *May* 1963)

Abstract—The determination of optimum conditions in a chemical plant comprising a number of interconnected units often presents considerable computational difficulties because of the large number of parameters which must be simultaneously varied. The method of dynamic programming permits the problem to be decomposed into a set of sub-problems of lower dimensionality, but is limited in application to systems consisting of simple sequential chains of units. The present work describes a classical variational approach which permits a similar dimensional decomposition to be effected in plants of arbitrarily complex structure. A number of systems which exemplify the main features of the method without undue algebraic complexity are discussed in detail.

INTRODUCTION

THE problem of choosing the available design and operating variables of a chemical plant in such a way as to optimize some specified performance criterion presents considerable computational difficulties. One reason for this is that the number of available variables is frequently large and correspondingly one is seeking a maximum (or minimum) value of a function in a space with a large number of dimensions. In cases where the plant has the very simple configuration of a set of units connected head to tail in a sequence, methods are available for decomposing the problem into a set of sub-problems with the dimensionalities appropriate to the separate units. These methods fall into two main classes, the first based on the algorithm of dynamic programming and the second based on classical variational calculus. The method of dynamic programming was developed and is fully described by BELLMAN [1]. It has subsequently been developed and applied to a very large number of problems by BELLMAN, his co-workers and others and is admirably translated into chemical engineering terms in the work of ARIS [2]. The variational methods have had a more conventional scientific history, having been developed more or less independently by a number of different workers. The continuous case, which arises, for example, in considering optimum temperature gradients in reactors, was treated independently by PONTRYAGIN and co-workers at the University of Moscow [3] and by SWINNERTON-DYER [4] in England, both arriving at the same method, which is now usually referred to as the Maximum Principle of Pontryagin. More recently HORN [5] has given a classical treatment of a discrete sequential problem based on Lagrangian multipliers which is the analogue, for the discrete case, of the Pontryagin Principle.

The work so far described is limited in scope to the treatment of simple sequences of units, while most chemical plants of realistic complexity are in the nature of interconnected networks of units involving by-pass streams, recycle streams, etc. The author is aware of only two attempts to generalize the dynamic programming procedure to handle these more complex cases [6, 7], both fallacious for reasons which have been stated elsewhere [8] and will be discussed further below. The purpose of the present paper is to provide a means of treating cases of arbitrarily complex topology or, to be more precise, a method of decomposing the over-all optimization problem into sub-problems with the dimensionalities of the individual units. The approach is a classical variational one and is therefore properly regarded as an extension of the methods of PONTRYAGIN, SWINNERTON-DYER and HORN to non-sequential problems rather than a generalization of the method of dynamic programming. It is felt worthwhile to describe the method in reasonable generality, so the present paper is necessarily rather abstract. However, it is hoped to illustrate it in more concrete form by working particular chemical engineering problems in subsequent publications.

THE ALGEBRAIC STRUCTURE OF OPTIMIZATION PROBLEMS

In general a chemical plant consists of a number of units, each with a set of inputs and outputs and a number of adjustable parameters. The inputs are process streams flowing into the unit, the outputs similar streams flowing out of it, while the adjustable parameters are variables which characterize its design and conditions of operation. Each unit will be identified by a number (n), and its inputs, outputs and adjustable parameters will be denoted by $\mathbf{y}_j^{(n)}$, $\mathbf{x}_i^{(n)}$ and $\mathbf{w}^{(n)}$ respectively. Each of these is to be regarded as a set of several quantities forming a vector and it must be emphasized that the suffixes i and j serve to distinguish between separate process streams for units with multiple inputs and outputs; they do not indicate components of a vector and any explicit reference to these would require a further suffix. When a unit has but a single input and output the suffixes i and j will normally be omitted. The numbers of components of vectors $\mathbf{y}_j^{(n)}$, $\mathbf{x}_i^{(n)}$ and $\mathbf{w}^{(n)}$ are not necessarily the same and they will be denoted by $E_j^{(n)}$, $D_i^{(n)}$ and $W^{(n)}$ respectively.

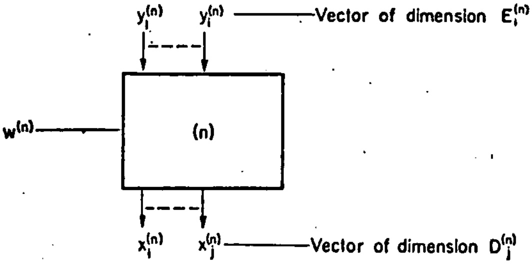

$y_l^{(n)}$
$y_i^{(n)}$
Vector of dimension $E_i^{(n)}$
$w^{(n)}$
(n)
$x_i^{(n)}$
$x_j^{(n)}$
Vector of dimension $D_j^{(n)}$

FIG. 1. Plant unit.

The definitions given above are illustrated in Fig. 1, which represents a single unit of the plant. Evidently a unit need not correspond to a particular physically distinct piece of equipment, but may comprise the contents of any control surface drawn within the plant in such a way as to intersect only lines carrying process streams. Thus a unit in the sense we employ the term, may have an internal structure and contain within itself a number of identifiable sub-units.

Each output is uniquely determined by the values of the inputs and adjustable parameters of the unit, and accordingly we shall write

$$\mathbf{x}_i^{(n)} = F_i^{(n)}[\mathbf{y}_j^{(n)}, \mathbf{w}^{(n)}] \qquad (1)$$

to indicate the functional relation between $\mathbf{x}_i^{(n)}$ and these variables. Of course, the symbol $F_i^{(n)}$ represents a set of $D_i^{(n)}$ functions, one for each component of $\mathbf{x}_i^{(n)}$, so the total number of equations of the form (1) associated with the plant is

$$v_1 = \sum_n \sum_i D_i^{(n)} \qquad (2)$$

The complete plant is formed by joining the outputs of one unit to the inputs of others to form a connected structure as typified by the example shown in Fig. 2. However, some of the inputs are left free of connexion to other units. These represent feeds to the plant and the corresponding vectors $\mathbf{y}_j^{(n)}$ take specified values. Similarly there are a number of free outputs, not connected to the inputs of other units, and these represent products of the plant. The complete topological structure can be specified by listing the variables with which each input is connected, thus obtaining a set of equations

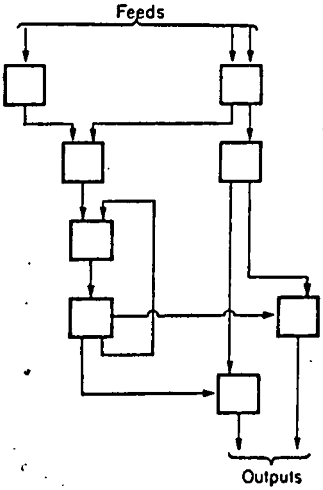


FIG. 2. Complete plant.

$$y_j^{(n)} = x_l^{(m)} \tag{3}$$

$$y_k^l = \bar{y}_k^{(l)} \tag{4}$$

vhere (3) refer to inputs connected to the outputs of other units while (4) refer to plant feeds, the $\bar{y}_k^{(l)}$ eing specified vectors. The total number of equaions (3) and (4) is

$\sum_n \sum_j E_j^{(n)}$, divided into

$$v_2 = \sum_n \sum_j E_j^{(n)} - G \tag{5}$$

quations of the form (3) and

$$v_3 = G \tag{6}$$

f the form (4), where G is the number of external eed stream variables.

The unit equations (1), together with the topoogical equations (3) and (4), provide a formally omplete mathematical description of the system.

The problem to be considered is that of finding alues for the adjustable parameters of the units hich maximize (or minimize) a function of the orm

$$P = \mathbf{c}\mathbf{x}_1^{(1)} - \sum_n G^{(n)}[\mathbf{w}^{(n)}] \tag{7}$$

here $\mathbf{c}$ is a constant row vector of $D_1^{(1)}$ compoents and the $G^{(n)}$ are given scalar functions of the ector arguments $\mathbf{w}^{(n)}$. Typically the first term epresents an income from sale of a product while e remaining terms represent the capital and runing charges associated with the chosen values of e adjustable parameters. The form of P given equation (7) can be generalized to include comonents of feeds and of outputs other than $\mathbf{x}_1^{(1)}$ in e first term, and linear combinations such as $\mathbf{c}\mathbf{x}_1^{(1)}$ ay be replaced by scalar functions of a general ature. However, the complications introduced are rely notational and contribute no genuine gain generality, so the simple form (7) will be retained re.

The most direct approach to the problem is to lve equations (1), (3) and (4) for $\mathbf{x}_1^{(1)}$ in terms of the $^{(n)}$ and the specified feeds $\bar{\mathbf{y}}_k^{(l)}$ (it is easily checked at (1), (3) and (4) provide sufficient equations for is purpose), substitute this value of $\mathbf{x}_1^{(1)}$ in uation (7), and seek the maximum of the resulting function of the variables $\mathbf{w}^{(n)}$. The number of these variables is

$$v_4 = \sum_n W^{(n)} \tag{8}$$

so this involves the simultaneous variation of v_4 variables in seeking the maximum, and this is a problem of daunting proportions if v_4 is at all large. The approach used here will be the classical one of seeking a stationary value of P with respect to small independent variations in the components of the vectors $\mathbf{w}^{(n)}$, making the assumption that the maximum value of P can be identified with such a stationary value. There are two difficulties associated with this approach: firstly P may take its greatest value at the boundary of the permitted region rather than a stationary point if the parameters $\mathbf{w}^{(n)}$ are constrained, and secondly P may have a number of stationary points of various types, only one of which corresponds to the greatest value of P. The first of these is a difficulty of practice rather than principle, since in principle constraints can always be "smoothed off" so that the maximum value of P is actually a stationary point. The second is a genuine difficulty of principle and can only be circumvented by examining each stationary value, when there is more than one, and finding which is largest.

According to equation (7) the small variation in P accompanying variations in the adjustable parameters is given by

$$dP = \mathbf{c}d\mathbf{x}_1^{(1)} - \sum_n \mathbf{g}^{(n)}d\mathbf{w}^{(n)} \tag{9}$$

where $\mathbf{g}^{(n)}$ is the row vector of $W^{(n)}$ components obtained by differentiating $G^{(n)}$ with respect to the components of $\mathbf{w}^{(n)}$. Similarly differentiation of equations (1), (3) and (4) gives

$$d\mathbf{x}_i^{(n)} = \mathbf{M}_{ij}^{(n)}d\mathbf{y}_j^{(n)} + \mathbf{N}_i^{(n)}d\mathbf{w}^{(n)} \tag{10}$$

$$d\mathbf{y}_j^{(n)} = d\mathbf{x}_l^{(m)} \tag{11}$$

and

$$d\mathbf{y}_k^{(l)} = 0 \tag{12}$$

where $\mathbf{M}_{ij}^{(n)}$ and $\mathbf{N}_i^{(n)}$ are the following matrices of partial derivatives

$$\{\mathbf{M}_{ij}^{(n)}[\mathbf{y}_k^{(n)}, \mathbf{w}^{(n)}]\}_{pq} = \frac{\partial[F_i^{(n)}]_p}{\partial[\mathbf{y}_j^{(n)}]_q} \tag{13}$$

$$\{\mathbf{N}_i^{(n)}[\mathbf{y}_k^{(n)}, \mathbf{w}^{(n)}]\}_{pr} = \frac{\partial[F_i^{(n)}]_p}{\partial[\mathbf{w}^{(n)}]_r} \tag{14}$$

with suffixes p, q and r referring to separate components of the vectors $\mathbf{F}_i^{(n)}$, $\mathbf{y}_j^{(n)}$ and $\mathbf{w}^{(n)}$ respectively. Thus $\mathbf{M}_{ij}^{(n)}$ is a matrix of $D_i^{(n)}$ rows and $E_j^{(n)}$ columns, while $\mathbf{N}_i^{(n)}$ is a matrix of $D_i^{(n)}$ rows and $W^{(n)}$ columns, each element of both matrices being a function of the variables $(\mathbf{y}_k^{(n)}, \mathbf{w}^{(n)})$ as indicated.

Equations (10), (11) and (12) are a set of $(\nu_1 + \nu_2 + \nu_3)$ simultaneous linear equations in the $(\nu_1 + \nu_2 + \nu_3)$ variables $(d\mathbf{x}_i^{(n)}, d\mathbf{y}_j^{(m)})$ and may be solved for any one of these variables in terms of the $d\mathbf{w}^{(n)}$. In particular, we may write

$$d\mathbf{x}_1^{(1)} = \sum_n \sum_i \mathbf{L}_{1i}^{(1)(n)} \mathbf{N}_i^{(n)} d\mathbf{w}^{(n)} \tag{15}$$

where $\mathbf{L}_{1i}^{(1)(n)}$ is a matrix of $D_1^{(1)}$ rows and $D_i^{(n)}$ columns which is a rational function of the matrices $\mathbf{M}_{ij}^{(n)}$ of a form determined by the topological structure of the plant. Equation (15) is simply the result of eliminating all the variables $d\mathbf{x}_i^{(n)}$ and $d\mathbf{y}_j^{(m)}$ except $d\mathbf{x}_1^{(1)}$ from equations (10), (11) and (12). The value of $d\mathbf{x}_1^{(1)}$ given by equation (15) may now be substituted into equation (9), which becomes

$$dP = \sum_n \left\{ \sum_i \mathbf{c}\mathbf{L}_{1i}^{(1)(n)} \mathbf{N}_i^{(n)} - \mathbf{g}^{(n)} \right\} d\mathbf{w}^{(n)}$$

The necessary and sufficient conditions for a stationary value of P are therefore

$$\sum_i \{\mathbf{c}\mathbf{L}_{1i}^{(1)(n)} \mathbf{N}_i^{(n)} - \mathbf{g}^{(n)}\} = 0 \quad (n = 1, 2, \ldots) \tag{16}$$

The direct approach to finding this stationary value is now to solve equations (1), (3) and (4) for the $\mathbf{x}_i^{(n)}$ and $\mathbf{y}_j^{(m)}$ in terms of the adjustable parameters $\mathbf{w}^{(n)}$, use these to express the matrices $\mathbf{M}_{ij}^{(n)}$ and $\mathbf{N}_i^{(n)}$ and hence the matrices $\mathbf{L}_{1i}^{(1)(n)}$, as functions of the parameters $\mathbf{w}^{(n)}$, then solve equations (16) as a set of ν_4 simultaneous equations in the ν_4 components of the vectors $\mathbf{w}^{(n)}$. This is the precise analogue of the direct approach to maximization described earlier, and once again involves the simultaneous variation of ν_4 variables, in this case to satisfy equations (16).

In dealing with sequential optimization problems, the method of dynamic programming permits the problem to be split into a number of separate problems of smaller dimensionality, which greatly facilitates the computation involved in the solution. In particular, if the plant units are connected in a simple sequence, the problem of simultaneous maximization in the $\sum_n W^{(n)}$ parameters can be reduced to a maximization in the $W^{(1)}$ parameters associated with the first unit, together with a maximization in the $W^{(2)}$ parameters associated with the second unit, and so on. This simplification is not obtained without cost, since a particular problem of interest can be solved only by imbedding it in a larger set of problems which must be solved at the same time. Attempts [6, 7] have been made to extend the method of dynamic programming to deal with structures more complicated than sequential chains but, as has been discussed elsewhere [8], these rest on a mathematical misconception and their results are not usually related in any way to the true solution. It will now be shown how the stationary value problem already set out can be split up into a set of problems of lower dimensionality by adopting an approach less direct than that employed hitherto.

In equations (16) it will be recalled that the matrices $\mathbf{L}_{1i}^{(1)(n)}$ are functions of the variables $\mathbf{y}_k^{(m)}$ and $\mathbf{w}^{(n)}$ associated with all the units of the plant. This arises from the functional dependence of the $\mathbf{L}_{1i}^{(1)(n)}$ on the matrices $\mathbf{M}_{ij}^{(n)}$. However, the matrix $\mathbf{N}_i^{(n)}$ is a function only of the variables $\mathbf{y}_j^{(n)}$ and $\mathbf{w}^{(n)}$ associated with the nth unit. Suppose now that we arbitrarily assume values for the components of all the vectors $\mathbf{c}\mathbf{L}_{1i}^{(1)(n)}$ and $\mathbf{y}_j^{(n)}$ associated with the plant, excepting those $\mathbf{y}$'s where values are externally specified through equation (4). The total number of variables whose values are assumed is

$$\sum_n \sum_i D_i^{(n)} + \sum_m \sum_j E_j^{(m)} - G$$

or $(\nu_1 + \nu_2)$. Using these assumed values, the left-hand side of the nth equation (16) depends only on the vector $\mathbf{w}^{(n)}$ with the same value of n, so equations (16) decompose into $W^{(1)}$ equations for the components of the vector $\mathbf{w}^{(1)}$, $W^{(2)}$ equations for the components of the vector $\mathbf{w}^{(2)}$, and so on. Once these equations have been solved, the values of the vectors $\mathbf{w}^{(n)}$ so determined can be used in equations (1), (3) and (4), which can then be solved for all the variables $\mathbf{x}_i^{(n)}$ and $\mathbf{y}_j^{(m)}$. These in turn determine the matrices $\mathbf{M}_{ij}^{(n)}$, and hence the vectors $\mathbf{cL}_{1i}^{(1)(n)}$. Thus we have arrived at calculated values of the $\sum_n \sum_i D_i^{(n)}$ vectors $\mathbf{cL}_{1i}^{(1)(n)}$ and the $\sum_m \sum_j E_j^{(m)} - G$ unspecified vectors $\mathbf{y}_j^{(m)}$, and these can be compared with the values originally assumed for these quantities. We are therefore finally faced with the iterative problem of successively adjusting the initially assumed values until they agree with the finally calculated values. This iteration is the price which has been paid for decomposing the problem into a set of separate problems of lower dimensionality, and at first sight it is a heavy one since the number of variables involved in the iteration is large for a plant of any complexity. At this point, however, it is possible to introduce the fact that a chemical plant is seldom a completely undirected pattern of interconnected units but, of its nature, is a largely sequential structure with a relatively small number of connexions which cannot be fitted into a sequential pattern. It will now be shown how this can be used to reduce the burden of iteration to manageable proportions.

Units with Sequential Structure

A unit with sequential structure will be defined as a unit with a single input and a single output stream, which is built up by sequential connexion of sub-units also of this type, as shown in Fig. 3. A plant with a large number of physically separate units can often be reduced to a relatively small number of units of this type by grouping together all sets of physical units connected sequentially. In this case we shall show that the number of vectors $\mathbf{cL}_{1i}^{(1)(n)}$ and $\mathbf{y}_j^{(m)}$ involved in the iterative process described above corresponds, not to the number of physical units in the plant, but rather to the number of compound units of sequential structure into which they can be grouped.

Consider one such compound unit, numbered R in Fig. 3 and comprising the sub-units 1, 2, ..., N connected sequentially. The equation (16) corresponding to this compound unit is

$$\mathbf{cL}_1^{(1)(R)}\mathbf{N}^{(R)} - \mathbf{g}^{(R)} = 0 \qquad (17)$$

where the suffixes i and j have been dropped, since the unit (and each sub-unit) has only one input and one output. The matrix $\mathbf{N}^{(R)}$ has $D^{(R)}$ rows and $W^{(R)}$ columns, where $D^{(R)}$ is the number of components of $\mathbf{x}^{(R)}$ and

$$W^{(R)} = W^{(1)} + W^{(2)} + \ldots + W^{(N)}.$$

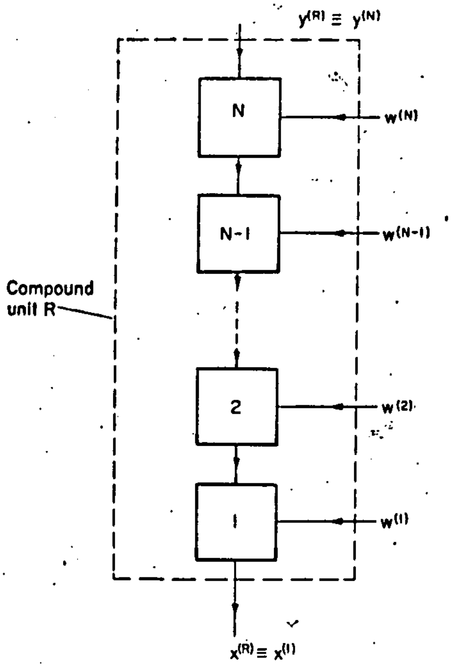


Fig. 3. Compound unit with sequential structure

The vector of adjustable parameters $w^{(R)}$ for the composite unit is the totality of the vectors $\mathbf{w}^{(1)}$, $\mathbf{w}^{(2)} \ldots \mathbf{w}^{(N)}$ for the sub-units and consequently has dimension $W^{(R)}$ equal to the sum of their dimensions, as indicated above. Now $\mathbf{N}^{(R)}$ is defined by the relation

$$d\mathbf{x}^{(R)} = \mathbf{N}^{(R)}d\mathbf{w}^{(R)} \quad (\text{with } \mathbf{y}^{(R)} \text{ constant}) \qquad (18)$$

But we can also write the following differential relations for the sub-units:

$$d\mathbf{x}^{(1)} = \mathbf{M}^{(1)}d\mathbf{x}^{(2)} + \mathbf{N}^{(1)}d\mathbf{w}^{(1)}$$

$$d\mathbf{x}^{(2)} = \mathbf{M}^{(2)}d\mathbf{x}^{(3)} + \mathbf{N}^{(2)}d\mathbf{w}^{(2)}$$

$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$

$$d\mathbf{x}^{(N)} = \mathbf{M}^{N}d\mathbf{y}^{(R)} + \mathbf{N}^{(N)}d\mathbf{w}^{(N)}$$

$$= \mathbf{N}^{(N)}\,d\mathbf{w}^{(N)} \quad (\text{with } \mathbf{y}^{(R)} \text{ constant})$$

Eliminating the variables $d\mathbf{x}^{(2)}, d\mathbf{x}^{(3)} \ldots d\mathbf{x}^{(N)}$ from these equations, we obtain

$$
\begin{aligned}
dx^{(1)} = N^{(1)}dw^{(1)} + M^{(1)}N^{(2)}dw^{(2)} + \\
+ M^{(1)}M^{(2)}N^{(3)}\,dw^{(3)} + \ldots + \\
+ M^{(1)}M^{(2)}\ldots M^{(N-1)}N^{(N)}dw^{(N)} \qquad (19)
\end{aligned}
$$

The matrix $N^{(R)}$ can be obtained in terms of matrices associated with the sub-units by comparing equations (18) and (19), and using this form for $N^{(R)}$ in equations (17), they are seen to break down into sets of equations of smaller dimensionality, as follows:

$$
\left.\begin{aligned}
&cL_1^{(1)(R)}N^{(1)} - g^{(1)} = 0 \qquad (W^{(1)}\text{ equations})\\
&cL_1^{(1)(R)}M^{(1)}N^{(2)} - g^{(2)} = 0\\
&\qquad\qquad\qquad\qquad (W^{(2)}\text{ equations})\\
&\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\\
&cL_1^{(1)(R)}M^{(1)}M^{(2)}\ldots M^{(N-1)}N^{(N)} - g^{(N)} = 0\\
&\qquad\qquad\qquad\qquad (W^{(N)}\text{ equations})
\end{aligned}\right\} \quad (20)
$$

Because of their structure, we shall see that these equations can be solved without having to assume values for variables such as $cL_1^{(1)(n)}$ and $y^{(n)}$ corresponding to the separate sub-units of the compound unit. In other words, it is necessary to assume only the input $y^{(R)}$ and the vector $cL_1^{(1)(R)}$ for each compound element of sequential structure in the plant and it is then possible to decompose the equations for the adjustable parameters into subsets of dimensionality corresponding to the separate sub-units, without introducing any iteration beyond that implied by the assumed values of $cL_1^{(1)(R)}$ and $y^{(R)}$ for each compound unit. To show this let us first take the case in which the compound unit R has no external connexions, in other words its input is the output of some other unit and its output is the input of some other unit. According to the procedure described at the end of the last section, one starts the problem by assuming values for the inputs to all the units (treating the compound unit as one unit) and this gives assumed values for both $y^{(R)}$ and $x^{(R)}$. The calculation can then be continued in one of two alternative ways.

In the first of these one assumes values for all the vectors $cL_{1l}^{(1)(n)}$ as already described, and amongst these is $cL_1^{(1)(R)}$. Using this assumed value of $cL_1^{(1)(R)}$ one can then proceed to solve the first of equations (20). Since $N^{(1)}$ is a function of the variables $y^{(1)}$, $w^{(1)}$ and $y^{(1)}$ is not known, it is necessary first to express $y^{(1)}$ in terms of $x^{(1)}$, one of the variables whose values have been assumed, by solution of the unit equations of the form (1). Thus we may write

$$y^{(1)} = F^{(1)}(x^{(1)}, w^{(1)}) \qquad (21)$$

and using this the left-hand side of the first equation (20) can be expressed as a function of $w^{(1)}$ and the specified vector $x^{(1)}$ only. Accordingly it can be solved for $w^{(1)}$, and the value so obtained determines $y^{(1)}$ through equation (21), and hence in turn the matrix $M^{(1)}$.

One may now proceed to deal with the second of equations (20) in the same way. Having solved the first, the vector $cL_1^{(1)(R)}M^{(1)}$ is known, and using the equation

$$y^{(2)} = F^{(2)}(x^{(2)}, w^{(2)})$$

the vector $N^{(2)}$, and consequently the left-hand side of the second equation (20), can be expressed as a function of $w^{(2)}$ and the vector $x^{(2)}$, which is known since it is equal to the vector $y^{(1)}$ already calculated. Thus the second equation (20) can be solved for $w^{(2)}$, which in turn determines $y^{(2)}$ and $M^{(2)}$, enabling the process to be continued to the third equation (20), and so on sequentially throughout the complete set. Thus all the equations of type (20) associated with the compound unit R can be solved without assuming values of any variables other than $cL_1^{(1)(R)}$ and $x^{(R)}$.

This method of solution has led us through the equations (20) in reverse order of the sub-units. It is equally possible to carry out the solution in the opposite sequence, starting with the equation corresponding to sub-unit N and working forwards to the one corresponding to sub-unit 1. In this case, instead of starting with assumed values for the $D^{(R)}$ components of the vector $cL_1^{(1)(R)}$, we assume values for the $D^{(R)}$ components of the vector $cL_1^{(1)(R)}M^{(1)}M^{(2)}\ldots M^{(N-1)}$. Using these, the left-hand side of the last of equations (20) is a function of $y^{(N)} \equiv y^{(R)}$ and $w^{(N)}$ only. Since $y^{(R)}$ is a unit input and has an assumed value, this equation can then be solved for $w^{(N)}$, which in turn determines $x^{(N)}$ and consequently $y^{(N-1)}$. Now the penultimate equation (20) may be written

$$[\mathbf{c}\mathbf{L}_1^{(1)(R)}\mathbf{M}^{(1)}\ldots\mathbf{M}^{(N-1)}]\mathbf{M}^{-1(N-1)}\mathbf{N}^{(N-1)} - \mathbf{g}^{(N-1)} = 0 \quad (22)$$

where $\mathbf{M}^{-1(N-1)}$ is the inverse of the (square) matrix $\mathbf{M}^{(N-1)}$, and is consequently a function of $\mathbf{y}^{(N-1)}$ and $\mathbf{w}^{(N-1)}$. However, $\mathbf{y}^{(N-1)}$ is known from the calculations on stage N, so the left-hand side of equation (22) is a known function of $\mathbf{w}^{(N-1)}$ only, and the equation can be solved for this variable. This in turn determines $\mathbf{x}^{(N-1)} \equiv \mathbf{y}^{(N-2)}$ and permits the process to be continued to the equation corresponding to the next block, and so through the sequence.

Thus we may solve equations (20) sequentially in either order, starting with assumed or known values either of the vectors $\mathbf{c}\mathbf{L}_1^{(1)(R)}$, $\mathbf{x}^{(R)}$ associated with the last unit, or of the vectors $\mathbf{c}\mathbf{L}_1^{(1)(R)}\mathbf{M}^{(1)}\ldots\mathbf{M}^{(N-1)}$, $\mathbf{y}^R$ associated with the first unit, and no further assumptions need be made. It follows that the optimization problem for the complete plant can be decomposed into problems of dimensionalities associated with the separate units, at the expense of introducing an iterative procedure which involves no more than two vectors for each of the compound units of sequential structure into which the separate plant units may be grouped. This represents a very great reduction in the burden of calculation for plants which are largely sequential in structure with a relatively small number of cross-connexions but, as we shall see, it is often possible to make use of particular properties of the structure of a given plant to reduce the amount of iteration required even further.

It should be noted that the vectors $\mathbf{c}\mathbf{L}_1^{(1)(R)}\mathbf{M}^{(1)}\mathbf{M}^{(2)}\ldots$ introduced here are precisely analogous to the Lagrangian multipliers used by HORN [5] in his treatment of the sequence of stirred tank reactors from the classical variational point of view. They are also the analogues for the discrete case of the auxiliary functions introduced in Pontryagin's method of dealing with the continuous sequential problem (e.g. the problem of optimum temperature gradients in reactors).

SIMPLE SEQUENTIAL PLANT

The analysis given above is perhaps best developed by applying it to a number of examples of increasing complexity, which will be done in this and the following sections.

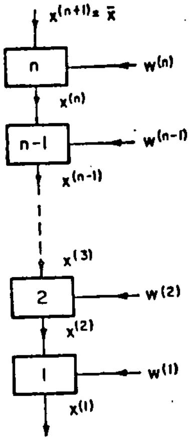


FIG. 4. Simple sequential plant.

The simplest case to consider is a plant consisting of a single sequence of units as shown in Fig. 3. In this case, however, $\mathbf{y}^{(N)}$ represents a given feed stream, $\mathbf{x}^{(1)}$ represents the output stream and is the vector which appears in the profit function P, and there are no other parts of the plant. It is then appropriate to drop the letter R and number the units as shown in Fig. 4. It is also convenient to incorporate the topological equations (3) implicitly by using a single symbol for a connected input and output, so that the vector $\mathbf{x}^{(i)}$ represents the output of the ith unit or the input of the (i-1)th. The specified properties of the feed stream are denoted by $\bar{\mathbf{x}}$. In this case equations (20) reduce to the form

$$\left.\begin{array}{ll}\mathbf{c}\mathbf{N}^{(1)} - \mathbf{g}^{(1)} = 0 & (W^{(1)}\text{ equations})\\ \mathbf{c}\mathbf{M}^{(1)}\mathbf{N}^{(2)} - \mathbf{g}^{(2)} = 0 & (W^{(2)}\text{ equations})\\ \cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot & \cdot\quad\cdot\quad\cdot\\ \mathbf{c}\mathbf{M}^{(1)}\mathbf{M}^{(2)}\ldots\mathbf{M}^{(n-1)}\mathbf{N}^{(n)} - \mathbf{g}^{(n)} = 0 & \\ & (W^{n}\text{ equations})\end{array}\right\} \quad (23)$$

No matrix $\mathbf{L}_{1i}^{(1)(n)}$ appears, since the output of the sequence is itself the plant output. Since the value of $\bar{\mathbf{x}}(= \mathbf{x}^{(n+1)})$ is given, the equations could be solved sequentially, starting at unit n, if the vector $\mathbf{c}\mathbf{M}^{(1)}\mathbf{M}^{(2)}\ldots\mathbf{M}^{(n-1)}$ were known. Alternatively, since no unknown matrix of the L-type appears, the equations could be solved sequentially starting at unit 1 if the vector $\mathbf{x}^{(1)}$ were known. The available information, namely the value of $\bar{\mathbf{x}}$, does not permit the calculation to be started from either end without further assumptions, and this situation is quite typical, as we shall see. One may start instead by assuming a value for $\mathbf{x}^{(1)}$ and then solve equations (23) successively in the order 1, 2, 3 ... as described in the previous section, at each stage making use of the inverse unit equations

$$\mathbf{x}^{(p+1)} = \bar{F}^{(p)}[\mathbf{x}^{(p)}, \mathbf{w}^{(p)}]$$

The calculated value of $\mathbf{w}^{(n)}$ then determines a value for $\mathbf{x}^{(n+1)}$ and in general this will not agree with the given value $\bar{\mathbf{x}}$, so the initially assumed value of $\mathbf{x}^{(1)}$ must be adjusted iteratively, repeating the calculations at each stage, until agreement is obtained.

Alternatively one may start by assuming a value $\mathbf{M}$ for the vector $\mathbf{c}\mathbf{M}^{(1)}\mathbf{M}^{(2)} \ldots \mathbf{M}^{(n-1)}$ and solve equations (23) in the order n, $n-1$, $n-2$, ... as previously described, successively generating the matrices $\mathbf{M}^{-1(n-1)}$, $\mathbf{M}^{-1(n-2)}$ ... needed to continue the calculation to the next stage. Finally a calculated value of the vector $\mathbf{c}$ may be deduced from the assumed vector $\mathbf{M}$ by the relation

$$(\mathbf{c})_{\text{calc}} = \mathbf{M} \, . \, \mathbf{M}^{-1(n-1)} \, . \, \mathbf{M}^{-1(n-2)} \, . \ldots \mathbf{M}^{-1(1)}$$

and this may be compared with the specified value of $\mathbf{c}$. The initially assumed vector $\mathbf{M}$ must then be adjusted iteratively until the two agree.

We may summarize this by saying that the first procedure optimizes the given profit function P for an input which is not necessarily the one specified, while the second starts from the specified input and optimizes a profit function which is not necessarily the one specified. These two procedures are identical with those introduced by HORN [5] in discussing a sequence of reactors, but HORN arrived at the result by using Lagrangian multipliers.

It should be noted that either of these methods, with any reasonably intelligent method of iterative adjustment, is much more economical in computation than the method of dynamic programming, and furthermore involves none of the intermediate tabulation at each stage which makes dynamic programming so demanding of storage space. It has often been claimed that the method of dynamic programming applied to a sequential problem greatiy reduces the amount of calculation required compared with the classical method of seeking a stationary value, and furthermore that the dynamic programming procedure is ideally suited to automatic computation. It is felt, however, that both these claims are mistaken. It has been shown here that the classical equations for a stationary value themselves decompose into sets of equations of lower dimensionality, thus reducing the problem in precisely the same way as dynamic programming sets out to do, but introducing an iterative process which is computationally much more economical than the intermediate tabulation involved in dynamic programming. The extent of the intermediate tabulation required in any problem of more than one dimension surely makes dynamic programming a method singularly ill-suited to automatic computation. It is, of course, perfectly true that it produces simultaneously the solution of a complete class of problems, but one would seldom be interested in a complete exploration of so much territory. The fact that one cannot solve a specified problem without such an exploration nevertheless reveals the staggering redundancy of the procedure. In the author's opinion the true virtue of dynamic programming—one which is seldom stressed—is the fact that it inevitably leads to the greatest (or least) value of the function to be optimized, irrespective of the possible existence of multiple stationary values. For this reason it would be of great interest to extend it to handle the non-sequential type of problems which are discussed (from the classical variational viewpoint) in the present paper.

A MORE COMPLICATED SEQUENTIAL PLANT

Fig. 5 shows a somewhat more complicated plant than was considered in the last section. Its

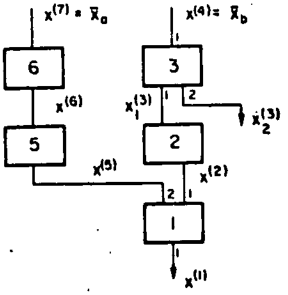


FIG. 5. More complicated sequential plant.

tructure is still sequential, as there are no paths orming closed loops, but it now has a branching orm. There are two feeds with specified properties $\bar{\mathbf{x}}_a$ and $\bar{\mathbf{x}}_b$ as indicated, and once again the topological equations are taken into account by using the same symbol for a connected input and output. Where a unit has more than one input or output, both inputs and outputs are numbered so that the suffix notation introduced earlier can be used. We then have

$$\begin{aligned} d\mathbf{x}^{(1)} = {} & \mathbf{N}^{(1)}d\mathbf{w}^{(1)} + \\ & + \mathbf{M}^{(1)}_{11}\{\mathbf{N}^{(2)}d\mathbf{w}^{(2)} + \mathbf{M}^{(2)}[\mathbf{N}^{(3)}_1 d\mathbf{w}^{(3)}]\} + \\ & + \mathbf{M}^{(1)}_{12}\{\mathbf{N}^{(5)}d\mathbf{w}^{(5)} + \mathbf{M}^{(5)}[\mathbf{N}^{(6)}d\mathbf{w}^{(6)}]\} \end{aligned}$$

and correspondingly equations (16) take the form

$$\left.\begin{aligned} &\text{(i)} \quad \mathbf{cN}^{(1)} - \mathbf{g}^{(1)} = 0 \qquad (W^{(1)} \text{ equations}) \\ &\text{(ii)} \quad \mathbf{cM}^{(1)}_{11}\mathbf{N}^{(2)} - \mathbf{g}^{(2)} = 0 \qquad (W^{(2)} \text{ equations}) \\ &\text{(iii)} \quad \mathbf{cM}^{(1)}_{11}\mathbf{M}^{(2)}\mathbf{N}^{(3)}_1 - \mathbf{g}^{(3)} = 0 \qquad (W^{(3)} \text{ equations}) \\ &\text{(iv)} \quad \mathbf{cM}^{(1)}_{12}\mathbf{N}^{(5)} - \mathbf{g}^{(5)} = 0 \qquad (W^{(5)} \text{ equations}) \\ &\text{(v)} \quad \mathbf{cM}^{(1)}_{12}\mathbf{M}^{(5)}\mathbf{N}^{(6)} - \mathbf{g}^{(6)} = 0 \qquad (W^{(6)} \text{ equations}) \end{aligned}\right\} \quad (24)$$

In order to solve equation (24)(i) for $\mathbf{w}^{(1)}$ it is necessary to express $\mathbf{N}^{(1)}$ in terms of $\mathbf{w}^{(1)}$ and quantities known or assumed. However, since unit 1 has two inputs and $\mathbf{N}^{(1)}$ depends, in general, on both of them, this cannot be done without assuming values for $\mathbf{x}^{(1)}$ and one of the inputs. Suppose we assume a value for $\mathbf{x}^{(2)}$. Then the unit equation for unit 1 can be solved to give $\mathbf{x}^{(5)}$ in terms of $\mathbf{x}^{(1)}$, $\mathbf{w}^{(1)}$ and the assumed value of $\mathbf{x}^{(2)}$. Using this in $\mathbf{N}^{(1)}$, the left-hand side of equation (24)(i) is expressed in terms of $\mathbf{w}^{(1)}$ and the two vectors $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ whose values have been assumed. The equation may then be solved for $\mathbf{w}^{(1)}$, which in turn determines $\mathbf{x}^{(5)}$. Now the second equation (24) can be solved for $\mathbf{w}^{(2)}$, which in turn determines $\mathbf{x}^{(3)}_1$ and permits the third equation to be solved for $\mathbf{w}^{(3)}$. Similarly the calculated value of $\mathbf{x}^{(5)}$ permits the fifth and sixth equations to be solved for $\mathbf{w}^{(5)}$ and $\mathbf{w}^{(6)}$. Finally one can calculate values for $\mathbf{x}^{(7)}$ and $\mathbf{x}^{(4)}_1$, which may be compared with the specified feed vectors $\bar{\mathbf{x}}_a$ and $\bar{\mathbf{x}}_b$. In order to obtain agreement between the calculated and specified values, the assumed values of $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ are available. The consistency of the unit equations will ensure that the total number of assumed quantities available to be varied is equal to the total number of components of the specified feed vectors, so that iterative adjustment is possible.

Plant with a By-Pass Stream

This is again a system of sequential type, but a branch from the main sequence rejoins it at a later stage as shown in Fig. 6. The suffix notation is used when necessary to distinguish separate inputs and outputs of the same unit, as in the previous case, and we have

$$\begin{aligned} d\mathbf{x}^{(1)}_1 = {} & \mathbf{N}^{(1)}_1 d\mathbf{w}^{(1)} + \mathbf{M}^{(1)}_{11}\{\mathbf{N}^{(2)}d\mathbf{w}^{(2)} + \\ & + \mathbf{M}^{(2)}[\mathbf{N}^{(3)}d\mathbf{w}^{(3)} + \mathbf{M}^{(3)}(\mathbf{N}^{(4)}_1 d\mathbf{w}^{(4)})]\} + \\ & + \mathbf{M}^{(1)}_{12}\{\mathbf{N}^{(6)}d\mathbf{w}^{(6)} + \mathbf{M}^{(6)}[\mathbf{N}^{(7)}d\mathbf{w}^{(7)} + \\ & + \mathbf{M}^{(7)}(\mathbf{N}^{(4)}_2 d\mathbf{w}^{(4)})]\} \end{aligned}$$

The corresponding equations of type (16) take the form:

$$\left.\begin{aligned} &\text{(i)} \quad \mathbf{cN}^{(1)}_1 - \mathbf{g}^{(1)} = 0 \\ &\text{(ii)} \quad \mathbf{cM}^{(1)}_{11}\mathbf{N}^{(2)} - \mathbf{g}^{(2)} = 0 \\ &\text{(iii)} \quad \mathbf{cM}^{(1)}_{11}\mathbf{M}^{(2)}\mathbf{N}^{(3)} - \mathbf{g}^{(3)} = 0 \\ &\text{(iv)} \quad \mathbf{cM}^{(1)}_{12}\mathbf{N}^{(6)} - \mathbf{g}^{(6)} = 0 \\ &\text{(v)} \quad \mathbf{cM}^{(1)}_{12}\mathbf{M}^{(6)}\mathbf{N}^{(7)} - \mathbf{g}^{(7)} = 0 \\ &\text{(vi)} \quad \mathbf{cM}^{(1)}_{11}\mathbf{M}^{(2)}\mathbf{M}^{(3)}\mathbf{N}^{(4)}_1 + \\ &\qquad + \mathbf{cM}^{(1)}_{12}\mathbf{M}^{(6)}\mathbf{M}^{(7)}\mathbf{N}^{(4)}_2 - \mathbf{g}^{(4)} = 0 \end{aligned}\right\} \quad (25)$$

Equation (25)(i) may be solved for $\mathbf{w}^{(1)}$, as in the previous example, after assuming values for $\mathbf{x}^{(1)}_1$ and one of the inputs to Unit 1. We may, for example, assume values for the components of $\mathbf{x}^{(2)}$. Together with the assumed value of $\mathbf{x}^{(1)}_1$ and the value of $\mathbf{w}^{(1)}$ obtained by solution of equation (25)(i), this then determines the vector $\mathbf{x}^{(6)}$.

The calculations may then be continued sequentially up each branch in the way already described, determining $\mathbf{w}^{(2)}$, $\mathbf{w}^{(3)}$, $\mathbf{w}^{(6)}$, $\mathbf{w}^{(7)}$ and the vectors $\mathbf{x}^{(4)}_1$ and $\mathbf{x}^{(4)}_2$.

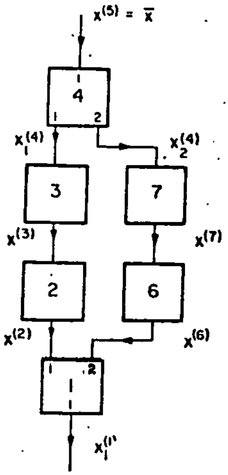


FIG. 6. Plant with bypass stream.

Now associated with unit 4 there will be two unit equations:

$$\mathbf{x}_1^{(4)} = F_1^{(4)}[\mathbf{x}^{(5)}, \mathbf{w}^{(4)}] \tag{26}$$

and

$$\mathbf{x}_2^{(4)} = F_2^{(4)}[\mathbf{x}^{(5)}, \mathbf{w}^{(4)}] \tag{27}$$

We may solve either of these for $\mathbf{x}^{(5)}$ as a function of $\mathbf{w}^{(4)}$, for example the former, giving

$$\mathbf{x}^{(5)} = \bar{F}_1^{(4)}[\mathbf{x}_1^{(4)}, \mathbf{w}^{(4)}] \tag{28}$$

and this function for $\mathbf{x}^{(5)}$ may then be used to express $\mathbf{N}_1^{(4)}$ and $\mathbf{N}_2^{(4)}$ as functions of $\mathbf{w}^{(4)}$ only, since $\mathbf{x}_1^{(4)}$ is already determined. Equation (25)(vi) may then be solved for $\mathbf{w}^{(4)}$ and insertion of the result in equation (27) determines, in turn, the components of the vector $\mathbf{x}_2^{(4)}$. In general these will not agree with the values obtained already by the sequential solution of the equations associated with units 6 and 7. Furthermore, $\mathbf{w}^{(4)}$ determines $\mathbf{x}^{(5)}$ through equation (28) and the components so determined will not, in general, agree with the specified feed vector $\bar{\mathbf{x}}$.

We see, therefore, that there are two mis-matches; one between the specified and calculated values of the vector associated with the feed and one between the values of the vector $\mathbf{x}_2^{(4)}$ calculated in two different ways. Correspondingly there are two vectors, $\mathbf{x}_1^{(1)}$ and $\mathbf{x}^{(2)}$, whose values have been assumed and may be adjusted iteratively to eliminate the mis-matches.

It will be recalled that in Section 2 it was shown in a general way that the stationary value problem for the plant as a whole could be decomposed into sub-problems with the dimensionalities of the separate units by introducing an iterative process involving vectors $c\mathbf{L}_{1i}^{(1)(n)}$ and $\mathbf{y}_j^{(m)}$ associated with each unit, while in Section 3 it was shown that the iterative work could be considerably reduced since it is necessary to introduce only one set of these variables for each compound unit of simple sequential structure. Now in fact, in the examples worked so far, the number of variables involved in the iteration has been smaller even than this. Indeed it has not been necessary to assume values for any vectors of the type $c\mathbf{L}_{1i}^{(1)(n)}$. By starting from the output appearing in the profit function and working backwards towards the feeds, it has proved possible to solve the problem in its dimensionally decomposed form with iterative adjustment only of vectors associated with plant streams. The reason for this is that the expressions for the matrices $\mathbf{L}_{1i}^{(1)(n)}$ in terms of the matrices $\mathbf{M}_{ij}^{(n)}$ are sequential in nature, permitting one matrix to be obtained from the previous one as the calculation proceeds through the plant. It is not difficult to see that this is a general property of structures of the type we have considered so far which do not contain any closed loop configurations. To be more precise we shall say that a structure contains closed loop configurations if variation of an output of a unit causes a consequential variation in one of its inputs, and we shall show that the power of the general method developed in Sections 2 and 3 is only fully revealed when dealing with such configurations.

A SIMPLE FEEDBACK LOOP

Fig. 7 shows a simple structure with a single closed loop configuration, using the same type of notation as in previous examples. As indicated, the units 2 and 6 may themselves consist of simple sequential chains of sub-units, but we shall first solve the problem treating them as single units. We then have

$$d\mathbf{x}_1^{(1)} = \mathbf{N}_1^{(1)} d\mathbf{w}^{(1)} + \mathbf{M}_{11}^{(1)} d\mathbf{x}^{(2)} \tag{29}$$

$$d\mathbf{x}^{(2)} = \mathbf{N}^{(2)} d\mathbf{w}^{(2)} + \mathbf{M}^{(2)}[\mathbf{N}_1^{(3)} d\mathbf{w}^{(3)} + \\ + \mathbf{M}_{11}^{(3)}\mathbf{N}^{(4)} d\mathbf{w}^{(4)} + \mathbf{M}_{12}^{(3)} d\mathbf{x}^{(6)}] \tag{30}$$

and

$$dx^{(6)} = N^{(6)}dw^{(6)} + M^{(6)}[N_2^{(1)}dw^{(1)} + + M_{21}^{(1)}dx^{(2)}] \quad (31)$$

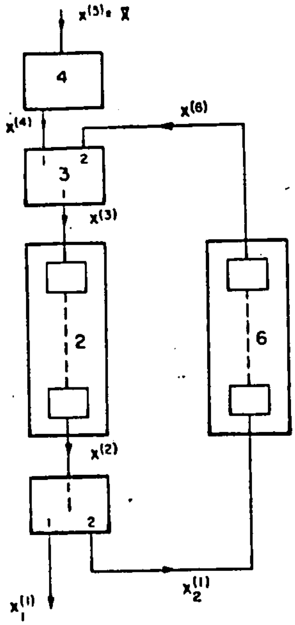


FIG. 7. Plant with simple feedback top.

The variables $dx^{(2)}$ and $dx^{(6)}$ may be eliminated from these to give the required equation for $dx_1^{(1)}$, which is

$$\begin{aligned} dx_1^{(1)} = {} & [N_1^{(1)}+M_{11}^{(1)}H^{-1}M^{(2)}M_{12}^{(3)}M^{(6)}N_2^{(1)}]dw^{(1)} + \\ & + M_{11}^{(1)}H^{-1}N^{(2)}dw^{(2)} + \\ & + M_{11}^{(1)}H^{-1}M^{(2)}N_1^{(3)}dw^{(3)} + \\ & + M_{11}^{(1)}H^{-1}M^{(2)}M_{11}^{(3)}N^{(4)}dw^{(4)} + \\ & + M_{11}^{(1)}H^{-1}M^{(2)}M_{12}^{(3)}N^{(6)}dw^{(6)} \end{aligned} \quad (32)$$

where the matrix H is given by

$$H = 1 - M^{(2)}M_{12}^{(3)}M^{(6)}M_{21}^{(1)} \quad (33)$$

and H^{-1} is its inverse. We may now write down the equations for the vectors $w^{(n)}$ in their standard form (16):

$$\left.\begin{aligned} &\text{(i)} \quad cL_{11}^{(1)(1)}N_1^{(1)} + cL_{12}^{(1)(1)}N_2^{(1)} - g^{(1)} = 0 \\ &\text{(ii)} \quad cL_1^{(1)(2)}N^{(2)} - g^{(2)} = 0 \\ &\text{(iii)} \quad cL_{11}^{(1)(3)}N_1^{(3)} - g^{(3)} = 0 \\ &\text{(iv)} \quad cL_1^{(1)(4)}N^{(4)} - g^{(4)} = 0 \\ &\text{(v)} \quad cL_1^{(1)(6)}N^{(6)} - g^{(6)} = 0 \end{aligned}\right\} \quad (34)$$

The matrices $L_{1i}^{(1)(n)}$ being given in terms of the M-matrices by comparing equation (32) with its alternative form

$$dx_1^{(1)} = \sum_n \sum_i L_{1i}^{(1)(n)} N_i^{(n)} dw^{(n)} \quad (15)$$

It is no longer possible to solve equations (34) sequentially without assuming values for the $L_{1i}^{(1)(n)}$, as it was in the previous cases, since each such matrix, except $L_{11}^{(1)(1)}$ depends on variables associated with all the units in the closed loop. Thus we now have a case in which there is no means of side-tracking the procedure described in Section 2. It is first necessary to assume values for the components of the vector $cL_{12}^{(1)(1)}$ ($cL_{11}^{(1)(1)}$ is already known) and for the components of the output vector $x_1^{(1)}$. $x^{(2)}$ can then be expressed in terms of $x_1^{(1)}$ and $w^{(1)}$ using the unit equations of unit 1, and hence the left-hand side of equation (34)(i) can be expressed as a function of $w^{(1)}$. The equation can then be solved and this determines $x^{(2)}$. It is not necessary to assume values for the components of the vector $cL_1^{(1)(2)}$, since it can be seen to be given by the following linear combination of $cL_{11}^{(1)(1)}$ and $cL_{12}^{(1)(1)}$:

$$cL_1^{(1)(2)} = cL_{11}^{(1)(1)}M_{11}^{(1)} + cL_{12}^{(1)(1)}M_{21}^{(1)} \quad (35)$$

Equation (34)(ii) can then be solved for $w^{(2)}$, which in turn determines $x^{(3)}$. It is then seen from equation (32) that $cL_{11}^{(1)(3)}$ can be obtained from $cL_1^{(1)(2)}$ according to

$$cL_{11}^{(1)(3)} = cL_1^{(1)(2)}M^{(2)} \quad (36)$$

since $M^{(2)}$ is known from the solution for unit 2. However, in order to solve equation (34)(iii) for $w^{(3)}$ it is necessary to express $N_1^{(3)}$ in terms of $w^{(3)}$ and quantities already known, and since $N_1^{(3)}$ is in general a function of both inputs to unit 3, this cannot be done without assuming values for some of the components of the associated vectors. For example, if $x^{(3)}$, $x^{(4)}$ and $x^{(6)}$ have the same dimensionality, it will suffice to assume values for the components of one of the two vectors $x^{(4)}$, $x^{(6)}$. Let us assume a value for $x^{(4)}$. Then the unit equation for unit 3 can be solved to give $x^{(6)}$ in terms of $x^{(3)}$, $w^{(3)}$ and the assumed value of $x^{(4)}$. Using this in $N_1^{(3)}$, the left-hand side of equation (34)(iii) is expressed in terms of $w^{(3)}$

and known (or assumed) quantities. It may therefore be solved for $w^{(3)}$, which in turn determines $x^{(6)}$. We next deal with equation (34)(v), which may be solved without assuming anything further, since $cL_1^{(1)(6)}$ is determined from $cL_{11}^{(1)(3)}$ according to

$$cL_1^{(1)(6)} = cL_{11}^{(1)(3)}M_{12}^{(3)} \tag{37}$$

and the matrix $M_{12}^{(3)}$ is known from the solution for unit 3. The left-hand side of equation (34)(v) can therefore be expressed as a function of $w^{(6)}$ and the equation can be solved, hence determining $x_2^{(1)}$. However, $x_2^{(1)}$ could also be calculated from the values of $x^{(2)}$ and $w^{(1)}$ already determined, and in general there will be a mis-match between the vectors calculated in these two different ways.

Finally we deal with equation (34)(iv). There is no need to assume a value for $cL_1^{(1)(4)}$ since it is determined from $cL_{11}^{(1)(3)}$ according to

$$cL_1^{(1)(4)} = cL_{11}^{(1)(3)}M_{11}^{(3)} \tag{38}$$

and the matrix $M_{11}^{(3)}$ is known from the solution for unit 3. A value of $x^{(4)}$ has been assumed earlier, so equation (34)(iv) may be solved for $w^{(4)}$, which in turn determines a value of $x^{(5)}$, and this will not, in general, match the specified feed vector $\bar{x}$.

Having completed a solution, we are in a position to calculate all the matrices $M_{ij}^{(n)}$, and hence the matrix H^{-1}, so using the expression for the vector $cL_1^{(1)(2)}$ in terms of these matrices, we may calculate the components of this vector. In general the value so obtained will not agree with that originally assumed in obtaining the solution. We are therefore in the position of having assumed values for three vectors, namely $x_1^{(1)}$, $x^{(4)}$ and $cL_1^{(1)(2)}$, in order to obtain a solution, and hence having arrived at a solution with three mis-matches. The assumed values must then be adjusted until the mis-matches are eliminated. The consistency of the unit equations will ensure that sufficient variables are available for this iteration.

Once again it has proved unnecessary to assume values for all the vectors $cL_{1i}^{(1)(n)}$ and $y_j^{(m)}$ in order to obtain a solution, and this is because the feed-back loop is itself a sequential structure, complicated only by the fact that its head is joined to its tail. Thus the presence of a closed loop increases the amount of iteration required compared with the branching structure considered earlier (in particular vectors $cL_{1i}^{(1)(n)}$ are drawn into the iterative process), but an intelligent use of the largely sequential nature of the structure reduces the number of variables involved in the iteration far below what might be expected from the general treatment of Section 2. Indeed it would be difficult to devise a structure so entangled that the full number of variables must be used in the iterative process.

If each of the units 2 and 6 is, in fact, a compound unit of sequential structure with many sub-units as indicated in Fig. 7, this does not increase the number of variables in the iteration. The matrices $N^{(2)}$ and $N^{(6)}$ merely decompose into a set of matrices associated with each individual sub-unit, as indicated in equations (18) and (19), and these can be calculated sequentially.

CONCLUSIONS

The method developed in this paper permits the problem of finding a stationary value of a specified objective function in a complex chemical plant to be decomposed into a set of sub-problems of lower dimensionality. To this extent it serves the same purpose as the dynamic programming algorithm but, unlike dynamic programming, its application is not restricted to simple sequential structures. Like all other classical variational calculations, the present method only ensures that the objective function will take a stationary value, and this is not necessarily its greatest value. It would therefore be very valuable to develop the method of dynamic programming itself in such a way that it could be applied to complex non-sequential structures, since dynamic programming always leads to the greatest value of the objective function. Attempts [6, 7] to do this known to the author fail, because they do not correctly consider independent variations in the adjustable parameters, as has been discussed elsewhere [8]. In particular, when a closed loop is present in the system, any variation in an adjustable parameter of a given unit leads to changes which are propagated round the loop and cause a change in one of the inputs of the unit. At no stage, therefore, is it permissible to consider variation of the parameters of a unit with fixed values of its inputs.

The examples which have been used to illustrate the procedure have deliberately been chosen to be simple, but the single feedback loop illustrates all the features of more complicated problems with multiple recycles. In particular, the reader will be able to convince himself that the method experiences no difficulty in dealing with structures such as interlaced feedback loops and even more complicated configurations.

Note added in proof: It is implied in the above paper that the optimization procedure proposed by Mitten and Nemhauser is fallacious. It is now realized that this view resulted from a misunderstanding of these authors' proposals. I am grateful to Dr. R. Aris for pointing out my error.

Notation

$\mathbf{c}$ Row vector used to form scalar objective function

$D_i^{(n)}$ Dimensionality of vector $\mathbf{x}_i^{(n)}$

$E_j^{(m)}$ Dimensionality of vector $\mathbf{y}_j^{(m)}$

$F_i^{(n)}$ Function relating $\mathbf{x}_i^{(n)}$ to variables $\mathbf{y}_k^{(n)}$ and $\mathbf{w}^{(n)}$

$F^{(n)}$ $F_i^{(n)}$ when unit has only one input and one output

$\mathcal{F}^{(n)}$ Function relating $\mathbf{y}^{(n)}$ to variables $\mathbf{x}^{(n)}$ and $\mathbf{w}^{(n)}$. Obtainable from $F^{(n)}$

G Number of specified feeds

$\mathbf{g}^{(n)}$ Vector obtained by differentiating $G^{(n)}$ with respect to components of $\mathbf{w}^{(n)}$

$G^{(n)}$ Scalar function of the vector argument $\mathbf{w}^{(n)}$

$\mathbf{H}$ Matrix defined by equation (33)

$\mathbf{L}_1{}^{(i^1)(n)}$ Matrix defined by equation (15)

$\mathbf{M}$ The vector $\mathbf{c}\mathbf{M}^{(1)}\mathbf{M}^{(2)} \ldots \mathbf{M}^{(n-1)}$

$\mathbf{M}_{ij}^{(n)}$ 'Matrix defined by equation (13), characteristic of unit n

$\mathbf{M}^{(n)}$ Matrix $\mathbf{M}_{ij}^{(n)}$ when unit has only one input and one output

$\mathbf{N}_i^{(n)}$ Matrix defined by equation (14), characteristic of unit n

$\mathbf{N}^{(n)}$ Matrix $\mathbf{N}_i^{(n)}$ when unit has only one input and one output

P Scalar objective function

$\mathbf{w}^{(n)}$ Vector of adjustable parameters for unit n

$W^{(n)}$ Dimensionality of vector $\mathbf{w}^{(n)}$

$\mathbf{x}_i^{(n)}$ Vector associated with ith output of unit n

$\mathbf{y}_j^{(m)}$ Vector associated with jth input of unit m

$\bar{\mathbf{y}}_k^{(1)}$ Vector associated with a specified feed

ν_1 Number $\sum\limits_n \sum\limits_i D_i^{(n)}$

ν_2 Number $\sum\limits_m \sum\limits_j E_j^{(m)} - G$

ν_3 Number G

ν_4 Number $\sum\limits_n W^{(n)}$

References

[1] Bellman R., *Dynamic Programming*. Princeton 1957.

[2] Aris R., *The Optimal Design of Chemical Reactors*. Academic Press, New York, 1961.

[3] Pontryagin L. S., *Usp. Mat. Nauk* 1959 **14** 3.

[4] Swinnerton-Dyer H. P. F., *Proc. Lond. Math. Soc.* 1957 **7** 568.

[5] Horn F., *Chem. Engng. Sci.* 1961 **15** 176.

[6] Rudd D. F. and Blum E. D., *Chem. Engng. Sci.* 1962 **17** 277.

[7] Mitten L. G. and Nemhauser G. L., *Chem. Engng. Prog.* 1963 **59** 52.

[8] Jackson R., *Chem. Engng. Sci.* 1963 **18** 215.

Résumé—La détermination des conditions optimales dans une usine chimique qui comprend un certain nombre d'unités en liaison entre elles présente souvent des difficultés de calcul considérables à cause du grand nombre de paramètres dont la variation doit être considérée simultanément.

La méthode de programmation dynamique permet de décomposer le problème en une série de problèmes secondaires, mais son application reste limitée aux systèmes formés de chaînes d'unités simples et droites.

L'auteur décrit ici une approche classique par variations, qui permet de réaliser une décomposition dimensionnelle analogue dans le cas d'unités aussi complexes que l'on veut. D'autre part, il étudie en détail quelques systèmes qui illustrent sans complexité algébrique excessive les caractéristiques principales de la méthode.

Chemical Engineering Science, 1964, Vol. 19, pp. 253–260. Pergamon Press Ltd., Oxford. Printed in Great Britain.

B6

A generalized variational treatment of optimization problems in complex chemical plants

R. Jackson

University of Edinburgh and Heriot-Watt College

(*Received* 28 *August* 1963; *in revised form* 4 *November* 1963)

Abstract—A method of decomposing optimization problems in topologically complex plants, described in a previous paper, is extended to deal with objective functions of a much more general type and to include units with a continuous infinity of adjustable parameters, such as tubular reactors. An alternative derivation of the formalism based on Lagrangian multipliers is also given.

Introduction

In a previous paper [1] a method was described for decomposing optimization problems in topologically complex plants into sub-problems with the dimensionalities associated with the separate units. The method was developed by a classical variational argument of a very straightforward type and, for simplicity, an objective function formed by linear combination of the elements of a single output vector was considered. The discussion was also limited to units with a finite number of adjustable parameters, and therefore excluded cases such as tubular reactors with an adjustable temperature gradient.

In the present paper it will be shown how this work can be generalized to deal with objective functions of a much more general type and to include units with a continuous infinity of adjustable parameters, such as tubular reactors with adjustable temperature gradients. At the same time an alternative derivation of the results based on the use of Lagrangian multipliers will be given. This is more concise and elegant than the derivation given in the earlier paper, but it provides a less appropriate introduction to the subject, since it is not so easily interpreted physically.

Statement of the Problem

As before [1] we shall consider a plant consisting of a set of interconnected units, each with a set of inputs and outputs and a number of adjustable parameters. A number n will be assigned to each unit, and its input and output streams will also be numbered. The properties of each stream are characterised by a set of physical quantities which may be regarded as the components of a vector, and similarly the adjustable operating and design variables for a given unit may be regarded as the components of a second vector. The vector associated with the i-th output stream of the n-th unit will be denoted by x_i^n and the vector of adjustable parameters for this unit by w^n. It should be noted that the suffix i serves to identify the output stream and does not refer to individual components of the vector x_i^n. The numbers n and i may be called the identifying indices of a stream.

In the previous paper [1] symbols y_j^n were introduced for the vectors associated with unit input streams, but here we shall introduce a different notation. Unit inputs are connected to outputs of other units (except those which form feeds to the plant as a whole) so, strictly speaking, the output vectors x_i^n suffice to describe all process streams in the plant, and it is only necessary to identify those inputs and outputs which are connected to each other. This can be done by listing the output stream connected to each input, and we shall use the notation $(\bar{n}, \bar{j})$ to indicate the identifying indices of the output stream which is connected to the j-th input of the n-th unit. The vector describing the physical properties of this stream is correspondingly denoted by $x_{\bar{j}}^{\bar{n}}$. Alternatively we could list the input stream connected to each output, using the notation $(\tilde{n}, \tilde{i})$ for the identifying indices of the input stream connected to the i-th output of the n-th unit. The corresponding vector is then denoted by $x_{\bar{i}}^{\bar{n}}$. This notation is illustrated in Fig. 1. There will, of course, be no values of $\bar{n}$ and $\bar{i}$ corresponding to outputs (n, i) which form plant products, and these will be referred to as free outputs.

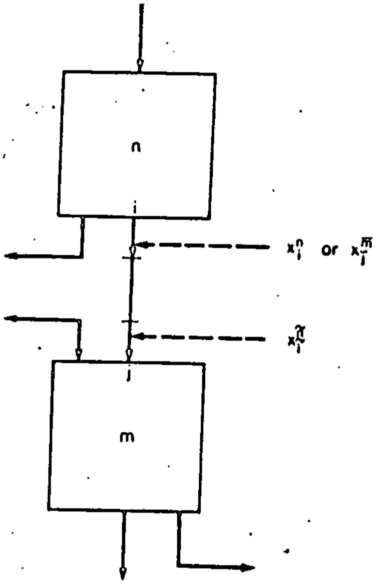


FIG. 1. Illustration of notation.

The relations between outputs, inputs and adjustable parameters for a given unit take the form of a set of equations

$$x_i^n = F_i^n(x_{\bar{j}}^{\bar{n}}, w^n) \tag{1}$$

which we shall refer to as the unit equations of the n-th unit. In general the functions F_i^n depend on the vectors $x_{\bar{j}}^{\bar{n}}$ associated with all the inputs of the n-th unit. There is no need to write separate equations identifying the vectors associated with streams which are connected, as in the previous work [1], since the topological structure of the plant is determined by the listed correspondence between (n, j) and $(\bar{n}, \bar{j})$ or between (n, i) and $(\bar{n}, \bar{i})$.

We shall consider the problem of finding values for the adjustable parameters of the units which maximise an objective function P of the form

$$P = H(x_i^n) - \sum_n G^n(w^n) \tag{2}$$

where H is a specified scalar function which may depend on all the output vectors x_i^n of the plant units, while the G^n are specified scalar functions of the vector arguments w^n. We shall not restrict ourselves to cases in which w^n has a finite number of dimensions, but shall include units such as tubular reactors with adjustable temperature gradients, for which w^n has a continuous infinity of components. In such cases the G^n are more properly regarded as functionals.

In the earlier paper [1] a simpler objective function of the form

$$P = cx_1^1 - \sum_n G^n(w^n) \tag{3}$$

was considered, but the generalization represented by equation (2) is desirable for two reasons. In the first place, the operating returns from the process may depend on more than one product stream; for example there may be more than one salable product, or one of the outputs may be an effluent which it is costly to disperse. Secondly, in addition to imposing constraints on the values taken by the adjustable parameters w^n, it is often desirable to constrain the values of certain quantities associated with the process streams themselves. For example, the properties of available materials of construction frequently impose upper limits on the permissible temperatures of certain streams. Such constraints can be imposed by including in P a term which takes very large negative values when the variables in question pass outside their permitted ranges, and the general form of the function H in equation (2) permits terms of this type to be included.

CONDITIONS FOR A STATIONARY VALUE OF P

Differentiation of equation (2) gives the first order variation in the objective function in the form

$$dP = \sum_n \sum_i h_i^n \, dx_i^n - \sum_n \partial G^n \tag{4}$$

where h_i^n is the row vector of partial derivatives

$$[h_i^n]_p = \frac{\partial H}{\partial [x_i^n]_p}. \tag{5}$$

The suffix p indicating the p-th component of a vector. The notation ∂G^n is introduced to indicate the first order variation in G^n produced by a small variation in the adjustable parameters w^n of unit n.

Interior : $\partial G_i^n = g_j^n \, dw^n$

[illegible]

The variations dx_i^n appearing on the right hand side of equation (4) are not independent since the unit equations (1) must be satisfied. Differentiating these gives the following set of equations relating small variations in the vectors associated with process streams

$$dx_i^n = \sum_j M_{ij}^n \, dx_j^{\bar{n}} + \partial x_i^n \tag{6}$$

where M_{ij}^n is the following matrix of partial derivatives

$$[M_{ij}^n]_{pq} = \frac{\partial [F_i^n]_p}{\partial [x_j^{\bar{n}}]_q}. \tag{7}$$

The notation ∂x_i^n is analogous to the notation ∂G^n introduced above and denotes the first order variation in x_i^n produced by a small variation in the adjustable parameters w^n with the inputs to unit n held constant.

Our problem is now to find the conditions that dP as given by equation (4), should vanish identically, when the dx_i^n are constrained to satisfy equations (6). The constraints may be introduced by the use of Lagrangian multipliers [2], which we shall denote by λ_i^n. The multiplier λ_i^n is associated with the equation (6) for dx_i^n, and correspondingly it is a row vector of the same dimensionality as the column vector x_i^n. Combining equations (4) and (6) by means of the Lagrangian multipliers, the condition for a stationary value becomes

$$\sum_n \sum_i h_i^n \, dx_i^n - \sum_n \partial G^n - \\ - \sum_n \sum_k \lambda_k^n \left\{ dx_k^n - \sum_j M_{kj}^n \, dx_j^{\bar{n}} - \partial x_k^n \right\} = 0. \tag{8}$$

In this equation the dx_i^n may be regarded as independent variations, while the ∂x_k^n and ∂G^n are independent for different values of n, since they then represent the effects of varying the adjustable parameters of different units. Collecting the terms in dx_i^n and the remaining terms separately, equation (8) may be written

$$\sum_n \sum_i \left\{ h_i^n - \lambda_i^n + \sum_k \lambda_k^{\bar{n}} M_{ki}^{\bar{n}} \right\} dx_i^n + \\ + \sum_n \left\{ \sum_k \lambda_k^n \, \partial x_k^n - \partial G^n \right\} = 0. \tag{9}$$

This must be satisfied for arbitrary and independent values of the dx_i^n, so the coefficient of each dx_i^n in the first double sum must vanish separately. In the second summation, each value of n gives the contribution arising from the variation of the adjustable parameters of a different unit, so again each term must vanish separately. Thus we have the following equations

$$\lambda_i^n = h_i^n + \sum_k \lambda_k^{\bar{n}} M_{ki}^{\bar{n}} \tag{10}$$

and

$$\sum_k \lambda_k^n \, \partial x_k^n - \partial G^n = 0. \tag{11}$$

When x_i^n is a free output, there is no unit $\bar{n}$ connected to it. Thus the second term on the right hand side of equations (10) is absent for values of n and i corresponding to a free output. Instead of modifying the form of the corresponding equations (10), however, we can achieve the desired end simply by introducing the formal definition $M_{ki}^{\bar{n}} = 0$ in cases where x_i^n is a free output.

The stationary value problem is therefore solved by choosing values of the vectors λ_i^n, x_i^n and w^n to satisfy equations (10) and (11), together with the unit equations (1). Those vectors x_i^n associated with plant feed streams are given, and form boundary conditions for equations (1). Again, the vectors λ_i^n corresponding to plant product streams are given by $\lambda_i^n = h_i^n$, since the $M_{ki}^{\bar{n}}$ vanish for free outputs, and these form boundary conditions for equations (10). It is characteristic of this type of problem that the boundary conditions for the λ_i^n and the x_i^n are given on different streams, and this makes an iterative solution procedure necessary even in the case of a plant with simple sequential structure.

Each formal equation (11) represents a set of equations involving the components of a single vector w^n associated with the n-th unit, and contains no adjustable parameters associated with other units. Thus, assuming it is possible to satisfy each separate equation

$$\sum_k \lambda_k^n \, \partial x_k^n - \partial G^n = 0$$

with a suitable choice of real values for the components of w^n, the problem has been decomposed into a set of sub-problems with the dimensionalities of the separate vectors w^n, and this has been paid

refer to notes

for by introducing the extra variables λ_i^n and equations (10).

Remembering the significance of ∂x_k^n and ∂G^n, it is seen that equations (11), regarded as equations for the components of the vectors w^n, simply state that the adjustable parameters of the n-th unit must be chosen so that

$$P^n = \sum_k \lambda_k^n x_k^n - G^n \qquad (12)$$

takes a stationary value, with constant values of all inputs to the unit. Then the solution to the problem as a whole may be stated in the pleasingly elegant form that each vector w^n must be chosen so that the sub-objective function

$$P^n = \sum_k \lambda_k^n x_k^n - G^n$$

for the corresponding unit takes a stationary value, the vectors λ_k^n and x_k^n being determined by equations (10) and (1) respectively, with the boundary conditions stated above. In this formulation it has nowhere been assumed that the dimensionality of the vectors w^n is finite, so the method can handle units with continuous ranges of adjustable parameters such as a tubular reactor with an adjustable temperature gradient. The stationary value subproblem for such a unit would, of course, be solved by the method of PONTRYAGIN [4] or SWINNERTON-DYER [5]. Indeed, if one considers the special case of a simple sequential chain of units and passes formally to the limit of an infinite number of units each generating an infinitesimal change in the stream vector, the present equations reduce to differential equations and the method becomes identical with that of PONTRYAGIN and SWINNERTON-DYER. To this extent, the Maximum Principle of PONTRYAGIN may be regarded as a special case of the relations given above, but of course the formal passage to the limit is permissible only if the limit exists, and SWINNERTON-DYER [5] has given a counter-example to show that this is not necessarily the case even in apparently innocuous situations.

The above description of the process corresponds to that given in the previous paper [1] and to a recent variational treatment of the simple sequential system by KATZ [6]. However, both in this description and in KATZ's paper, it is implicitly assumed that it is possible to find real values of the components of w^n to make P^n stationary for each value of n, and this is by no means always the case. For example, the sub-objective function P^n may be monotonic in certain components of w^n even when the objective function P has a perfectly satisfactory stationary value, and in this case it is not possible to adjust the components of w^n to make P^n stationary, so the procedure just described breaks down. This possibility will be clarified later by means of a simple example, but meanwhile the nature of the iterative process involved in the solution will be illustrated by an example for which it will be assumed that each P^n has a stationary value.

Finally, it should be remarked that the use of Lagrangian multipliers to take account of the restrictions imposed on the dx_i^n by the unit equations and the plant structure was introduced by HORN [3] in discussing a simple sequence of stirred tank reactors. The present derivation may therefore be regarded as the extension of HORN's method to systems of arbitrarily complex non-sequential structure.

APPLICATION TO A SYSTEM WITH RECYCLE

The method developed above will be illustrated by describing its application to a system which was also discussed in the earlier paper [1], namely a simple recycle configuration. The system is shown in Fig. 2, which also serves to define the notation. The objective function will be taken to be

$$P = cx_1^1 + ex^6 - \sum_n G^n \qquad (13)$$

which depends on the components of two different stream vectors.

The calculation may conveniently be started at the free output x_1^1. The sub-problem associated with unit 1 has the objective function

$$P^1 = \lambda_1^1 x_1^1 + \lambda_2^1 x_2^1 - G^1$$

according to equation (12). The row vector λ_1^1 is simply equal to c, as can be seen from equation (10) for the free output x_1^1, but a value for λ_2^1 must be assumed to start the calculation. If we also

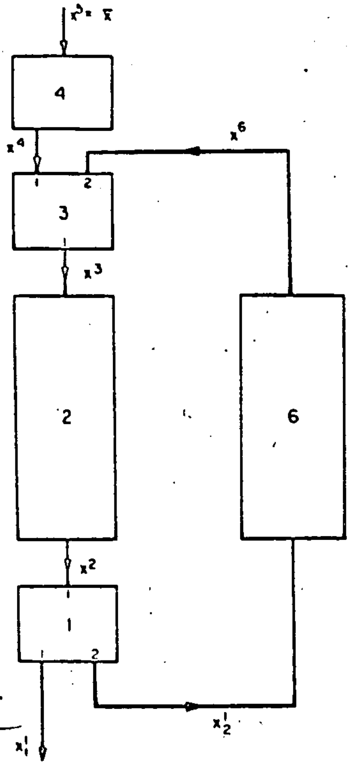


FIG. 2. Simple recycle system.

assume a value for the vector x_1^1, we must then choose w^1 so that P^1 takes a stationary value for fixed x^2, subsequently adjusting the value of x^2 until x_1^1 agrees with its assumed value. Thus the solution of the sub-porblem for unit 1 yields values of x^2 and w^1, together with the partial derivative matrices M_{11}^1 and M_{21}^1 and the vector x_2^1, from assumed values of λ_2^1 and x_1^1.

According to equation (10) we have

$$\lambda^2 = \lambda_1^1 M_{11}^1 + \lambda_2^1 M_{21}^1$$

since h^2 vanishes as x^2 does not appear in the objective function P. We may therefore proceed to choose values of x^3 and w^2 which make

$$P^2 = \lambda^2 x^2 - G^2$$

stationary and give a value of x^2 equal to that previously calculated. This in turn determines the matrix M^2 and permits λ^3 to be calculated through equations (10) giving

$$\lambda^3 = \lambda^2 M^2$$

Values of x^4, x^6 and w^3 must now be chosen which make

$$P^3 = \lambda^3 x^3 - G^3$$

stationary and give a value of x^3 equal to that found in the calculation for unit 2. This can clearly be done in an infinite number of different ways, since the total number of components of x^4 and x^6 available for adjustment will be greater than the number of components of x^3. Thus it is necessary to fix arbitrarily the values of some of the components of x^4 or x^6. If x^4, x^6 and x^3 all have the same dimensionality, for example, we could specify a value of x^4 and choose x^6 and w^3 to make P^3 stationary and x^3 equal to its previously calculated value. This, in turn, would determine the matrix M_{12}^3.

Equations (10) now permit λ^6 to be calculated, remembering that $h^6 = e$ since x^6 appears in P. Thus

$$\lambda^6 = e + \lambda^3 M_{12}^3.$$

Values for x_2^1 and w^6 must then be chosen to make

$$P^6 = \lambda^6 x^6 - G^6$$

stationary and x^6 equal to its previously determined value. These in turn determine the matrix M^6, which may be used in equations (10) to give λ_2^1

$$\lambda_2^1 = \lambda^6 M^6.$$

Finally the sub-problem for unit 4 is solved. λ^4 is given by

$$\lambda^4 = \lambda^3 M_{11}^3$$

and x^5 and w^4 are chosen to make

$$P^4 = \lambda^4 x^4 - G^4$$

stationary and x^4 equal to the value already assumed.

Inspection of the course of calculation just described shows that three mis-matches have arisen. First, values of x_2^1 have been obtained by solution of the sub-problems for both units 1 and 6, and these values will not, in general, agree. Secondly, a value of λ_2^1 is obtained from λ^6 and M^6 after solution of the sub-problem for unit 6, and this will not, in general, agree with the value for λ_2^1 assumed at the beginning of the calculation.

Thirdly, a value for x^5 is determined by solution of the sub-problem for unit 4 and this will not, in general, agree with the specified feed vector. Corresponding to these three mis-matches we have assumed values for three vectors, namely x_1^1, λ_2^1 and x^4, and these must be adjusted until the mis-matches are eliminated.

It is not difficult to check that the above calculation is identical with that described previously [1] for this system, except for a small added complication in the present case due to the presence of an extra term ex^6 in the objective function. The row vectors λ_i^n introduced here correspond to the row vectors cL_{1i}^{1n} which appeared in the previous work.

The iterative solution of the problem can, of course, be started in many different ways, of which the above is but a single example. In practice one's choice would need to be guided by consideration of the convergence of the numerical process of iterative adjustment.

Cases in which the Above Procedure May Fail

It was remarked earlier that the process as described so far will fail if any sub-objective function P^n is monotonic in one or more of the components of w^n. This is by no means impossible, even when P itself has a stationary value [7], as may be shown by a simple example.

Consider the sequence of two units shown in Fig. 3, the unit equation for each unit being inscribed in the corresponding block of the block diagram, which also serves to define the notation. All vectors in this case have just one component. With the objective function $P = x_1$, direct calculation gives

$$P = A - (x_3 + w_2)^2 - w_1^2$$

and this has the stationary maximum value A when $w_1 = 0$, $w_2 = -x_3$. Now let us attempt to find this stationary value by following the procedure described earlier in this paper. For unit 1 we have

$$\lambda_1\,\partial x_1 = \left(\frac{\partial x_1}{\partial w_1}\right)_{x_2} dw_1 = -2w_1\,dw_1 = 0$$

when $w_1 = 0$

and with this value of w_1

$$x_2 = \sqrt{(A - x_1)} \quad \text{and} \quad M_1 = \left(\frac{\partial x_1}{\partial x_2}\right)_{w_1} = -2x_2$$
$$= -2\sqrt{(A - x_1)}.$$

We then have $\lambda_2 = \lambda_1 M_1 = -2\sqrt{(A - x_1)}$, and the equation of type (11) for unit 2 is

$$\lambda_2\,\partial x_2 = \lambda_2\left(\frac{\partial x_2}{\partial w_2}\right)_{x_3} = \lambda_2 = -2\sqrt{(A - x_1)} = 0$$

which cannot be satisfied by choice of w_1, since w does not appear on the left hand side. The corresponding sub-objective for unit 2 is

$$P_2 = \lambda_2 x_2 = -2\sqrt{(A - x_1)}(x_3 + w_2)$$

which is monotonic in w_2, and has no stationary value. Thus the simple procedure breaks down at this stage.

However, the value of x_1 is still available, and we can satisfy the equation $\lambda_2\,\partial x_2 = 0$ by taking $x_1 = A$, when the left hand side vanishes for all values of w_2. w_2 then becomes an available variable in place of x_1, and it must be chosen so that the boundary conditions at inlet are satisfied. Since $x_1 = A$ implies that $x_2 = 0$, we must take $w_2 = -\bar{x}_3$, where x_3 is the specified value of $\bar{x}_3$, and we have then found the same solution as was originally obtained by inspection of the form of P.

The description of the general procedure can be extended to embrace cases such as this. For simplicity consider a simple sequence of units as shown

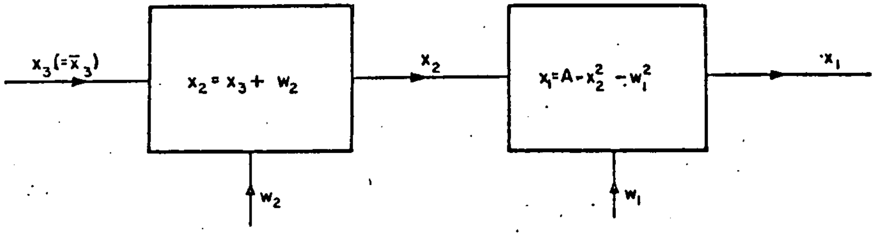


Fig. 3. A case in which the simple procedure fails.

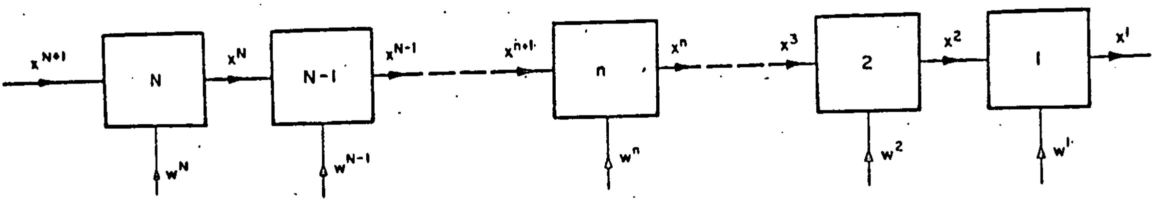


FIG. 4. Simple sequence.

in Fig. 4, which serves to define the notation. Let D be the dimensionality of the x-vectors and let each vector w^n have finite dimensionality W^n. Then at unit 1 we have $D + W^1$ variables (i.e. w^1 and x^1 or w^1 and x^2) with which to satisfy the W^1 equations.

$$\lambda^1 \, \partial x^1 - \partial G^1 = 0 \tag{14}$$

Thus D variables remain available and, in principle, it does not matter which D we choose to regard as available and which we regard as "used" in satisfying equations (14). (More generally, equations (14) confine x^1 and w^1 to a D-dimensional subspace of the $D + W^1$ dimensional product space.) Let us denote by y^1 a set of D variables regarded as still available after satisfying equations (14). Then x^2 and λ^2 can be expressed as functions of y^1, and the stationary equations for unit 2, namely

$$\lambda^2 \, \partial x^2 - \partial G^2 = 0 \tag{15}$$

have left hand sides which are functions of the $D + W^2$ variables w^2, y^1. They may be satisfied by any W^2 of these variables, the remainder being regarded as still available, and the particular W^2 variables chosen may or may not be the components of w^2.

In some cases, as in the example just worked, w^2 (or some of its components) does not appear in the left hand sides of equations (15), in which case it is *necessary* to incorporate some components of y^1 in order to have enough variables to satisfy the equations. Even when all the components of w^2 appear on the left hand sides of equations (15), it is possible that these equations cannot be satisfied for real w^2, in which case adjustment of y^1 may modify the coefficients in such a way that a real solution exists.

After the stationary value equations (15) for unit 2 have been satisfied, there still remain D variables undetermined, which we shall denote by y^2. If components of w^2 only were used in solving equations (15), y^2 will be identical with y^1, but if some components of y^1 were used in the solution these components will be absent from y^2 and will be replaced by the components of w^2 not used in solving equations (15).

In this way we can proceed to unit 3, and so on throughout the chain to unit N. After satisfying the stationary equations (11) for unit N, there will remain a D-dimensional vector y^N of undertermined variables, and adjustment of these will permit the boundary conditions at entry to be satisfied. In general y^N will comprise components of y^1 together with certain components of the w-vectors along the chain.

It is clear that the alternatives available at each stage may permit several different solutions to be obtained, in which case each will correspond to a different stationary value of P. When P has a unique stationary value, however, only one of the alternatives will permit the solution to be continued throughout the chain.

The situation described in the present section does not exhaust the possible difficulties which may arise in attempting to carry through a solution, but in many cases the simplest procedure described, using only w-vectors to satisfy the stationary value equations (11), is successful. The basic equations (1), (10) and (11) are universally applicable, and one is attempting to find a procedure which enables them to be satisfied with a minimum amount of iterative calculation.

Finally, although the method described here can be applied to a very large class of optimisation problems, it does not necessarily follow that it is always the most effective way of solving them. It is frequently possible to take advantage of certain features of the equations defining a particular problem to reduce the amount of iterative calculation far below what would be required in a blind application of the method.

CONCLUSIONS

The present formalism extends that described in an earlier paper [1] so that it can be applied to very general objective functions in plants of arbitrarily complex topological strucutre. It is also able to handle units with continuous ranges of adjustable parameters, either alone or connected to other units with discrete sets of parameters. Thus methods of reducing the dimensionality of optimization problems, which have previously been useful only in sequential structures, have been extended so as to be applicable to systems of any structure.

Acknowledgement—The author would like to thank Mr. C. L. BROOK for his invaluable and constructive criticism of the paper in draft form.

NOTATION

c Specified row vector used in forming objective functic

e Specified row vector used in forming objective functic

F_i^n Function defining unit equation for n-th unit

G^n Cost associated with adjustable variables for the n- unit

H Scalar function of the variables x_i^n appearing in tl objective function

h_i^n Partial derivative vector with components $[h_i^n]_p =$ $\dfrac{\partial H}{\partial [x_j^n]_p}$

M_{ij}^n Matrix of partial derivates with elements $[M_{ij}^n]_{pq} =$ $\dfrac{\partial [F_i^n]_p}{\partial [x_j^n]_q}$

n Number identifying a given unit

P Objective function

P^n Sub-objective function for the n-th unit

x_i^n Vector associated with the i-th output stream of t n-th unit

$x_j^{\bar{n}}$ Vector associated with the output stream connected the j-th input of the n-th unit

$x_k^{\bar{n}}$ Vector associated with the input stream connected the k-th output of the n-th unit

w^n Vector whose components are the adjustable desi and operating variables for the n-th unit.

REFERENCES

[1] JACKSON R., *Chem. Engng. Sci.* In press.

[2] ARIS R., *The Optimal Design of Chemical Reactors*, Academic Press, New York, 1961.

[3] HORN F., *Chem. Engng. Sci.* 1961 15 176.

[4] PONTRYAGIN L. S., *Usp. Mat. Nauk*, 1959 14 3.

[5] SWINNERTON-DYER H. P. F., *Proc. Lond. Math. Soc.*, 1957, 7 568.

[6] KATZ S., *Industr. Engng. Chem.* (Fundamentals) 1962 1 226.

[7] BROOK C. L., Private Communication 1963.

Résumé—L'auteur qui a déjà décrit dans un article précédent une méthode de décomposition des problèmes d'optimisation dans les ensembles topologiquement complexes, développe sa théorie pour pouvoir traiter des fonctions objectives d'un type plus général et pour inclure les unités comportant une infinité continue de paramètres ajustables, telles que les réacteurs tubulaires. Il donne une autre dérivation des formules, basée sur les multiplicateurs de Lagrange.

Reprinted from

CHEMICAL
ENGINEERING SCIENCE

Le Journal International de Génie Chimique

PERGAMON PRESS

OXFORD · LONDON · NEW YORK · PARIS

Chemical Engineering Science, 1965, Vol, 20, pp. 405-413. Pergamon Press Ltd., Oxford. Printed in Great Britain

A variational solution of unsteady state optimization problems in complex chemical plants

R. JACKSON

(University of Edinburgh and Heriot-Watt College)

(*Received* 14 *September* 1964)

Abstract—Variational methods described in previous papers for dealing with optimization problems in complex chemical plants operating under steady conditions are developed so that they may be applied to unsteady state problems of a very general nature. Previous results are shown to be special cases of the new equations, and the relation of the method to the well known method of gradients is discussed.

INTRODUCTION

IN PREVIOUS papers [1,2] a variational method of determining optimum conditions for the steady state operation of a complex plant has been developed. The method is analogous to, but weaker than, Pontryagin's Maximum Principle, and permits the over-all optimization problem to be decomposed into sub-problems, one associated with each of the units which are interconnected to form the plant. Similar results have since been re-derived by DENN and ARIS [3]. Stronger results, more strictly similar to the maximum principle, were derived by FAN and WANG [4], but it has since been shown by HORN and the present writer [5,6] that their derivation is erroneous, and their results are correspondingly untrue. DENN and ARIS [7] have also demonstrated the error of the stronger result by a method which does not differ essentially from that of HORN and JACKSON [6].

All the papers referred to above deal with the optimization of steady state operation. In the present paper it will be shown how analogous results may be obtained for time varying conditions. Results obtained earlier [1,2] will be shown to be special cases of the more general results now presented. The close relation between the weak maximum principle and the well known method of gradients will also be demonstrated.

THE DYNAMIC PROBLEM

As in earlier work [1,2] we shall consider a plant consisting of a set of units which are interconnected by process streams flowing from one to another. No restriction is placed on the nature of the network of connexions, which may be as complicated as we please, including multiple and interlocking recycle loops, cross feeds, bypass streams etc. Each unit will be assigned an identifying number n and its input and output streams will also be numbered. Thus the pair of indices (n,i) will identify the ith process stream associated with the nth unit. The physical and chemical properties characterizing each process stream may be regarded as the components of a vector x^n_i, where the suffixes n and i identify a particular unit and a particular stream. Neither is used to distinguish components of the vector, which would require a further index. In the same way, the adjustable design and operating variables for a unit may be regarded as the components of a second vector w^n. We shall further assume that the performance of a unit depends on the values of certain other parameters, for example heat transfer coefficients and catalyst activities, whose values are not freely available to the designer or operator but may change with time in a manner depending on the mode of operation of the unit. Their values at any particular time may be regarded as the components of a vector $\phi^n(t)$. Finally, each unit may be subject to disturbances quite outside the control of the operator, for example, the temperature of cooling water may change with changes in ambient temperature, and these disturbances may be regarded as the components of a time-dependent vector $d^n(t)$.

The interconnexions of the units will be specified by means of a notation introduced in an earlier paper [2], namely by listing the identifying indices of each pair of process streams which are connected together to form a link. We shall use the notation $(\bar{n},\bar{j})$ to indicate the identifying indices of the output stream which is connected to the jth input of the nth unit. A similar notation $(\tilde{n},\tilde{i})$ will be used for the identifying indices of the input stream connected to the ith output of the nth unit. The structure of interconnexions can then be specified completely by listing corresponding values of (n,i) and $(\bar{n},\bar{j})$, or alternatively corresponding values of (n,j) and $(\tilde{n},\tilde{i})$. Thus a list of the form

$$(\bar{2}, \bar{1}) \equiv (3, 2)$$
$$(\bar{5}, \bar{4}) \equiv (1, 1)$$
$$\cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot$$

would indicate that output stream number 2 of unit number 3 is connected to input stream number 1 of unit number 2; output stream number 1 of unit number 1 is connected to input stream number 4 of unit number 5, and so on. It could be written in the alternative form

$$(2, 1) \equiv (\tilde{3}, \tilde{2})$$
$$(5, 4) \equiv (\tilde{1}, \tilde{1})$$
$$\cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot$$

and either of these lists would provide a complete topological specification of the plant. The notation is illustrated by Fig. 1.

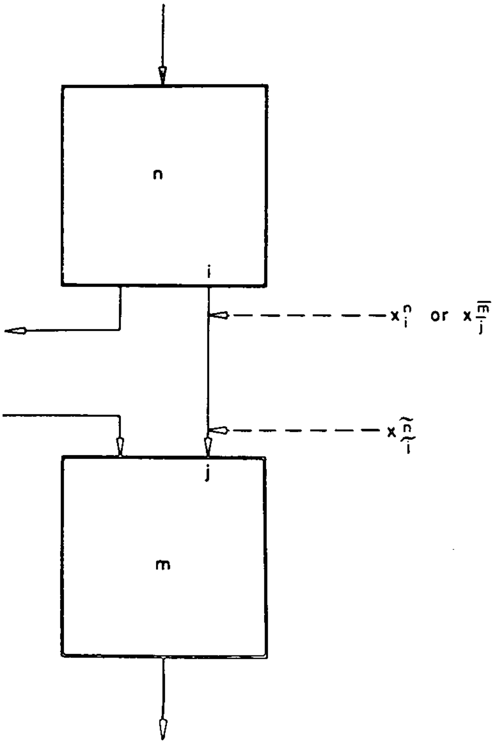

n
i
x_i^n or $x_{\bar{j}}^{\bar{m}}$
$x_{\tilde{i}}^{\tilde{n}}$
j
m

FIG. 1. Illustration of Notation

The behaviour of each unit within the plant will be completely described by differential equations of the form

$$\frac{dx_i^n}{dt} = F_i^n(x_k^n, x_{\bar{j}}^{\bar{n}}, \phi^n, d^n, w^n) \tag{1}$$

where F_i^n is a vector function of all the components of input stream vectors $x_{\bar{j}}^{\bar{n}}$ and output stream vectors x_k^n associated with this unit, as well as the parameters ϕ^n, the disturbances d^n and the design and operating variables w^n. Any set of ordinary differential equations of any order can be reduced to a set of simultaneous first-order equations of the form (1) by introducing suitably defined auxiliary variables, so the first-order form of equation (1) does not imply that the physical differential equations governing the behaviour of the unit are necessarily of the first order in their most compact form.

The parameters ϕ^n are also time dependent, and we shall assume that they change at a rate which depends on current conditions in the corresponding plant unit. Thus

$$\frac{d\phi^n}{dt} = E^n(x_{\bar{j}}^{\bar{n}}, \phi^n, d^n, w^n) \tag{2}$$

It is not difficult to see what conditions must be specified in order to determine a solution of equations (1) and (2). Clearly the initial values $x_i^n(0)$ and $\phi^n(0)$ of all plant stream variables and unit parameters must be specified, while the disturbances $d^n(t)$ and operating parameters $w^n(t)$ must be specified as functions of time throughout the interval 0 → T of interest. Certain streams enter the plant as external feeds and do not form links between pairs of units, and the corresponding stream vectors $x_i^n(t)$ must also be specified as functions of time. With these specifications equations (1) and (2), together with the list of interconnexions described above, suffice to determine the complete behaviour of the plant in the time interval 0 → T.

Now the components of the vectors $w^n(t)$ are assumed to be to some extent available. For certain of these parameters, for example, those representing design variables, it may only be possible to fix a value which is subsequently retained at all times. In other cases, for example operating variables, the parameter may be freely available for adjustment as a function of time. In either case we shall be interested in exercizing the available freedom of choice to maximize some criterion of performance defined over the time interval 0 → T. More specifically, we shall consider the problem of choosing the w^n so as to maximize an objective function of the form

$$P=\int_0^T \sum_n \left(\sum_i H_i^n(x_i^n) - G^n(w^n)\right) dt \qquad (3)$$

Where H_i^n is a specified scalar function of the variables associated with the ith output of the nth unit, and $G^n(w^n)$ represents the cost of a particular design and operating policy for the nth unit.

This general formulation clearly enables us to deal with practical problems of determining optimum start-up procedures, of investigating optimum operating policies for chemical reactors with decaying catalysts, of controlling the operation of a plant against variations in feed or externally imposed disturbances, and so on; in fact a very wide range of dynamic optimization problems.

Adjoint Equations

In this section some simple algebraic manipulations of general sets of linear algebraic equations will be set out for subsequent use in solving the optimization problem.

Suppose a set of N variables ξ_j is related to a set of N variables θ_i by linear algebraic equations of the form

$$\sum_{j=1}^{N} \alpha_{ij}\xi_j = \theta_i \qquad (i = 1, 2, \dots N) \qquad (4)$$

Consider the problem of expressing a certain linear combination of the ξ's

$$P = \sum_{k=1}^{N} c_k \xi_k \qquad (5)$$

as a linear combination of the θ's in the form

$$P = \sum_{k=1}^{N} \zeta_k \theta_k \qquad (6)$$

If equations (4) are multiplied from the left by a matrix β which is the inverse of α, we obtain

$$\sum_{i=1}^{N} \sum_{j=1}^{N} \beta_{ki}\alpha_{ij}\xi_j = \sum_{i=1}^{N} \beta_{ki}\theta_i$$

but since

$$\sum_{i=1}^{N} \beta_{ki}\alpha_{ij} = \delta_{kj},$$

where δ_{kj} is the Kronecker delta, this becomes

$$\xi_k = \sum_{i=1}^{N} \beta_{ki}\theta_i \qquad (7)$$

If equations (4) were differential equations rather than algebraic equations and i were a continuous variable such as time, β_{ki} would be a Greens function, so the inverse matrix of coefficients in these algebraic equations is the analogue of the Greens function for differential equations.

From equations (5) and (7)

$$P = \sum_{i=1}^{N} \left(\sum_{k=1}^{N} c_k \beta_{ki}\right)\theta_i$$

and comparing this with equation (6), it is seen that

$$\zeta_i = \sum_{k=1}^{N} c_k \beta_{ki}$$

Multiplying from the right by the matrix α, this then gives

$$\sum_{i=1}^{N} \zeta_i \alpha_{ij} = c_j \qquad (8)$$

We shall call the ζ_i the *adjoints* of the variables ξ_i and equations (8) are said to be adjoint to equations (4). The desired form (6) for P is therefore obtained by solving the adjoint equations and using the adjoint variables in equation (6).

These elementary general results, applicable to any set of linear algebraic equations, will be used in the next sections to develop a solution of the dynamic optimization problem.

Expression of the variation in P in terms of variations in the adjustable parameters

Consider the effect of a small change in the adjustable parameters of the units from $w^n(t)$ to

$w^n(t) + \delta w^n(t)$. This induces corresponding increments δx_i^n and $\delta\phi^n$ in the vectors x_i^n and ϕ^n, where the increments, if small, are related by the incremental forms of equations (1) and (2), namely

$$\frac{d}{dt}\,\delta x_i^n = \sum_k \mathscr{F}_{ik}^n \delta x_k^n + \sum_j F_{ij}^n \delta x_j^{\bar{n}} + K_i^n \delta\phi^n + f_i^n \delta w^n \quad (9)$$

and

$$\frac{d}{dt}\,\delta\phi^n = \sum_j E_j^n \delta x_j^{\bar{n}} + J^n \delta\phi^n + e^n \delta w^n \qquad (10)$$

where $\mathscr{F}_{ik}^n, F_{ij}^n, K_i^n, f_i^n, E_j^n, J^n$ and e^n

are matrices of partial derivatives with the following elements

$$\left.\begin{aligned} (\mathscr{F}_{ik}^n)_{pq} &= \frac{\partial(F_i^n)_p}{\partial(x_k^n)_q}\,; (F_{ij}^n)_{pq} = \frac{\partial(F_i^n)_p}{\partial(x_j^{\bar{n}})_q} \\ (K_i^n)_{pb} &= \frac{\partial(F_i^n)_p}{\partial(\phi^n)_b}\,;(f_i^n)_{pr} = \frac{\partial(F_i^n)_p}{\partial(w^n)_r} \end{aligned}\right\} \qquad (11)$$

and

$$(E_j^n)_{aq} = \frac{\partial(E^n)_a}{\partial(x_j^{\bar{n}})_q}\,; (J^n)_{ab} = \frac{\partial(E^n)_a}{\partial(\phi^n)_b}\,; (e^n)_{ar} = \frac{\partial(E^n)_a}{\partial(w^n)_r} \quad (12)$$

(The extra suffixes p, q, a, b and r in these expressions serve to distinguish components.)

The increment in the objective function P is seen from equation (3) to be

$$\delta P = \int_0^T \sum_n \left(\sum_j h_i^n \delta x_i^n - g^n \delta w^n\right) dt \qquad (13)$$

where h_i^n and g^n are row vectors with components

$$(h_i^n)_p = \frac{\partial H_i^n}{\partial(x_i^n)_p} \quad \text{and} \quad (g^n)_r = \frac{\partial G^n}{\partial(w^n)_r} \qquad (14)$$

Though there are certainly more elegant methods of dealing with these equations, perhaps the most straightforward procedure at this point is to take finite differences with respect to time, thus reducing equations (9) and (10) to a set of simultaneous algebraic equations involving the variations δx_i^n and $\delta\phi^n$ at different instants of time. The expression (13) for δP then reduces to a sum and we can use the general formulae of Section 3 to express this sum in terms of the independent variations δw^n. Writing $t = s\delta t$, where δt is a small increment in time which we shall later allow to tend to zero, and taking the simplest finite difference approximations to differential and integral operators, equations (9), (10) and (13) become

$$\begin{aligned} \delta x_i^{ns} - \delta x_i^{n,s-1} - \delta t \sum_k \mathscr{F}_{ik}^{ns} \delta x_k^{ns} - \delta t \sum_j F_{ij}^{ns} \delta x_j^{\bar{n}s} - \\ - \delta t K_i^{ns} \delta\phi^{ns} = \delta t f_i^{ns} \delta w^{ns} \quad (15) \end{aligned}$$

$$\begin{aligned} \delta\phi^{ns} - \delta\phi^{n,s-1} - \delta t J^{ns} \delta\phi^{ns} - \delta t \sum_j E_j^{ns} \delta x_j^{\bar{n}s} \\ = \delta t e^{ns} \delta w^{ns} \quad (16) \end{aligned}$$

and

$$\delta P = \sum_{s=1}^{S} \sum_n \left(\sum_i h_i^{ns} \delta x_i^{ns} - g^{ns} \delta w^{ns}\right) \delta t \qquad (17)$$

where $T = S\delta t$.

Equations (15) and (16) are a set of simultaneous linear algebraic equations in the variables δx_i^{ns} and $\delta\phi^{ns}$ precisely analogous to equations (4) in the variables ξ_j. Again, the first term on the right-hand side of equation (17), namely

$$\delta P_1 = \sum_{s=1}^{S} \sum_n \sum_i \delta t h_i^{ns} \delta x_i^{ns} \qquad (18)$$

is a linear combination of variables precisely analogous to the right-hand side of equation (5). Thus, writing down adjoint equations for the present case corresponding to equations (8) of the general case, we can determine the coefficients in an expression of δP_1 as a linear combination of the variables δw^{ns} by comparison with equation (6).

We shall introduce variables λ_i^{ns} and μ^{ns} adjoint to δx_i^{ns} and $\delta\phi^{ns}$ respectively. To write down adjoint equations corresponding to equations (8) we then have to pick out the coefficients corresponding to the α_{ij} of equations (4). Now for a given value of j, the α_{ij} are the coefficients of all the terms in ξ_j which appear on the left-hand sides of equations (4). Similarly, to write an equation adjoint to one of equations (15) or (16), we must pick out all the terms in which a given variable δx_i^{ns} or $\delta\phi^{ns}$ respectively appears on the left-hand sides of these equations. The sum of the coefficients of these terms multiplied by the corresponding adjoint variables then forms the left-hand side of the corresponding adjoint equation, as in (8).

Let us consider first a variable δx_i^{ns}, picking out those terms on the left-hand sides of equations (15) and (16) in which it appears. If we also write down the right-hand sides of the corresponding equations so as to identify these equations, we obtain the following terms, provided (n,i) are not the identifying indices of a "free "output stream unconnected to any other unit, and provided also that $s \neq S$.

(i) A term $(1 - \delta t\, \mathscr{F}^{ns}_{ii})\delta x^{ns}_i$
in an equation with r.h.s $\delta t\, f^{ns}_i \delta w^{ns}$

(ii) Terms $-\delta t \mathscr{F}^{ns}_{ki}\delta x^{ns}_i$, $(k \neq i)$
in equations with r.h.s.'s $\delta t f^{ns}_k \delta w^{ns}$

(iii) Terms $-\delta t F^{\tilde{n}s}_{j\bar{\imath}}\delta x^{ns}_i$, $(j = 1,2,\ldots)$
in equations with r.h.s.'s $\delta t f^{\tilde{n}s}_j \delta w^{\tilde{n}s}$

(iv) A term $-\delta t E^{\tilde{n}s}_{\bar{\imath}}\delta x^{ns}_i$
in an equation with r.h.s. $\delta t e^{\tilde{n}s}\delta w^{\tilde{n}s}$

(v) A term $-\delta x^{ns}_i$
in an equation with r.h.s. $\delta t f^{n,s+1}_i \delta w^{n,s+1}$

When (n,i) is a "free" output not connected to any other unit, only the terms (i), (ii) and (v) appear, and when $s = S$ there is no term of the form (v).

In a similar manner we can pick out those terms on the left-hand sides of equations (15) and (16) in which a particular variable $\delta\phi^{ns}$ appears, again identifying the corresponding equations by noting their right-hand sides. Provided $s \neq S$ this gives the following terms

(i) Terms $-\delta t K^{ns}_i \delta\phi^{ns}$, $(i = 1,2,\ldots)$
in equations with r.h.s.'s $\delta t f^{ns}_i \delta w^{ns}$

(ii) A term $(1 - \delta t J^{ns})\delta\phi^{ns}$
in an equation with r.h.s. $\delta t e^{ns}\delta w^{ns}$

(iii) A term $-\delta\phi^{ns}$
in an equation with r.h.s. $\delta t e^{n,s+1}\delta w^{n,s+1}$

However, when $s = S$ there is no term of the form (iii).

Having picked out these terms, and remembering that equations (18) correspond to the general equations (5), the adjoint equations can now be written down directly by analogy with equations (8). They are

$$\lambda^{ns}_i(1 - \delta t\mathscr{F}^{ns}_{ii}) - \delta t \sum_{k\neq i} \lambda^{ns}_k \mathscr{F}^{ns}_{ki} - \lambda^{n,s+1}_i - \\ - \delta t \sum_j \lambda^{\tilde{n}s}_j F^{\tilde{n}s}_{j\bar{\imath}} - \delta t \mu^{\tilde{n}s} E^{\tilde{n}s}_{\bar{\imath}} = \delta t h^{ns}_i \qquad (19)$$

when (n,i) is not a free output and $s \neq S$.

$$\lambda^{ns}_i(1 - \delta t\mathscr{F}^{ns}_{ii}) - \delta t \sum_{k\neq i} \lambda^{ns}_k \mathscr{F}^{ns}_{ki} - \lambda^{n,s+1}_i = \delta t h^{ns}_i \qquad (20)$$

when (n,i) is a free output and $s \neq S$

$$\lambda^{ns}_i(1 - \delta t\mathscr{F}^{ns}_{ii}) - \delta t \sum_{k\neq i} \lambda^{ns}_k \mathscr{F}^{ns}_{ki} - \delta t \sum_j \lambda^{\tilde{n}s}_j F^{\tilde{n}s}_{j\bar{\imath}} - \\ - \delta t \mu^{\tilde{n}s} E^{\tilde{n}s}_{\bar{\imath}} = \delta t h^{ns}_i \qquad (21)$$

when (n,i) is not a free output and $s = S$

$$\lambda^{ns}_i(1 - \delta t\mathscr{F}^{ns}_{ii}) - \delta t \sum_{k\neq i} \lambda^{ns}_k \mathscr{F}^{ns}_{ki} = \delta t h^{ns}_i \qquad (22)$$

when (n,i) is a a free output and $s = S$

Together with

$$\mu^{ns}(1 - \delta t J^{ns}) - \mu^{n,s+1} - \delta t \sum_i \lambda^{ns}_i K^{ns}_i = 0 \qquad (23)$$

when $s \neq S$, and

$$\mu^{ns}(1 - \delta t J^{ns}) - \delta t \sum_i \lambda^{ns}_i K^{ns}_i = 0 \qquad (24)$$

when $s = S$

We also have, by comparison with equation (6)

$$\delta P_1 = \sum_{s=1}^{S} \delta t \sum_n \left(\sum_i \lambda^{ns}_i f^{ns}_i + \mu^{ns} e^{ns}\right)\delta w^{ns} \qquad (25)$$

The required results can finally be obtained by passing to the limit $\delta t \to 0$, when equations (19) to (24) reduce to

$$\frac{d\lambda^n_i}{dt} + \sum_k \lambda^n_k \mathscr{F}^n_{ki} + \sum_j \lambda^{\tilde{n}}_j F^{\tilde{n}}_{j\bar{\imath}} + \mu^{\tilde{n}} E^{\tilde{n}}_{\bar{\imath}} = -h^n_i \qquad (26)$$

valid when (n,i) is not a free output, together with

$$\frac{d\lambda^n_i}{dt} + \sum_k \lambda^n_k \mathscr{F}^n_{ki} = -h^n_i \qquad (27)$$

valid when (n,i) is a free output, with the boundary conditions

$$\lambda^n_i(T) = 0 \qquad [\text{all } (n, i)] \qquad (28)$$

and

$$\frac{d\mu^n}{dt} + \mu^n J^n + \sum_i \lambda^n_i K^n_i = 0 \qquad (29)$$

with boundary condition

$$\mu^n(T) = 0 \qquad (\text{all } n) \qquad (30)$$

In equations (26) to (30), the suffix s has been dropped, since all quantities are understood to depend on the continuous variable t. The boundary conditions (28) are obtained from equations (21) and (22) on passing to the limit $\delta t \to 0$, and the conditions (30) follow in a similar way from equation (24).

On passing to the limit in equation (25) and incorporating, once more, the second term on the right hand side of equation (13), the variation in P is seen to be given by

$$\delta P = \int_0^T \sum_n \left(\mu^n e^n + \sum_i \lambda_i^n f_i^n - g^n\right) \delta w^n \, dt \quad (31)$$

which expresses δP, as required, in terms of the variations $\delta w^n(t)$ in the available adjustable parameters.

SOLUTION OF THE OPTIMIZATION PROBLEM

If each of the functions $w^n(t)$ is freely available for variation, necessary condition for P to take a stationary maximum value is that δP should vanish for all variations $\delta w^n(t)$; in other words that

$$\mu^n e^n + \sum_i \lambda_i^n f_i^n - g^n = 0$$
$$(n = 0, 1, 2 \dots) \text{ (all } 0 \leqslant t \leqslant T) \quad (32)$$

which represents a set of equations to determine the components of w^n. Remembering the definitions of f_i^n, e^n and g^n given in equations (11), (12) and (14), the condition (32) may be expressed in the following alternative form.

"P will take a stationary value if and only if each of the quantities

$$P^n(t) = \mu^n E^n + \sum_i \lambda_i^n F_i^n - G^n \quad (33)$$

is stationary with respect to variations in $w^n(t)$ at each value of t, the variables λ_i^n and μ^n and all other variables on which E^n and F_i^n depend, other than w^n, being regarded as constants."

It should be noted that although a stationary value of P results when each P^n is stationary for all t, there is no relation, in general, between the natures of these stationary values. What we have found is therefore a "weak" maximum principle analogous to that developed earlier (1),(2) for the case of steady state operation. As in the steady state a "strong" maximum principle is not generally true [5,6].

An alternative and possibly more practical way of making use of equation (31) to solve the dynamic optimization problem is to regard it as giving the gradient of the objective function P in the function space of the set of functions $w^n(t)$. As the idea of a gradient in function space is not yet very familiar to chemical engineers, it is probably worthwhile digressing briefly to say a little about it. The concept appears to have been first introduced into chemical engineering by HORN [8], but it has also been used by KELLEY, BRYSON and other workers in the field of flight path control [9,10]. However, the use of the gradient in function space in handling variational problems was described a good deal earlier in the well known textbook of COURANT (11).

The idea may be introduced by comparing a function of many variables

$$P_1 = F(x_1, x_2, \dots x_n) \quad (34)$$

with an integral whose integrand depends on a function of the variable of integration.

$$P_2 = \int_0^T G[t, x(t)] \, dt \quad (35)$$

P_1 is a function of the finite set of variables x_1, $x_2, \dots x_n$, while P_2 may be regarded as a function of an infinite set of variables, namely the values of $x(t)$ at each value of t. The discrete valued parameter i which distinguishes different variables x_i of the finite set then corresponds to the continuous valued parameter t which distinguishes different variables $x(t)$ of the infinite set. In the finite case one often uses geometrical language, saying that P_1 is a function of position in the space of the variables $x_1, x_2, \dots x_n$, and by analogy we can say that P_2 is a function of position in the "function space" of $x(t)$. This will be a "space" with an infinite number of dimensions, each point of which corresponds to a particular function $x(t)$ defined in $0 \leq t \leq T$.

Now consider small variations δx_i in the variables x_i of equation (34) and $\delta x(t)$ in the function $x(t)$ of equation (35).

We can then write

$$\delta P_1 = f_1 \delta x_1 + f_2 \delta x_2 + \dots + f_n \delta x_n \quad (36)$$

where the numbers $(f_1, f_2, \dots f_n)$ are the components of the gradient of P_1 in the space of the variables $x_1, x_2, \dots x_n$. They have the property that, for all displacements of equal magnitude (i.e.$(\delta x_1)^2 + (\delta x_2)^2 + \dots + (\delta x_n)^2 = \text{const.}$), that for which each δx_i is proportional to the corresponding f_i gives the largest increase in P_1; in other words displacements $\delta x_1 = kf_1$, $\delta x_2 = kf_2, \dots \delta x_n = kf_n$ lie along the line of steepest ascent of P_1. Just as

(36) gives an incremental form of equation (34), it may be possible to express δP_2, corresponding to the variation $\delta x(t)$, in the form

$$\delta P_2 = \int_0^T g[t, x(t)]\delta x(t)\, dt \qquad (37)$$

Comparing equations (36) and (37), and regarding integration in (37) as analogous to summation in (36), it is seen that the function $g[t, x(t)]$ could appropriately be called the gradient of P_2 in the function space of $x(t)$. It is not difficult to show that, if we regard as of equal magnitude all variations $\delta x(t)$ for which $\int_0^T[\delta x(t)]^2\, dt$ is the same, then the largest increase in P_2 results from that variation in which $\delta x(t)$ is proportional to $g[t, x(t)]$ at each value of t. In other words, if we use the above quadratic measure of the "length" of a displacement in the function space, then displacements

$$\delta x(t) = kg[t, x(t)]$$

lie in the direction of steepest ascent of P_2.

Now we are interested in the variational problem of finding a function $x(t)$ which maximizes P_2, and the above suggests that this might be attacked by improving an initial guess $x_0(t)$ according to a steepest ascent procedure, replacing $x_0(t)$ by

$$x_1(t) = x_0(t) + kg[t, x(t)]$$

and recalculating the gradient g after proceeding some distance along this line. Such a procedure is strictly analogous to the well known method of steepest ascents for maximizing a function of a finite number of variables.

Returning now to equation (31), we see that it is essentially of the same form as equation (37) except that many functions of time, namely all the components of the vectors $w^n(t)$, are involved rather than the single function $x(t)$. However, it is still true to say that the functions

$$p^n(t) = \mu^n e^n + \sum_i \lambda_i^n f_i^n - g^n \qquad (38)$$

represent the gradient of P, in the sense just described, in the space of the functions $w^n(t)$. One could therefore start from a guess $w_0^n(t)$ at the adjustable parameters and obtain the greatest increase in P for a small modification of the guess by taking

$$w^n(t) = w_0^n(t) + k\Big[\mu^n e^n + \sum_i \lambda_i^n f_i^n - g^n\Big] \qquad (39)$$

As k is increased, one moves up the line of steepest ascent at the point corresponding to the initial guess $w_0^n(t)$, and at any stage the functions $w_1^n(t)$, given by equations (39) with $k = k_1$, may serve as a basis for a new estimate of the gradients (38). One may then proceed up the new steepest ascent line, continuing the process of alternate climbing and recalculation of the gradient, until P is no longer significantly increased.

It remains to indicate how the plant equations and adjoint equations may be solved in order to compute the gradient (38) corresponding to any particular set of values of the functions $w^n(t)$. With the postulated form of the functions $w^n(t)$, the given initial conditions $x_i^n(0)$ and $\phi^n(0)$ and the boundary conditions specified for those x_i^n which correspond to external feeds to the plant, the plant equations (1) and (2) may be integrated forwards in time to determine all the variables x_i^n and ϕ^n throughout the interval $0 \leq t \leq T$. The adjoint equations (26), (27) and (29) may then be integrated backwards in time from the terminal conditions (28) and (30). This is possible since the coefficients in these equations are determined once the variables x_i^n and ϕ^n have been found. The adjoint variables λ_i^n and μ^n are then known for all t, and the solution of the plant equations determines the quantities e^n and f_i^n for all t, so the required gradient can be calculated from equation (38) at each value of t.

Earlier, in describing the problem, it was recognized that certain of the components of the vectors $w^n(t)$ might not be independently adjustable at all values of t. For example, if a particular vector w^n consists entirely of design variables for the nth unit, the values of these variables must be chosen once for all, and cannot be made to depend on time. In this case δw^n in equation (31) is independent of time and can be taken outside the integral so that the corresponding contribution to δP is

$$\delta w^n \int_0^T \Big(\mu^n e^n + \sum_i \lambda_i^n f_i^n - g^n\Big)\, dt$$

and the components of the gradient corresponding to the components of w^n are simply

$$\int_0^T \Big(\mu^n e^n + \sum_i \lambda_i^n f_i^n - g^n\Big)\, dt \qquad (39)$$

which can be computed when the plant equations and adjoint equations have been solved. In the optimizing adjustments the changes δw^n are then made proportional to the quantities (39), and are independent of time.

THE STEADY STATE PROBLEM

Finally we shall show how the equations derived in the present paper degenerate into those previously obtained [1,2] when conditions in the plant do not vary with time. In the case of steady state operation, the process stream vectors are related by algebraic equations of the form

$$x_i^n = R_i^n(x_j^{\bar{n}}, w^n) \tag{40}$$

corresponding to equations (1) in the earlier treatment of the steady state [2], with a slight change in notation to avoid confusion. The problem is then to choose (constant) values for the vectors w^n which will maximize an objective function of the form

$$P_s = T \sum_n \left(\sum_i H_i^n(x_i^n) - G^n(w^n) \right) \tag{41}$$

where the constant factor T has been introduced to make P_s formally identical with P of the present paper.

The simplest way to derive the solution of this problem as a special case of the general dynamical problem is to replace equations (40) formally by dynamic equations of the form

$$\frac{dx_i^n}{dt} = R_i^n(x_j^{\bar{n}}, w^n) - x_i^n \tag{42}$$

The steady state solution of these equations, obtained by setting $dx_i^n/dt = 0$, clearly satisfies equations (40). There are no variables of the type ϕ^n in this case so the adjoint equations, whose general form is given by (26), reduce to

$$\frac{d\lambda_i^n}{dt} - \lambda_i^n + \sum_j \lambda_j^{\bar{n}} M_{ji}^{\bar{n}} = -h_i^n \tag{43}$$

where

$$[M_{ij}^n]_{pq} = \frac{\partial (R_i^n)_p}{\partial (x_j^{\bar{n}})_q}$$

The variation in the objective function is given by equation (31), which simplifies to

$$\delta P = \int_0^T \sum_n \left(\sum_i \lambda_i^n N_i^n - g^n \right) \delta w^n \, dt \tag{44}$$

where

$$(N_i^n)_{pr} = \frac{\partial (R_i^n)_p}{\partial (w^n)_r}$$

Now if there is no time variation, we can set $dx_i^n/dt = d\lambda_i^n/dt = 0$ in equations (42) and (43), and the integrand of equation (44) is independent of time. Thus these equations reduce to

$$x_i^n = R_i^n({}_j^{\bar{n}}x, w^n) \tag{40}$$

$$\lambda_i^n = h_i^n + \sum_j \lambda_j^{\bar{n}} M_{ji}^{\bar{n}} \tag{45}$$

$$\delta P = T \sum_n \left(\sum_i \lambda_i^n N_i^n - g^n \right) \delta w^n \tag{46}$$

Equations (40) are the steady state plant equations, equations (45) are indentical with the adjoint equations for the steady state problem previously derived [2], and equation (46) yields the following condition for a stationary value of P

$$\sum_i \lambda_i^n N_i^n - g^n = 0 \qquad (n = 1, 2, \ldots) \tag{47}$$

which is the same as that found in the earlier work [1,2]. Thus we have demonstrated that the formalism of the present paper includes the steady state formalism presented in previous papers as a special case.

CONCLUSION

The methods developed in the present paper are applicable to a very wide range of practical dynamical optimization problems. In particular one might mention the determination of optimum start-up conditions, the optimum operation of catalytic reactors with decaying catalysts, the optimum adjustment of operating conditions to compensate for variations in feedstock and the control of plants against time dependent external disturbances. Some of these particular problems will be examined in more detail in subsequent publications.

REFERENCES

[1] JACKSON R., *Chem. Engng. Sci.* 1964 **19** 9.
[2] JACKSON R., *Chem. Engng. Sci.* 1964 **19** 253.
[3] DENN M. M. and ARIS R., *Amer. Inst. Chem. Engrs. J.* In press.
[4] FAN L. T. and WANG C. S., *J. Electron Control* 1964 **16**, 189.
[5] HORN F. and JACKSON R., *Industr. Engng. Chem.* (*Fundamentals*). 1965.
[6] HORN F. and JACKSON R. J., *Electron. Control.* In press.
[7] DENN M. M. and ARIS R. Private communication. 1964.
[8] HORN F. *Chem. Ing. Tech.* 1960 **32** 382.
[9] KELLEY H. J. in *Optimization Techniques* 1962 (Edited by LEITMANN G.) Academic Press, New York.
[10] BRYSON A. E. and DENHAM W. F., *Trans Amer. Soc. Mech. Engrs.* (*J. Appl. Mech.*) 1962 **29**, 2.
[11] COURANT R. and HILBERT D., *Methods of Mathematical Physics* 1953 Vol. I. Interscience, New York.

Group C

In many optimization problems of practical interest there is a finite and often small number of variables whose values are available to be adjusted. However, this is not always the case, and sometimes the available quantity may be a function of one or more independent variables. For example, in time-dependent optimization problems the form of certain variables as functions of time may be available to be adjusted. Again, in tubular chemical reactors it may, in principle, be possible to adjust the temperature as a function of position along the tube. Mathematically such problems belong to the Calculus of Variations, and in chemical engineering systems they most frequently arise in a form conveniently attacked by the Maximum Principle of Pontryagin, which is a result in the Calculus of Variations. The publications of this group are concerned with such problems.

The solution of a variational problem by means of the Maximum Principle is by no means a straightforward procedure as this principle gives only a necessary condition for optimality. Thus mathematical structures which satisfy the Principle are not necessarily solutions of the problem, and one sometimes finds that the Maximum Principle may be satisfied in more than one way, even though the problem itself has a unique solution. Such difficulties are frequently regarded as mere mathematical curiosities by engineers, but in publication C1 it is shown that they arise in acute form in a very well known and apparently simple problem in chemical reaction engineering. Indeed when this problem was first analysed in 1956, the solution given was incorrect in certain cases for this very reason.

Publication C2 deals with the problem of choosing the control variables in a chemical reactor, as functions of time, to bring the system into its final operating state in the most economic possible way. The system treated is an exothermic reaction operated autothermically with recycle of heat to the feed stream/

stream. This has an interesting ignition phenomenon with effects which are reflected in the optimum startup procedure, and is also of practical interest since it represents a simple mathematical model of a Haber ammonia synthesis reactor.

Publications C3 and C4 are concerned with a problem in which the adjustable variable, the local temperature in a tubular reactor, is a function of two independent variables, namely time and position in the reactor. In catalytic reactors the activity of the catalyst packing frequently decays in use and the rate of decay depends, among other things, on the temperature. Thus the choice of temperature profile in the reactor must at all times be a compromise between securing the highest instantaneous yield of the desired product and preserving the activity of the catalyst. The problem is therefore to select the optimum temperature, as a function of position and time throughout the length of the reactor and the life of the catalyst, to secure the largest total yield of the desired product. The principal difficulty here is to develop a practical computational procedure to approximate the solution of an apparently very complex problem. The basic theory of such a procedure and early attempts to implement it are described in publication C3, while publication C4 reports the finally successful computational method and its results.

In the course of this work it became apparent that variational problems in two independent variables,with hyperbolic partial differential equations as side conditions, raise some interesting mathematical questions. These are taken up in publications C5 and C6 which develop a first order variational theory and a maximum principle respectively for problems of this class. The principal new result of interest arises when the boundary of the domain of interest in the plane of the independent variables includes finite segments parallel/

parallel to the characteristics of the hyperbolic equations. These generate interesting and unexpected singularities in the solution within the domain.

When two successive chemical reactions are required to convert a feedstock into a final product, and when each reaction is catalysed by a different catalyst, it is common industrial practice to carry out the reaction in two separate stages in physically distinct reactors. The first is used to convert the feedstock to the intermediate product and the second to convert the intermediate to the final product, and each contains its respective catalyst. In 1965 Gunn and Thomas suggested that it may be better to blend the two catalysts in a single reactor, and investigated the optimum blend. In publication C7 a generalisation of Gunn and Thomas's problem is investigated. The proportions of the two catalysts present in the blend is now permitted to vary from point to point along the reactor in any way, and the problem is to find the optimum catalyst blend profile as a function of position in the reactor. For the simplest class of two successive reactions it turns out that this problem can be solved explicitly without recourse to numerical computation and the result is rather simple. The optimum policy is to use a reactor of a certain length containing only the first catalyst, followed by a second section of determined length containing a uniform blend of the two catalysts in determined proportions, and a terminal section containing only the second catalyst.

Finally, in publication C8, the present writer and a number of other workers draw together their joint and separate investigations of a system which has now been very thoroughly explored, namely the single exothermic reversible reaction.

C1

Reprinted from

*CHEMICAL
ENGINEERING SCIENCE*

Le Journal International de Génie Chimique

PERGAMON PRESS

OXFORD · LONDON · NEW YORK · PARIS

Chemical Engineering Science, 1965, Vol. 20, pp. 911–920. Pergamon Press Ltd., Oxford. Printed in Great Britain

Optimum temperature profiles in tubular reactors: an exploration of some difficulties in the use of Pontryagin's Maximum Principle

I. COWARD and R. JACKSON

University of Edinburgh and Heriot-Watt College

(*Received* 21 *December* 1965)

Abstract—By considering the well known problem of determining optimum temperature profiles for the successive first order reactions

$$A \xrightarrow{1} B \xrightarrow{2} C$$

carried out in a tubular reactor it is shown that the application of Pontryagin's Maximum Principle is not straightforward, even in a case as simple as this. In particular, it is stressed that the Maximum Principle provides necessary but not sufficient conditions for optimum operation, and that the distinction is of more than academic importance.

INTRODUCTION

As a result of current interest in optimisation problems the Maximum Principle of Pontryagin has recently received a great deal of attention. Experience [1] has shown that there may be computational difficulties in obtaining a solution to a given problem which satisfies Pontryagin's condition, but quite apart from these there are mathematical difficulties of a more fundamental type. It is well known [4] that Pontryagin's principle provides only a necessary condition for the maximisation or minimisation of the quantity of interest, so there exists the possibility of finding solutions which satisfy Pontryagin's condition but do not maximise or minimise this quantity. It is the purpose of this paper to illustrate by means of a well known example the sort of difficulties which can arise.

We shall consider the system of two successive first order reactions

$$A \xrightarrow{1} B \xrightarrow{2} C$$

originally studied by BILOUS and AMUNDSON [2]. This example is sufficiently simple for the structure of the possible solutions to be deducible largely by general reasoning without resort to detailed numerical work, and it will be seen that attempts at a direct numerical solution by the usual methods would be unlikely to succeed in the absence of the information obtainable by general reasoning. There is no reason to suppose that the difficulties revealed in this example are not present in more complicated problems, where they can no longer be elucidated by the type of general reasoning applied here. Consequently the success of a blind numerical attack, the only method available in such cases, is doubtful even if it runs into no purely numerical difficulties of the type discussed by ROSENBROCK and STOREY [1].

In their study of the reactions $A \rightarrow B \rightarrow C$, BILOUS and AMUNDSON [2] sought temperature profiles which would maximise the yield of substance B in a tubular reactor of given length.

Rather surprisingly the optimum temperature profile was found to be independent of the relative activation energies E_1 and E_2 of the two reactions, and to decrease monotonically from inlet to outlet, both for $E_1 < E_2$ and $E_1 > E_2$. This is physically reasonable when $E_1 < E_2$ but less reasonable when $E_1 > E_2$, and the authors expressed some reservations about the validity of their result in the latter case. This point was later examined, using the method of dynamic programming, by ARIS [3], who showed that their doubts were well founded since the profile derived was not optimal when $E_1 > E_2$. In this paper we shall use the Maximum Principle to re-examine the problem of maximising the exit concentration of B, and we shall also consider the conditions necessary to minimise this quantity. Apart from its value in illustrating the difficulties of the Maximum Principle in a simple way, this second problem is of some physical interest in the case where the object of the reaction is to produce pure C, and any unreacted A may be separated and recycled.

PONTRYAGIN'S MAXIMUM PRINCIPLE [4]

For the sake of continuity we shall use the notation of HORN [5]. Pontryagin's Maximum Principle solves the following problem: given the set of simultaneous differential equations

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = v_i(x_1, x_2 \ldots x_n, T) \qquad i = 1, 2, \ldots n \tag{1}$$

with specified initial conditions

$$x_i(0) = a_i \qquad i = 1, 2, \ldots n \tag{2}$$

choose $T(t)$ in the interval $0 \leqslant t \leqslant \theta$ so that

$$P = \sum_{i=1}^{n} c_i x_i(\theta) \tag{3}$$

is maximised (minimised), where the c_i are given constants.

The solution requires that we solve simultaneously the set of n differential equations

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = v_i(x_1, x_2 \ldots x_n, T) \qquad i = 1, 2 \ldots n \tag{1}$$

together with their adjoints

$$\frac{\mathrm{d}\lambda_i}{\mathrm{d}t} = - \sum_{j=1}^{n} \frac{\partial v_j}{\partial x_i} \lambda_j \qquad i = 1, 2 \ldots n \tag{4}$$

with the boundary conditions

$$x_i(0) = a_i \qquad i = 1, 2 \ldots n \tag{2}$$

$$\lambda_i(\theta) = c_i \qquad i = 1, 2 \ldots n. \tag{5}$$

If P is to be maximised (minimised) it is then necessary that $T(t)$ should be so chosen that, for each t, the Hamiltonian H is maximised (minimised) where

$$H = \sum_{j=1}^{n} \lambda_j v_j \tag{6}$$

It is apparent that our problem—the maximisation (minimisation) of the exit concentration of B with a given reaction time θ—is of this form. The components of the vector $\mathbf{x}$ are given in this case by $x_1 = a$, $x_2 = b$, where a and b are the concentrations of A and B respectively, and the differential equations governing them are

$$\frac{\mathrm{d}a}{\mathrm{d}t} = v_1(a, b, T) = -k_1 a \tag{7a}$$

$$\frac{\mathrm{d}b}{\mathrm{d}t} = v_2(a, b, T) = k_1 a - k_2 b \tag{7b}$$

with the boundary conditions at inlet

$$a(0) = a_0 \tag{8a}$$

$$b(0) = b_0 \tag{8b}$$

The reaction velocity constants are assumed to be given by the Arrhenius expression

$$k_i = p_i e^{-E_i/RT}$$
$$= p_i e^{-e_i/T}$$

($i = 1$, 2) and t represents residence time from the instant of entry to the reactor.

The differential equations governing the adjoint variables are

$$\frac{\mathrm{d}\lambda_1}{\mathrm{d}t} = -\lambda_1 \frac{\partial v_1}{\partial x_1} - \lambda_2 \frac{\partial v_2}{\partial x_1} = k_1\lambda_1 - k_1\lambda_2 \tag{9a}$$

$$\frac{\mathrm{d}\lambda_2}{\mathrm{d}t} = -\lambda_1 \frac{\partial v_1}{\partial x_2} - \lambda_2 \frac{\partial v_2}{\partial x_2} = k_2\lambda_2 \tag{9b}$$

with boundary conditions at outlet

$$\lambda_1(\theta) = 0 \tag{10a}$$

$$\lambda_2(\theta) = 1 \tag{10b}$$

since b is the quantity to be maximised or minimised at exit.

The temperature at every point in the reactor is to be chosen so as to maximise (minimise) the Hamiltonian H where

$$H = \lambda_1 v_1 + \lambda_2 v_2 = \lambda_1(-k_1 a) + \lambda_2(k_1 a - k_2 b)$$
$$= k_1 a(\lambda_2 - \lambda_1) - k_2 b \lambda_2 \tag{11}$$

and λ_1, λ_2, a and b are to be regarded as constants in the maximisation (minimisation). This will be referred to as Pontryagin's maximising (minimising) condition. Pontryagin's condition is sometimes replaced by the weaker condition that the temperature be chosen everywhere so that

$$\frac{\partial H}{\partial T} = \lambda_1 \frac{\partial v_1}{\partial T} + \lambda_2 \frac{\partial v_2}{\partial T}$$
$$= \frac{k_1 e_1}{T^2} a(\lambda_2 - \lambda_1) - \frac{k_2 e_2}{T^2} b\lambda_2 = 0 \tag{12}$$

where it is understood that, if the temperature required to satisfy this equation lies outside the permissible range $T_* \leqslant T \leqslant T^*$, then the temperature at this point is either T_* or T^* depending on which inequality is violated. It is important to realise that equation (12) may give any type of stationary value of the Hamiltonian and that consequently care is needed if the weaker condition is used.

HORN [6] has pointed out that, when only two independent reactions are involved, as here, λ_1 and λ_2 can be eliminated between equations (7), (9) and (12) to give an explicit expression for $\mathrm{d}T/\mathrm{d}t$ in terms of x_1, x_2 and T. However, since this result derives from the weak condition (12) it should be used with caution. The result of following this procedure in our problem is

$$\frac{\mathrm{d}T}{\mathrm{d}t} = -\frac{k_1 T^2}{e_2} \cdot \frac{a}{b}. \tag{13}$$

A precisely equivalent result is given by BILOUS and AMUNDSON in their equation (15) and it is this differential equation which gives the falling temperature profile irrespective of the relative activation energies. The results of integrating equations (7), (8) and (13) numerically will be referred to later.

SOME PROPERTIES OF EQUATIONS (7)–(11)

For convenience in the future discussion we shall establish at this point those properties of our equations which will be used later.

LEMMA 1. *λ_2 is positive throughout the interval $0 \leqslant t \leqslant \theta$.*

Proof. It is apparent from equation (9b) that $\mathrm{d}\lambda_2/\mathrm{d}t$ and λ_2 have the same sign and therefore that λ_2 is either positive and monotonic increasing with increasing time or negative and monotonic decreasing. Since $\lambda_2 = 1$ at $t = \theta$ the result follows.

LEMMA 2. *If we define $\alpha(t) = a(\lambda_2 - \lambda_1)$ and $\beta(t) = b\lambda_2$ then both α and β increase monotonically as we go forwards along the reactor no matter how the temperature varies.*

Proof. From equations (7) and (9)

$$\frac{\mathrm{d}}{\mathrm{d}t}\{a(\lambda_2 - \lambda_1)\} = ak_2(T)\lambda_2 = \frac{\mathrm{d}\alpha}{\mathrm{d}t} \tag{14}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\{b\lambda_2\} = ak_1(T)\lambda_2 = \frac{\mathrm{d}\beta}{\mathrm{d}t} \tag{15}$$

LEMMA 3. *If a temperature profile satisfies the Pontryagin maximising (minimising) condition throughout the reactor the value of the Hamiltonian along the profile is constant.*

This is actually a general result [4] not restricted to our particular system.

LEMMA 4. *The only possible stationary values of the Hamiltonian are as follows:*

(*a*) *If $E_1 < E_2$ a single maximum at a finite value of the temperature.*

(*b*) *If $E_1 > E_2$ a single minimum at a finite value of the temperature.*

Proof. It has already been established (LEMMA 1) that β is positive (or zero if $b = 0$) and that both α and β increase with time (LEMMA 2). It is apparent that the Hamiltonian $H(T) = k_1(T)\alpha - k_2(T)\beta$ is zero when $T = 0$ and also that, assuming $\beta > 0$, it is negative and monotone decreasing with increasing temperature for $\alpha \leqslant 0$; for $\alpha > 0$ we must investigate the two cases $E_1 < E_2$ and $E_1 > E_2$ separately.

Case 1: $E_1 < E_2$

$$H(T) = k_1(T)\left\{\underline{\alpha} - \beta\frac{k_2(T)}{k_1(T)}\right\}$$

$$T^2\frac{\partial H}{\partial T} = k_1(T)e_1\left\{\alpha - \beta\frac{k_2(T)e_2}{k_1(T)e_1}\right\}.$$

In this case k_2/k_1 increases with increasing temperature, so if the sign of the Hamiltonian or its derivative changes with increasing temperature it will do so from positive to negative. The frequency factors p_1 and p_2 in the two rate constants may be such that there is no stationary value and the Hamiltonian is monotonic increasing with increasing temperature; if this is not so there is a stationary maximum. These two alternatives are shown in Fig. 1 and it is clear that there is no possibility of a stationary minimum.

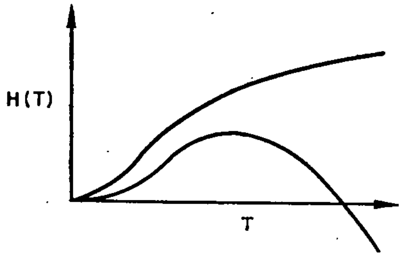


FIG. 1. $E_1 < E_2$. The Hamiltonian as a function of temperature takes one of two possible forms ($\alpha > 0$)

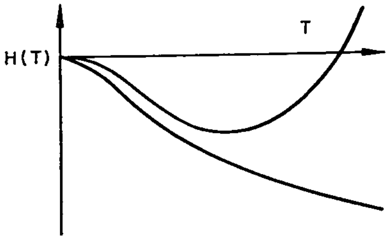


FIG. 2. $E_1 > E_2$. The Hamiltonian as a function of temperature takes one of two possible forms ($\alpha > 0$)

Case 2: $E_1 > E_2$

In a precisely similar way it can be shown that, in this case, there are the alternatives shown in Fig. 2, and it is clear that there is no possibility of a stationary maximum.

LEMMA 5. *When the Hamiltonian as a function of temperature has a stationary value the value of the temperature corresponding to this stationary value decreases with increasing time.*

Proof. In general

$$\left(\frac{\partial H}{\partial T}\right)_{T=T'} = \frac{k_1(T')e_1}{(T')^2}\,\alpha(t) - \frac{k_2(T')e_2}{(T')^2}\,\beta(t)$$

and from equations (14) and (15)

$$\frac{\partial}{\partial t}\left(\frac{\partial H}{\partial T}\right)_{T=T'} = \frac{k_1(T')e_1}{(T')^2}\,ak_2(T_m)\lambda_2 - \frac{k_2(T')e_2}{(T')^2}\,ak_1(T_m)\lambda_2$$

where T_m is the temperature which gives the Hamiltonian at time t a stationary value.

In particular, taking $T' = T_m$ we have

$$T_m^2\frac{\partial}{\partial t}\left(\frac{\partial H}{\partial T}\right)_{T=T_m} = ak_1(T_m)k_2(T_m)\lambda_2(e_1 - e_2) \quad (16)$$

Now if $e_1 > e_2$ we have shown in LEMMA 4 that the Hamiltonian may have a stationary minimum, in which case $(\partial H/\partial T)_{T=T_m} = 0$. Equation (16) then shows that the time rate of change of $\partial H/\partial T$ at a constant temperature equal to the minimising temperature is positive, since a, k_1, k_2, λ_2 are all positive, and hence the minimising value of the temperature must decrease with increasing time. If $e_1 < e_2$, on the other hand, the Hamiltonian may have a stationary maximum and the right hand side of equation (16) is negative. It then follows in the same way that the maximising value of the temperature must decrease with increasing time.

APPLICATION OF PONTRYAGIN'S CONDITION TO THE DETERMINATION OF TEMPERATURE PROFILES

We shall now consider the problem of maximising or minimising the exit concentration of B in a tubular reactor of given length for the kinetic scheme given in the introduction.

Example 1. $E_1 < E_2$. *Minimise the exit concentration of B.*

When $E_1 < E_2$ we have seen in LEMMA 4 that the Hamiltonian cannot have a stationary minimum, so it follows that the Hamiltonian is minimised when the temperature takes one or other of the values bounding the range of interest, which we shall take to be $0 \leqslant T \leqslant T^*$. Consequently any temperature profile satisfying Pontryagin's minimising condition must consist entirely of segments on which the temperature is either zero or T^*, and we can confine our attention to such profiles.

The argument is most easily followed by considering the course of the reaction in an isothermal reactor at T^* and plotting trajectories of a against b with time as a parameter, as shown in Fig. 3. We are interested only in the region $a + b \leqslant 1$ within the triangle OAB and we shall assume further that substance C is not present in the feed at $t = 0$, so that all initial conditions (a_0, b_0) lie on the line AB.

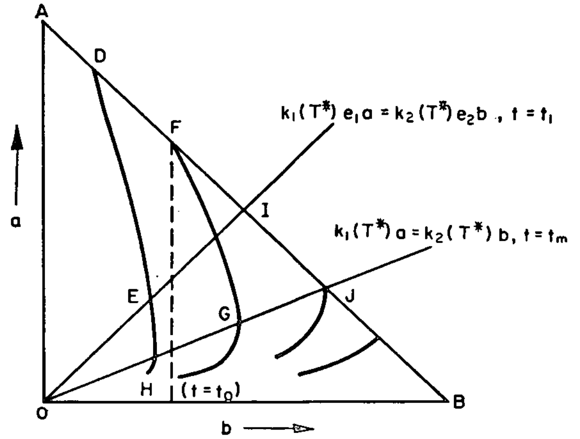


FIG. 3. $E_1 < E_2$. The course of the reaction $A \rightarrow B \rightarrow C$ in an isothermal reactor at T^* for various initial conditions.

The course of the reaction is then represented by a trajectory starting from initial conditions on AB. When $k_1(T^*)a_0 > k_2(T^*)b_0$, corresponding to initial conditions on the line AJ, it is not difficult to show that b passes through a maximum at a certain time $t = t_m$ which depends on the inlet concentrations of A and B, then decreases to zero as the time tends to infinity. When $k_1(T^*)a_0 < k_2(T^*)b_0$ on the other hand, both a and b decrease monotonically with increasing time and approach zero as the time tends to infinity. Hence the form of the trajectories is as indicated in Fig. 3.

In a numerical search for a solution satisfying Pontryagin's minimising condition one could proceed by guessing the values of a and b at the reactor exit $t = \theta$ and integrating equations (7) and (9) backwards along the reactor, choosing the temperature at each point to minimise the Hamiltonian. On reaching the reactor inlet $t = 0$ the values of a and b would then give the inlet conditions corresponding to the solution obtained. These would not, in general, agree with the given inlet boundary conditions (8) and iteration would be necessary to solve this problem. The same procedure will be followed here, but the simplicity of the system is such that the results we require can be obtained with the aid of Fig. 3 without detailed calculation.

First consider terminal conditions $[a(\theta), b(\theta)]$ lying in the triangle OAJ, so that $k_1(T^*)a(\theta) > k_2(T^*)b(\theta)$. Then

$$H(T^*, \theta) = k_1(T^*)a(\theta) - k_2(T^*)b(\theta) > 0$$

and it follows that $H(T, \theta)$ is minimised by taking $T = 0$. If we integrate backwards at $T = 0$, a, b, λ_1 and λ_2 all remain unchanged on moving into $t < \theta$; consequently the Hamiltonian as a function of temperature remains identical with $H(T, \theta)$ and is always minimised by $T = 0$. Thus we obtain the isothermal temperature profile $T = 0$ and the corresponding inlet concentrations $a(0) = a(\theta)$ and $b(0) = b(\theta)$. Since we are interested only in initial conditions on the diagonal AB in Fig. 3 the only relevant terminal conditions in the triangle OAJ lie on the line AJ. Thus, for any initial conditions on the line AJ one possible Pontryagin minimising profile is $T = 0$ everywhere, corresponding to no reaction.

When the terminal conditions lie in the triangle OJB we have $k_1(T^*)a(\theta) < k_2(T^*)b(\theta)$, and correspondingly $H(T^*, \theta) < 0$. Thus $H(T, \theta)$ is minimised by taking $T = T^*$, and if we integrate backwards in time with this value of T the value of $H(T^*, t)$ remains constant and equal to $H(T^*, \theta)$ (LEMMA 3). Thus $H(T, t)$ continues to be minimised by $T = T^*$ for all t and Pontryagin's minimising condition is satisfied by the isothermal temperature profile $T = T^*$ when the terminal conditions lie in OJB. The corresponding reaction trajectories are shown in the diagram and may be traced back until they meet the diagonal AB to give the initial conditions $[a(0), b(0)]$ corresponding to any terminal conditions $[a(\theta), b(\theta)]$.

It is seen that $[a(\theta), b(\theta)]$ may be chosen to correspond to any initial point on the line JB for any reactor length θ, so for such initial conditions the isothermal temperature profile $T = T^*$ always satisfies Pontryagin's minimising condition. A terminal condition $[a(\theta), b(\theta)]$ in OJB can be found to correspond to an initial condition on AJ such as point F only if the reactor is sufficiently long. Clearly θ must be greater than t_m if the isothermal reaction trajectory for $T = T^*$ starting at point F is to terminate within OJB, but provided that this is the case, the profile $T = T^*$ satisfies the Pontryagin minimising condition. Thus, for initial conditions on the line AJ and reactor length $\theta > t_m$, we have shown that both the isothermal profile $T = 0$ and the isothermal profile $T = T^*$ satisfy Pontryagin's minimising condition. In fact both give local minima of $b(\theta)$ in the function space of $T(t)$, and which gives the absolute minimum depends on the length of the reactor. In Fig. 3 it is seen that, on the isothermal trajectory $T = T^*$ through F, b returns to its initial value at time $t = t_0$ corresponding to point H. If $\theta > t_0$, $b(\theta)$ is reduced below b_0 and the isothermal profile $T = T^*$ yields the smallest value of $b(\theta)$. If $\theta < t_0$ on the other hand, the isothermal profile $T = 0$ yields $b(\theta) = b_0$, and this is the smallest obtainable value of $b(\theta)$.

This situation is further complicated if we consider terminal conditions lying on the line $0J$, so that $k_1(T^*)a(\theta) = k_2(T^*)b(\theta)$ and $H(T^*, \theta) = H(0, \theta) = 0$. $H(T, \theta)$ is then minimised by choosing either $T = 0$ or $T = T^*$. With either choice $H(T^*, t)$ retains the value zero on passing back along the reactor from $t = \theta$, so the alternatives $T = 0$ or $T = T^*$ still minimise $H(T, t)$ and the Hamiltonian is minimised at all times by any temperature profile composed entirely of segments at $T = 0$ and at $T = T^*$. In order that the trajectory in the (a, b) plane representing the course of the reaction should terminate at initial conditions on the line AB it is clearly necessary that the total length of the segments on which $T = T^*$ should be t_m, and this is possible only if $\theta \geqslant t_m$. Thus, when $\theta > t_m$ we have a third possibility which also satisfies Pontryagin's minimising conditions for feeds along *AJ*, namely an infinite set of discontinuous profiles consisting of segments at $T = T^*$ and segments at $T = 0$, the segments at T^* having total length t_m. However, the corresponding value of $b(\theta)$ is not even a local minimum in the space of $T(t)$; indeed it is obviously a local maximum for those particular variations consisting of changes in the total length of the segments at T^*.

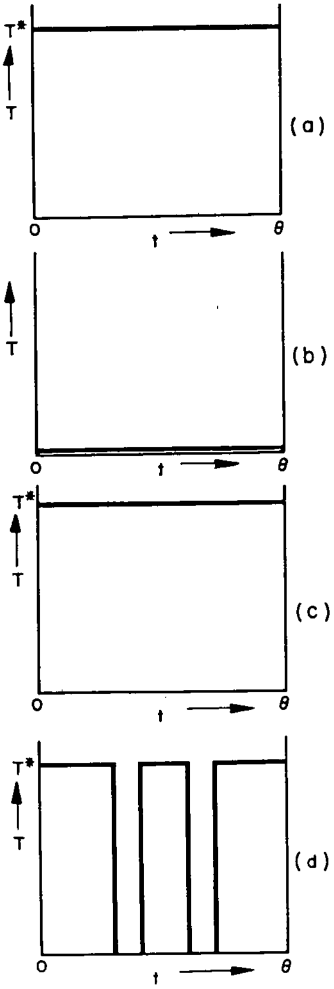


FIG. 4. Temperature profiles satisfying Pontryagin's condition to minimise $b(\theta)$ when $E_1 < E_2$

(a) $k_1(T^*)a_0 < k_2(T^*)b_0$ (all θ) Nonstationary absolute minimum

(b) $k_1(T^*)a_0 > k_2(T^*)b_0$ (all θ) Nonstationary local minimum. Absolute minimum if $\theta < t_0$

(c) $k_1(T^*)a_0 > k_2(T^*)b_0$ ($\theta > t_m$) Nonstationary local minimum. Absolute minimum if $\theta > t_0$

(d) $k_1(T^*)a_0 > k_2(T^*)b_0$ ($\theta \geqslant t_m$) Total length of segments at T^* is t_m. Not even a local minimum.

This situation is summarised in Fig. 4. Of course the cases in which the whole reactor is at the absolute zero are of no practical interest, and it is quite easy to pick out those results which are of physical value. Nevertheless, this example serves to emphasise in a simple way that Pontryagin's condition is only necessary, and not sufficient, so that it is quite possible to find temperature profiles which satisfy the condition but are not the required solution of the problem. Later we shall meet a case where it is less obvious on physical grounds which results should be rejected.

Example 2. $E_1 > E_2$. *Maximise the exit concentration of B.*

When $E_1 > E_2$ we have seen in LEMMA 4 that the Hamiltonian cannot have a stationary maximum, so it follows once again that the Hamiltonian is maximised when the temperature takes one or other of

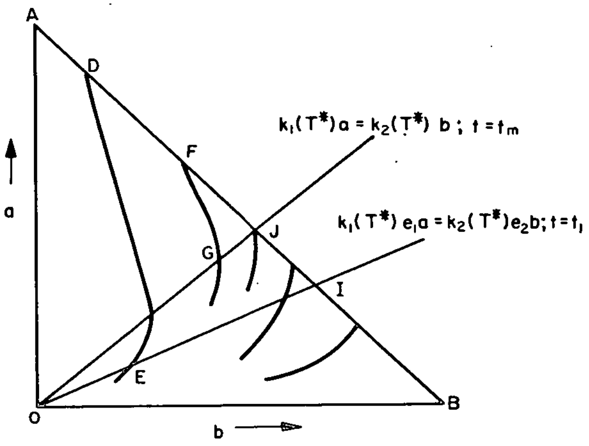

A
D
F
G
J
I
E
O
B
a
b
$k_1(T^*)a = k_2(T^*)\,b\,;\ t = t_m$
$k_1(T^*)\,e_1a = k_2(T^*)e_2b; t = t_1$

FIG. 5. $E_1 > E_2$. The course of the reaction $A \rightarrow B \rightarrow C$ in an isothermal reactor at T^* for various initial conditions.

the boundary values zero or T^*. The argument here follows closely that already developed in Example 1 and will not be given in detail. The conclusions are quoted below and reference should be made to Fig. 5.

For initial conditions on *AJ* (Fig. 5) the only Pontryagin maximising profile when $\theta \leqslant t_m$ is the isothermal profile $T = T^*$, while when $\theta > t_m$ we have an infinite set of discontinuous profiles consisting of segments at $T = T^*$ and segments at $T = 0$, the segments at T^* having total length t_m. For initial conditions on *JB* the Pontryagin maximising condition is satisfied only by the isothermal profile $T = 0$, corresponding to no reaction. The possibilities are summarised in Fig. 6.

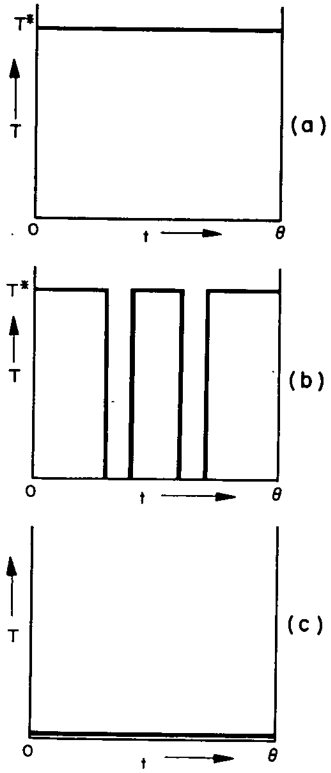


FIG 6. Temperature profiles satisfying Pontryagin's condition to maximise $b(\theta)$ when $E_1 > E_2$.
(a) $k_1(T^*)\, a_0 > k_2(T^*)b_0$ $(\theta \leqslant t_m)$ Nonstationary absolute maximum
(b) $k_1(T^*)a_0 > k_2(T^*)b_0$ $(\theta > t_m)$ Total length of segments at T^* is t_m. Nonstationary absolute maximum.
(c) $k_1(T^*)a_0 < k_2(T^*)b_0$ (all θ) Nonstationary absolute maximum

Example 3. $E_1 > E_2$. *Minimise the exit concentration of B.*

When $E_1 > E_2$ we have seen in LEMMA 4 that the Hamiltonian may have a stationary minimum value in the temperature interval of interest and accordingly there may be Pontryagin minimising temperature profiles for which the Hamiltonian takes a stationary minimum value at some, or all, times.

A consideration of the Hamiltonian at $t = \theta$ shows that it is monotonic decreasing with increasing temperature in the range $0 \leqslant T \leqslant T^*$ when the exit concentrations $[a(\theta), b(\theta)]$ are such that

$$k_1(T^*)e_1a(\theta) < k_2(T^*)e_2b(\theta).$$

Following the argument given in LEMMA 5 we can then say that the Hamiltonian at any previous time is also monotone decreasing with increasing temperature in the range of interest when the exit concentrations satisfy this inequality. Thus, for terminal conditions in the triangle *OIB* of Fig. 5 the Pontryagin minimising condition is satisfied by the isothermal profile $T = T^*$ and the course of the reaction is then given by the trajectories on the diagram. For feeds represented by points along the line *AI* it is seen from Fig. 5 that isothermal reactions at T^* will give exit concentrations within the triangle *OIB* only if $\theta > t_1$. Thus, for feeds along *AI* one Pontryagin minimising possibility is the isothermal profile $T = T^*$ when $\theta > t_1$: this possibility does not exist when $\theta < t_1$. For feeds along *IB* isothermal reaction at T^* will always give terminal conditions in this area no matter what the value of θ.

We now turn to the question of the existence of temperature profiles for which the Hamiltonian has a stationary minimum at some, or all, times. It is apparent that the Hamiltonian at $t = \theta$ can have a stationary minimum in the range $0 \leqslant T \leqslant T^*$ only if

$$k_1(T^*)e_1a(\theta) > k_2(T^*)e_2b(\theta) \qquad (17)$$

in other words, only if the terminal conditions lie in the triangle *OAI*. Thus for feeds along *AI* we now have the possibility of a Pontryagin minimising temperature profile which gives the Hamiltonian a stationary minimum value at some, or all, times, whatever the value of θ, so long as the exit concentrations lie within the triangle *OAI*. This possibility

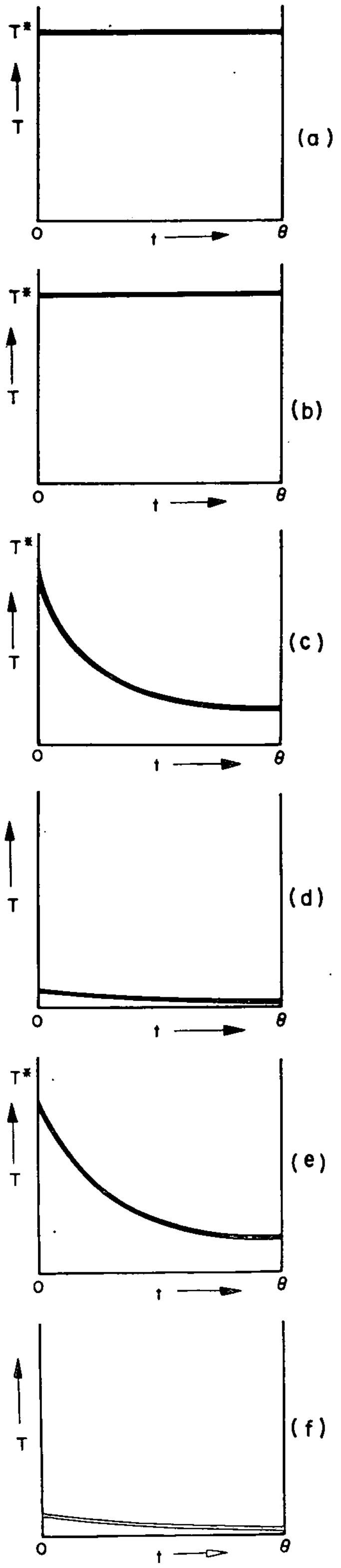


FIG. 7. Temperature profiles satisfying Pontryagin's condition to minimise $b(\theta)$ when $E_1 > E_2$.

(a) $k_1(T^*)e_1a_0 < k_2(T^*)e_2b_0$ (all θ) Nonstationary absolute minimum

(b) $k_1(T^*)e_1a_0 > k_2(T^*)e_2b_0$ ($\theta > t_1$) Nonstationary local minimum

(c) $k_1(T^*)e_1a_0 > k_2(T^*)e_2b_0$ ($\theta > t_1$) Saddle point

(d) $k_1(T^*)e_1a_0 > k_2(T^*)e_2b_0$ ($\theta > t_1$) Stationary local minimum

(e) $k_1(T^*)e_1a_0 > k_2(T^*)e_2b_0$ ($\theta < t_1$) Saddle point

(f) $k_1(T^*)e_1a_0 > k_2(T^*)e_2b_0$ ($\theta < t_1$) Stationary absolute minimum

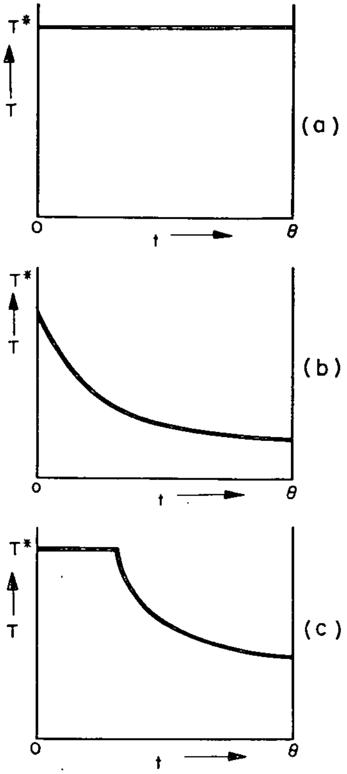


FIG. 8. Temperature profiles satisfying Pontryagin's condition to maximise $b(\theta)$ when $E_1 < E_2$.

(a) $k_1(T^*)e_1a_0 > k_2(T^*)e_2b_0$ ($\theta < t_1$) Nonstationary maximum

(b) $k_1(T^*)e_1a_0 < k_2(T^*)e_2b_0$ (all θ) or $k_1(T^*)e_1a_0 > k_2(T^*)e_2b_0$ (θ sufficiently large) Stationary maximum

(c) $k_1(T^*)e_1a_0 < k_2(T^*)e_2b_0$ (all θ) or $k_1(T^*)e_1a_0 > k_2(T^*)e_2b_0$ (θ sufficiently large) Nonstationary maximum.

for feeds along AI represents a second minimising alternative when $\theta > t_1$ and the only possibility when $\theta < t_1$.

The most convenient method of finding such solutions involves the use of equation (13). The procedure is as follows:

(i) The inlet temperature $T(0)$ is guessed and equations (7) and (13) are integrated forwards from the initial conditions a_0, b_0, $T(0)$.

(ii) The final values $a(\theta)$, $b(\theta)$, $T(\theta)$ obtained in this way must correspond to a stationary value of $H(T, \theta)$, so we must have

$$k_1[T(\theta)]e_1 a(\theta) = k_2[T(\theta)]e_2 b(\theta). \qquad (18)$$

In general this relation will not be satisfied and we must adjust the initial guess $T(0)$ until it is. Note that it will not be possible to find an admissible temperature satisfying equation (18) unless the inequality (17) is satisfied.

To clarify the situation we shall quote the results of some numerical calculations on a particular example. If we assume the existence of a temperature profile which gives a stationary minimum of the Hamiltonian throughout the reactor it can be calculated numerically by the procedure outlined above. This was done for the example given by BILOUS and AMUNDSON [2] for the case $E_1 > E_2$. In this case $\theta > t_1$ (with $T^* = 400$°K) and *two* falling temperature profiles were found which satisfied the condition for a stationary minimum everywhere. The first of these was indentical with that given by BILOUS and AMUNDSON for maximising the exit concentration of B, but the results obtained on perturbing this profile indicate that it is actually a saddle-point. The second, which corresponded to a low value of the inlet temperature $T(0)$ and fell so slowly as to be virtually isothermal, was found to give a local minimum. Thus, for feeds along AI it would appear from this example that when $\theta > t_1$ we may have no less than three possible Pontryagin minimising profiles, namely the isothermal profile $T = T^*$ and two falling profiles, both of which give a stationary minimum value of the Hamiltonian throughout. Of these three possibilities, the intermediate temperature falling profile identical with BILOUS and AMUNDSON's result is a saddle point and therefore cannot give an absolute minimum, while the other two are both local minima and either one may give the absolute minimum. In this particular example the absolute minimum is given by the isothermal profile $T = T^*$.

For feeds along AI when $\theta < t_1$ the isothermal profile $T = T^*$ has already been shown to be inadmissible and we may tentatively conclude that there will again be two falling temperature profiles of which the lower gives the absolute minimum. This situation is summarised in Fig. 7.

Example 4. $E_1 < E_2$. Maximise the exit concentration of B.

When $E_1 < E_2$ we have seen in LEMMA 4 that the Hamiltonian may have a stationary maximum value in the temperature interval of interest and accordingly there may be Pontryagin maximising temperature profiles for which the Hamiltonian takes a stationary maximum value at some, or all, times. The argument here follows closely that already developed in considering Example 3 and will not be given in detail. The conclusions are quoted below and reference should be made to Fig. 3.

It is easily demonstrated that the Hamiltonian is monotone increasing with temperature within the interval $0 \leqslant T \leqslant T^*$ throughout the reactor for any reaction trajectory which terminates within the triangle OAI. For initial conditions on AI, therefore, the Pontryagin maximising profile when $\theta < t_1$ is the isothermal profile $T = T^*$. For reaction trajectories terminating within the triangle OBI, corresponding to initial conditions on IB or initial conditions on AI when θ is sufficiently large, it is possible to find an admissible exit temperature satisfying equation (18) and we must seek solutions for which the Hamiltonian takes a stationary maximum at some, or all, times. The procedure outlined in Example 3 was applied to the example given by BILOUS and AMUNDSON with $E_1 < E_2$ and a result identical with theirs was obtained.

The situation is summarised in Fig. 8. Which of the cases 8(*b*) and 8(c) is found in any given example depends on the value specified for T^*.

CONCLUSIONS

It is felt that the examples considered here are sufficient to dispel any illusion that the search for optimum conditions using Pontryagin's Maximum Principle is simply a matter of following through a prescribed procedure, with the possible impediment of computational difficulties. Although the problem considered is a very simple one, of its type, we have found cases in which there are several profiles, all satisfying Pontryagin's conditions, only one of which is the required solution (in particular Example 3), cases in which there are an infinite number of profiles, none of which is the solution, in addition to the one which is (Example 1), and cases in which there are an infinite number of profiles, all satisfying Pontryagin's condition and all providing valid solutions of the problem (Example 2). Some of these cases arise from the constraints imposed on the temperature profile, but in Example 3 we met a case in which there were two profiles, each satisfying Pontryagin's condition and each lying entirely between the permitted bounds of variation of T.

Of course the physical meaning of some of the extraneous solutions is quite easy to see in the present case; for example the infinite set of discontinuous temperature profiles found in Example 2 are simply a result of making the reactor too long. However, examples of similar situations in a rather less elementary problem can also be found in the recent work of SIEBENTHAL and ARIS [7].

NOTATION

a, b	Concentrations of A and B
a_0, b_0	The given initial concentrations of A and B
$a(0), b(0)$	Initial concentrations of A and B corresponding to guessed values of the exit concentrations of A and B; not necessarily equal to a_0, b_0
$a(\theta), b(\theta)$	Exit concentrations of A and B
E_1, E_2	Activation energies for reactions 1 and 2
e_1, e_2	E_1/R and E_2/R respectively
H	The Hamiltonian, defined by equation (11)
$H(T)$	The Hamiltonian at any given point in the reactor regarded as a function of temperature only
$H(T, t)$	The Hamiltonian regarded as a function of both temperature and time along the reactor
k_1, k_2	Velocity constants for reactions 1 and 2
$k_1(T), k_2(T)$	Velocity constants for reactions 1 and 2 as functions of temperature
t	Time along the reactor measured from $t = 0$ at inlet
t_m	Time at which the concentration of B is a maximum in an isothermal reactor at T^* with feed concentrations a_0, b_0
t_1	Time at which concentrations of A and B are such that $k_1 e_1 a = k_2 e_2 b$ in an isothermal reactor at T^* with feed concentrations a_0, b_0
t_0	Time at which the concentration of B is again equal to b_0 in an isothermal reactor at T^*
T	Temperature
T_m	Temperature which maximises (minimises) the Hamiltonian at time t
T^*	Upper temperature limit
α, β	Numerical quantities in the expression for $H(T)$
$\alpha(t), \beta(t)$	The same quantities regarded as functions of time along the reactor in expressions for $H(T, t)$; defined in LEMMA 3
λ_1, λ_2	Variables adjoint to a, b; defined by equations (9) and (10)
θ	The given reaction time

REFERENCES

[1] ROSENBROCK H. H. and STOREY C., *Computational Techniques for Chemical Engineers*. In press (Pergamon Press Ltd.).

[2] BILOUS O. and AMUNDSON N. R., *Chem. Engng. Sci.* 1956 **5** 81, 115.

[3] ARIS, R., *Chem. Engng Sci.* 1960 **13** 18.

[4] ROZONOER L. I., *Automation and Remote Control* 1959 **20** 1288, 1405, 1517.

[5] HORN F., *Chem. Engng Sci.* 1961 **15** 176.

[6] HORN F., *Chem. Engng Sci.* 1961 **14** 77.

[7] SIEBENTHAL C. D. and ARIS R., *Chem. Engng Sci.* 1964 **19** 729.

Chemical Engineering Science, 1966, Vol. 21, pp. 241–260. Pergamon Press Ltd., Oxford. Printed in Great Britain.

Optimum startup procedures for an autothermic reaction system

R. JACKSON

University of Edinburgh and Heriot-Watt College

(*Received* 14 *June* 1965; *in revised form* 13 *August* 1965)

Abstract—Certain exothermic reactions of considerable commercial importance make use of regenerative heat exchange between reactant and product streams in such a way that they are thermally self-sustaining, or autothermic, when operating steadily. Nevertheless they cannot be started up without supplying heat to the reactants from an external source, though the external heating may be withdrawn once "ignition" has been achieved. This feature makes the determination of the correct startup procedure a problem of some interest.

In this paper it is shown how the idea of optimum startup can be given a precise quantitative formulation, and Pontryagin's maximum principle is used to determine the optimum startup procedure.

INTRODUCTION

THE AMMONIA synthesis reaction

$$N_2 + 3H_2 = 2NH_3$$

provides an example of a commercially important reaction which is carried out in such a way that it is thermally self-sustaining, or autothermic. At atmospheric temperature the rate of reaction is quite negligible and a temperature of several hundred degrees centigrade is necessary if it is to proceed at a useful speed. However, since it is strongly exothermic, the heat of reaction may be used to preheat the reactant mixture to a temperature sufficiently high to maintain the required reaction rate. This is accomplished by means of a heat exchanger in which the incoming reactants are contacted with the hot gases leaving the reactor.

Systems of this type are well known to exhibit an ignition phenomenon, rather like flames. If cold reactants are fed to the reactor there is a neglibible amount of reaction and insufficient heat is generated to raise the temperature of the reactants significantly. However, if the reactants are sufficiently preheated the reaction is much more vigorous and the heat generated suffices to maintain the necessary temperature at inlet without any further need for external heating. Thus, although the heater plays no part in the steady operation of the system, its presence is essential for startup. A good account of ignition phenomena of the type just described has been given by VAN HEERDEN [1].

In the simplest type of autothermic system the reactor is adiabatic and its operation and startup are controlled by regulating the external heat supply to the reactant preheater and the fraction of the hot gas leaving the reactor which passes through the regenerative heat exchanger. The startup procedure adopted is often influenced by special features of the particular reaction considered; for example, in the case of ammonia synthesis, startup with a new charge of catalyst must commence with a period of catalyst reduction. Nevertheless the mathematical techniques now available to handle problems of this sort can be illustrated by considering a system where optimum startup is a simple compromise between the desirability of reaching steady operating conditions quickly and the cost of the external heating required to do this. This includes, of course, as a limiting case, the very common situation in which heating costs are of little account and it is vital to bring the system to its steady operating conditions as quickly as possible.

Startup problems in chemical engineering closely resemble the problem of guiding a missile to a specified target, and perhaps the most successful technique developed to deal with this type of problem is the maximum principle of Pontryagin, of which a good elementary account can be found in the papers of ROZONOER [2]. In the present paper it will be shown that the maximum principle leads to a complete solution of the optimal startup problem if one assumes that conditions in the reactor approximate to perfect mixing. At the other extreme, the approximation of no axial mixing raises some interesting mathematical problems which will be discussed elsewhere.

Recently SIEBENTHAL and ARIS [3] have used the maximum principle to discuss the optimal control of some simple reaction systems, and the reader might find it useful as a preliminary to read the first of their papers in order to familiarize himself with the method, as illustrated by examples rather simpler than the one considered here.

THE MATHEMATICAL MODEL

We shall consider the system shown schematically in Fig. 1. For simplicity the first order exothermic reversible reaction $A \rightleftharpoons B$ will be treated, though extension to more complicated cases presents no difficulty. The reaction is carried out in a reactor which is assumed to approximate to an adiabatically isolated, perfectly mixed vessel which can hold V moles of the reaction mixture. The incoming reactants may be heated both by heat exchange with the product stream and by heat supplied externally to a heater. The molar flow rate of the reaction mixture is F and it enters the cold side of the heat exchanger at a temperature T_0. The temperature is raised to T_1 in the exchanger, then further to T_1' in the heater, and finally to a value T at the reactor exit. The stream leaving the reactor, containing a mole fraction y of the product, is then split so that part of it, represented by the flow f, passes through the hot side of the exchanger to preheat the incoming reactants.

In considering the startup procedure we shall be interested in time-varying conditions, so we must write down the differential equations representing unsteady state mass and heat balances for the system, namely

$$\frac{dy}{dt'} = r(y, T) - y/\tau \tag{1}$$

and

$$\gamma\tau\frac{dT}{dt'} + T = T_1' + \Delta T_{\text{ad}}\tau r(y, T) \tag{2}$$

Here $r(y, T)$ denotes the reaction rate, which depends on the composition and temperature of the reaction mixture, as indicated, and τ denotes the ratio $V{:}F$, the mean residence time in the reactor. If ΔH is the heat of reaction (negative for an exothermic reaction) and C the molar specific heat of the reaction mixture, the symbol ΔT_{ad} is introduced to represent the ratio $(-\Delta H/C_g)$: physically ΔT_{ad} is the temperature rise accompanying complete reaction under adiabatic conditions. The thermal capacity of the reactor contents would be VC_g if they consisted solely of the reaction mixture, but very often the reactor is packed with catalyst so that its thermal capacity takes a different value VC. The ratio $C{:}C_g$ is then denoted by γ. The dependence of ΔH and C_g on the temperature and composition of the mixture is neglected. Finally, note that t' is used to indicate time, the symbol t being reserved for a dimensionless measure of time introduced later.

Equations (1) and (2) do not provide a complete description of the system since they contain the temperature T_1', which is determined by conditions in the heater and exchanger. Strictly speaking,

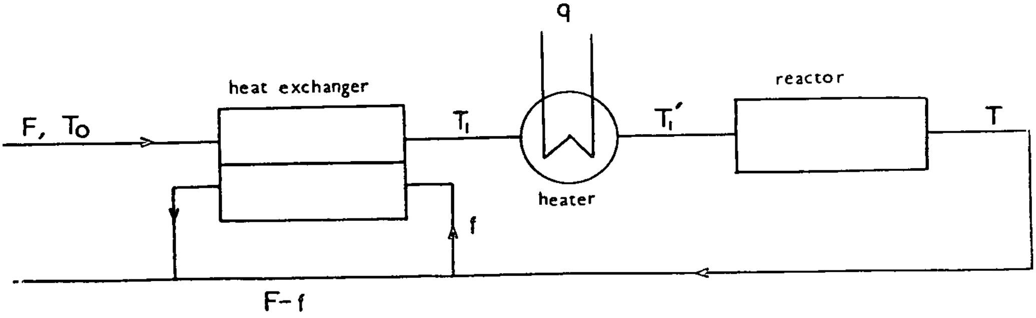


FIG. 1. Thermally regenerative reaction system.

dynamic equations analogous to (2) should be written for the exchanger and heater, but very often the reactor has a much larger thermal capacity than these units, and consequently its speed of response to changing conditions is much slower. The states of the heater and exchanger then approximate closely to instantaneous steady states at all times, and if the exchanger is assumed to be purely countercurrent in operation the temperatures T_0, T_1 and T are related by a well known equation, which may be written in the form

$$T_1 = uT_0 + (1-u)T \tag{3}$$

where

$$u = \frac{1 - f/F}{1 - f/F \exp[-\alpha(F/f - 1)]} \tag{4}$$

The constant α is characteristic of the exchanger and the heated gas stream and is given by $\alpha = hA/C_gF$, where h is the mean overall heat transfer coefficient and A the area of heat transfer surface available. The reactor inlet temperature T_1' is then related to T_1 by

$$T_1' = T_1 + q \tag{5}$$

where q has the dimensions of temperature and is proportional to the heat supplied to the reaction mixture in the heater.

Using Eqs. (3) and (5), T_1' can now be eliminated from equation (2). At the same time it is convenient to introduce a dimensionless measure of time $t = t'/\tau$, and a dimensionless reaction rate $R = \tau r$. Equations (1) and (2), then reduce to

$$\frac{dy}{dt} = R(y, T) - y \tag{6}$$

and

$$\gamma\frac{dT}{dt} = q - u(T - T_0) + \Delta T_{ad}R(y, T) \tag{7}$$

which provide our mathematical model for the dynamical behaviour of the system.

The reaction is controlled by varying the heat supply q and the flow f through the hot side of the exchanger. However, it can be seen from Eq. (4) that u is a monotone decreasing function of f/F, decreasing from $u = 1$ when $f/F = 0$ to $u = 1/(1+\alpha)$ when $f/F = 1$. Thus there is a one-to-one correspondence between values of u and f/F and it is convenient to regard u as the control variable rather than f/F, since it enters very simply into Eq. (7). Variations of u and q are bounded both above and below. We have already seen that

$$u_{\min} \leqslant u \leqslant 1 \quad \text{with} \quad u_{\min} = 1/(1+\alpha) \tag{8}$$

and, for a given design of heater, q will be bounded above by some value $q_{\max}$, so that

$$0 \leqslant q \leqslant q_{\max} \tag{9}$$

Given any initial state $(y(0), T(0))$, together with a specification of the two control variables q and u as functions of time, Eqs. (6) and (7) can be integrated to give the behaviour of the system in $t > 0$. We shall be concerned, in particular, with the behaviour in response to manipulations of q and u subject to the constraints (8) and (9).

Steady State Operation

Necessary conditions for steady operation are obtained by equating $\frac{dy}{dT}$ and $\frac{dT}{dt}$ to zero in Eqs. (6) and (7), giving the following two equations

$$R(y, T) - y = 0 \tag{10}$$

and

$$q - u(T - T_0) + \Delta T_{ad}R(y, T) = 0 \tag{11}$$

which may be solved for the two unknowns y and T once the form of the function R is known. For the first order reversible reaction $A \rightleftharpoons B$, we have

$$R(y, T) = K_1(T)(1-y) - K_2(T)y \tag{12}$$

where

$$\left.\begin{aligned} K_1(T) &= \tau k_{01}\, e^{-e_1/T} \\ K_2(T) &= \tau k_{02}\, e^{-e_2/T} \end{aligned}\right\} \tag{13}$$

assuming the usual Arrhenius form of temperature dependence for the velocity constants. Since the reaction is exothermic, $e_2 > e_1$. With R given by Eq. (12), Eq. (10) may be solved for y, giving

$$y = \frac{K_1}{1 + K_1 + K_2} \tag{14}$$

and using this to eliminate y from Eq. (11)

$$u(T - T_0) - q = \Delta T_{ad}\left(\frac{K_1}{1 + K_1 + K_2}\right) \tag{15}$$

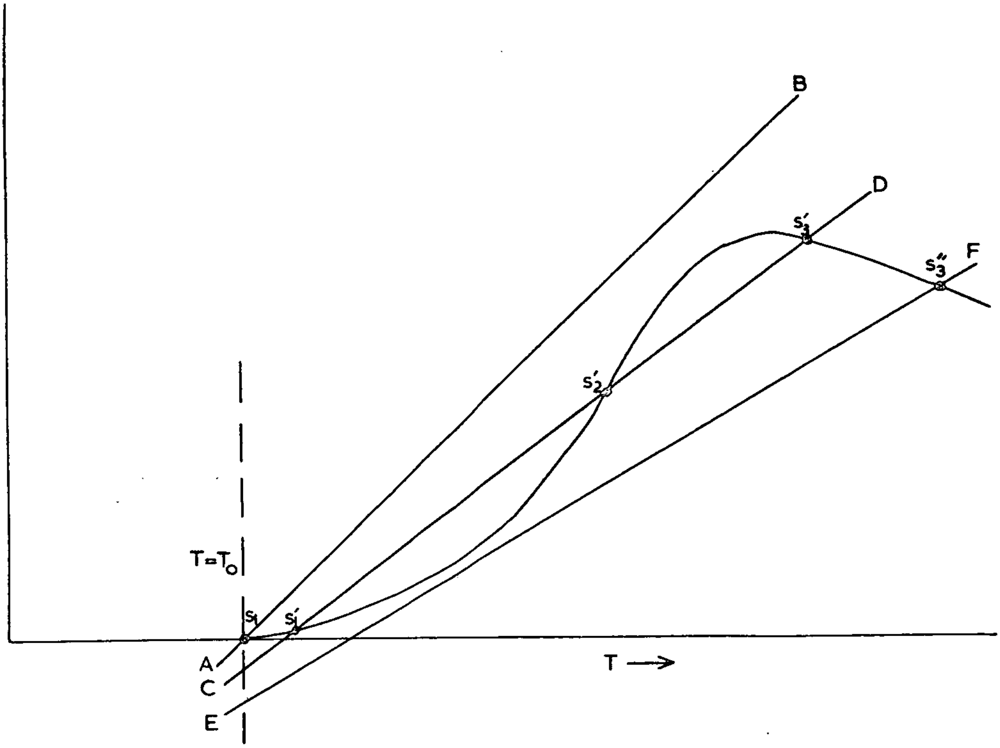


FIG. 2. Possible steady states.

It is not difficult to sketch the right hand side of Eq. (15) as a function of T making use of Eqs. (13), and the result is shown as the curve in Fig. 2. The left-hand side of Eq. (15) is a straight line of slope u which intersects the vertical line $T = T_0$ a distance q below the axis $T = 0$. An intersection of this straight line and the curve gives a value of T corresponding to a possible steady state and it is seen from Fig. 2 that there may be one or three such points, depending on the values of u and q. When T_0 and q are small and u is large, corresponding to the line AB, there is a single steady state s_1 at a low temperature very near to T_0. Correspondingly, the extent of reaction is very small. If q is increased and u decreased a line such as CD is obtained, intersecting the curve at the three points s_1', s_2' and s_3'. s_1' once again corresponds to a low temperature and very little reaction, but at s_3' the temperature is high and the extent of reaction large. We shall see later that s_2' represents an unstable condition which can only formally be regarded as a steady state at all. When q is increased and u decreased still further, a line such as EF is eventually obtained, intersecting the curve at a single point s_3'' corresponding to a high temperature and large extent of reaction.

This behaviour corresponds to the well known "ignition" phenomenon in reactors of this type which was discussed by VAN HEERDEN [1]. A rigorous account of this must be deferred until we consider the dynamical equations in more detail, but an intelligent guess at what happens can be based on the steady state equations, as illustrated by Fig. 3. Suppose the system starts in the low temperature steady state s_1 with $q = 0$ and $u = u_1$. If u is decreased, corresponding to an increase in the flow of hot gas through the exchanger, s_1 moves as indicated by the arrow to higher values of T as the straight line pivots on the point. P. When u reaches the value u_2, the steady state has progressed to s_2, and with any further decrease in u two of the steady states are lost, leaving only the high temperature steady state s_3. We may therefore surmise that, at this point, the system suddenly "ignites" and settles in state s_3.

If the flow of hot gas to the exchanger is then decreased, causing u to increase once more, s_3 moves continuously to lower values of T, as indicated by the arrow, until it reaches the position s_4, corresponding to $u = u_1$. With any further increase in u two of the steady states are lost, leaving only the low temperature state s_1, so we may surmise that the system returns to this state and the reaction is effectively extinguished. At no stage in the cycle has the system settled at the central intersection of the line and the curve, and this is a consequence of the instability of the corresponding state noted above.

The ease with which autothermic operation can be attained depends on the kinetics and thermodynamics of the reaction, the size of the heat exchanger and reactor, and the inlet temperature T_0. If T_0 is high, clearly a high conversion can be obtained with very little heating or heat exchange, and when T_0 is sufficiently high the straight line and the curve in Figs. 2 and 3 never intersect more than once. If T_0 is low, on the other hand, and the heat exchanger is small, the slope u_2 in Fig. 3, which is necessary to cause ignition, may be smaller than the smallest value attainable, $u = 1/(1 + \alpha)$, even with all the hot gas passing through the exchanger. It is then necessary that q should have a finite value to achieve autothermic operation. If ΔT_{ad} and τ are larger the ordinates of the curve in Fig. 3 are large, and it is even possible for the line and curve to intersect three times when $u = 1$ and $q = 0$, corresponding to no heating, either externally or by exchange. Thus a well mixed adiabatic reactor may operate steadily in an autothermic state without any heat exchange if the heat of reaction and capacity of the reactor are sufficiently large.

In practical systems of the present type such as ammonia synthesis converters, autothermic operation without heat exchange is not usually possible, but the size of the heat exchanger is such that $u = u_{\min} = 1/(1 + \alpha)$ corresponds to a line like CD in Fig. 2, intersecting the curve in three points. Steady autothermic operation in a state such as s_3' is then possible, but heat must be supplied externally to initiate the ignition which leads to the autothermic state. This is precisely analogous to the situation encountered in the combustion of fuels, where the fuel must first be heated to initiate the reaction, which is subsequently maintained by its own heat output.

It is clear that the state corresponding to the highest point of the curve in Figs. 2 and 3 gives maximum conversion of reactants to products and,

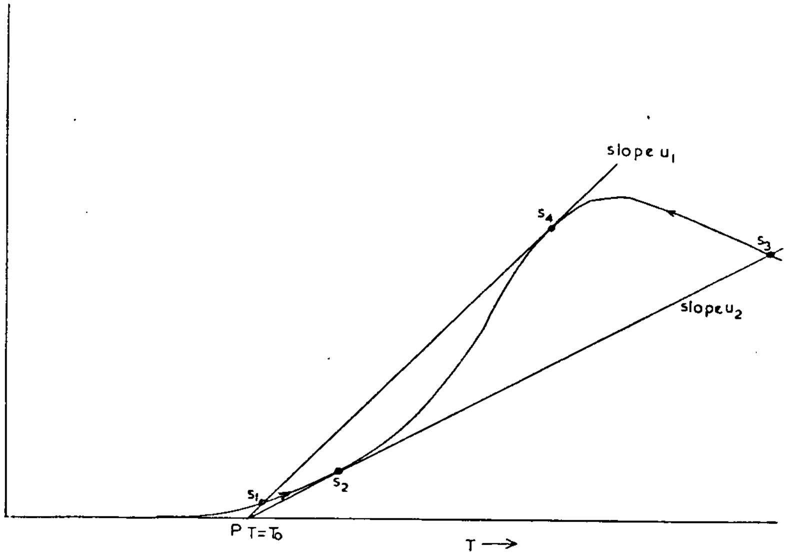


FIG. 3. Illustration of the ignition phenomenon.

provided this is a possible autothermic state, it will represent the optimum operating conditions.

THE OPTIMUM STARTUP PROBLEM AND THE MAXIMUM PRINCIPLE

If reactant gas is circulated through the system without any external heat supply or heat exchange, so that $u = 1$, $q = 0$, the steady state achieved is represented by s_1 in Fig. 2, and the conversion achieved is negligible. However, it is required to operate the reactor in an autothermic state such as s'_3, with y and T taking given steady values y_s and T_s respectively. The problem, then, is to manipulate the variables u and q in such a way that the system is transferred from its initial state (y_0, T_0) to the final autothermic state $(y_s\, T_s)$ in the best possible way. To formulate this problem precisely we must define what is meant by the "best possible way" and express this mathematically in terms of an objective function which it is desired to maximise or minimise.

Let us suppose that the startup procedure commences at $t = 0$ and that the reactor attains its final steady state at time $t = \theta$. Let u_s and q_s be the values of u and q for steady operation in the final state (q_s will normally be zero). Then if p_1 is the selling price per mole of product and p_2 the cost per unit of external heating, the net profit during the startup period is

$$P_a = p_1 F \int_0^\theta y\, dt - p_2 \int_0^\theta q\, dt$$

If the system had operated in its final steady state throughout this period, the profit would have been

$$P_b = p_1 F y_s \theta - p_2 q_s \theta$$

so the loss of profit due to startup is

$$P' = P_b - P_a = p_1 F \int_0^\theta (y_s - y)\, dt - p_2 \int_0^\theta (q_s - q)\, dt$$

and the plant is started up as economically as possible if this quantity is minimised. Since p_1, p_2 and F are constants the algebra can be simplified slightly by seeking to minimise the alternative objective function

$$P = \int_0^\theta [(y_s - y) - c(q_s - q)]\, dt \qquad (16)$$

with $c = p_2/p_1 F$. This differs from P' only by a constant factor. Mathematically, then, our object is to choose $u(t)$ and $q(t)$ in the time interval $0 \leqslant t \leqslant \theta$, with θ unspecified, subject to the constraints $u_{\min} \leqslant u \leqslant 1$, $0 \leqslant q \leqslant q_{\max}$, in such a way as to transfer the system from the state (y_0, T_0) to $(y_s\, T_s)$ with a minimum value for P. The variation of y and T is constrained by the boundary conditions and the differential Eqs. (6) and (7) which they satisfy.

The maximum principle of Pontryagin provides a technique which is well adapted to the solution of problems of this type, and before proceeding any further we will briefly summarise this result. A fuller account may be found in the papers of ROZONOER [2].

Consider a set of simultaneous differential equations

$$\frac{dx_i}{dt} = f_i(x_j, w_p) \qquad (i = 1, 2, \dots n) \qquad (17)$$

with the following boundary conditions

$$\text{and} \qquad \left.\begin{aligned} x_i(0) &= x_{oi} \qquad (i = 1, 2, \dots n) \\ x_i(\theta) &= x_{fi} \qquad (i \epsilon I) \end{aligned}\right\} \qquad (18)$$

where I is a subset of the numbers 1, 2, $n \dots$ It is required to choose the functions $w_p(t)$ in $0 \leqslant t \leqslant \theta$, possibly subject to constraints of the form $a_p \leqslant w_p \leqslant A_p$, in such a way as to minimise a specified linear combination

$$P = \sum_{i \notin I} \alpha_i x_i(\theta) \qquad (19)$$

of those variables $x_i(\theta)$ whose values are not fixed by the boundary conditions (18). To solve this problem one introduces a set of variables λ_i, adjoint to the x_i and defined by the differential equations they satisfy, namely

$$\frac{d\lambda_i}{dt} = -\sum_{j=1}^{n} \lambda_j \frac{\partial f_j}{\partial x_i} \qquad (20)$$

together with the boundary conditions

$$\lambda_i(\theta) = -\alpha_i \qquad (i \epsilon I) \qquad (21)$$

Using these variables we may define a Hamiltonian function

$$H = \sum_{i=1}^{n} \lambda_i f_i(x_j, w_p) \tag{22}$$

in terms of which it is possible to give necessary conditions for the minimisation of P. The form of these conditions depends on whether θ is specified or is itself available to be varied in minimising P. If θ is specified, the variables w_p must be chosen so that H is maximised at each t, the values of the λ_i and x_j being regarded as constant in this maximisation. This condition is also necessary if θ is not specified, but then it is additionally necessary that the maximum value of H should vanish at all times.

$$\max_{w_p} H(\lambda_i, x_j, w_p) = 0 \qquad (\text{all } t) \tag{23}$$

At first sight our problem does not take the form used in stating the maximum principle, since our objective function (16) is an integral over the time interval of interest, while the objective function (19) involves only the values of variables at the terminal time. However, it is easily reduced to the desired form by introducing a new variable z defined by the differential equation

$$\frac{dz}{dt} = y_s - y - c(q_s - q) \tag{24}$$

together with the boundary condition

$$z(0) = 0 \tag{25}$$

Our objective function (16) is then simply $P = z(\theta)$, which is of the desired form (19). Equation (24), together with the mass and heat balance equations

$$\frac{dy}{dt} = R(y, T) - y \tag{6}$$

and

$$\gamma\frac{dT}{dt} = q - u(T - T_0) + \Delta T_{ad}R(y, T) \tag{7}$$

then corresponds to the differential Eqs. (17) introduced in stating the maximum principle. The boundary conditions corresponding to (18) are the specified values of (y_0, T_0) and $(y_s\ T_s)$, together with condition (25) on z. Introducing variables λ_1, λ_2 and λ_3 adjoint to z, y and T respectively, adjoint equations analogous to (20) can be written down, namely,

$$\frac{d\lambda_1}{dt} = 0 \tag{26}$$

$$\frac{d\lambda_2}{dt} = \lambda_1 - \lambda_2(R_1 - 1) - \lambda_3\Delta T_{ad}R_1/\gamma \tag{27}$$

and

$$\frac{d\lambda_3}{dt} = -\lambda_2R_2 - \lambda_3(\Delta T_{ad}R_2 - u)/\gamma \tag{28}$$

where $R_1 = \partial R/\partial y$ and $R_2 = \partial R/\partial T$
These are subject to the single boundary condition

$$\lambda_1(\theta) = -1 \tag{29}$$

corresponding to (21) in the general case. Finally, the Hamiltonian is given by

$$H = \lambda_1[y_s - y - c(q_s - q)] + \lambda_2[R(y, T) - y] + \lambda_3[q - u(T - T_0) + \Delta T_{ad}R(y, T)]/\gamma \tag{30}$$

Thus the problem is reduced completely to a form in which the maximum principle may be applied.

Some General Features of the Solution

A good deal can be learned about the optimum startup procedure simply by inspection of the Hamiltonian (30). From Eqs. (26) and (29) it is clear that $\lambda_1 = -1$ for all t, so the Hamiltonian may be written

$$H = (\lambda_3/\gamma - c)q - \lambda_3(T - T_0)u/\gamma + (y - y_s + cq_s) + \lambda_2[R(y, T) - y] + \lambda_3\Delta T_{ad}R(y, T)/\gamma \tag{31}$$

showing that it varies monotonically with each of the control variables u and q. Since these variables are to be chosen to maximise H their values will depend on the signs of the factors by which they are multiplied, so in the case of q we must take $q = q_{\max}$ if $\lambda_3 > \gamma c$ and $q = 0$ if $\lambda_3 < \gamma c$. Since we are interested principally in temperatures higher than T_0, the value of u depends on the sign of λ_3, with $u = 1$ if $\lambda_3 < 0$ and $u = u_{\min}$ if $\lambda_3 > 0$. These results can be summarised as follows:

$$\left.\begin{array}{l} \text{When } \lambda_3 > \gamma c \text{ then } q = q_{\max} \text{ and } u = u_{\min} \\ \text{When } \gamma c > \lambda_3 > 0 \text{ then } q = 0 \text{ and } u = u_{\min} \\ \text{When } 0 > \lambda_3 \text{ then } q = 0 \text{ and } u = 1 \end{array}\right\} \tag{32}$$

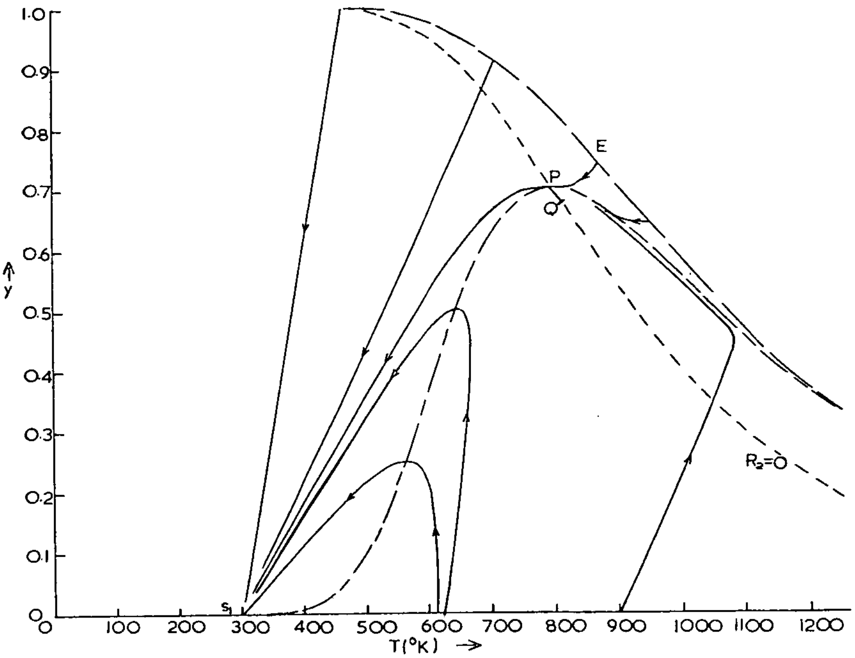


FIG. 4. Solution trajectories for $u = 1{\cdot}0, q = 0$.

q and u may take values in the interior of their permissible ranges, while satisfying the maximum principle, only if $\lambda_3 = \gamma c$ or $\lambda_3 = 0$ respectively and we shall investigate these possibilities further in a later section.

The situation in which the control variables must take one or other of their limiting values in order to satisfy the maximum principle is very familiar in the theory of optimum control, and is said to correspond to "bang-bang" operation. The name refers to the fact that the optimum control policy may be divided into a sequence of time intervals in each of which the control variables are constant at one or other of their extreme values, with sudden changes from one set of extreme conditions to another at the end points of the intervals.

Neglecting, for the moment, the possibility of q or u taking values in the interior of their permitted ranges, it is seen that the physical behaviour of the system during optimum startup will be represented by segments of three sets of solutions of differential Eqs. (6) and (7), corresponding to the three pairs of values of q and u given by conditions (32). These segments must be joined together in such a way that the switches from one to another occur when λ_3 passes through the values 0 and γc, where λ_3 is obtained by solving the adjoint Eqs. (27) and (28). The solutions of the physical Eqs. (6) and (7) may be plotted parametrically in the (y, T)-plane with time as a parameter as shown in Figs. 4, 5 and 6, which represent solutions for the three pairs of values of q and u determined by conditions (32). Specifically, these diagrams represent solutions of Eqs. (6) and (7) with the rate equations given by Eqs. (12) and (13) and the following numerical values of the kinetic constants.

$$\tau k_{01} = 2417, \qquad \tau k_{02} = 2{\cdot}683 \times 10^5,$$

$$e_1 = 5000°\text{K}, \qquad e_2 = 10{,}000°\text{K}$$

It was also assumed that $\Delta T_{ad} = 500°K$, $T_0 = 300°\text{K}$, $\gamma = 1$, $u_{\min} = 0{\cdot}5$ and $q_{\max} = 100°\text{K}$. The trajectories shown are based on numerical solutions of the differential equations with a little interpolation, but actually it is quite easy to see their general features without any detailed calculation, using the method of isoclines [4].

Reference to Fig. 5 clears up a point raised earlier in the discussion. On this diagram there are seen to

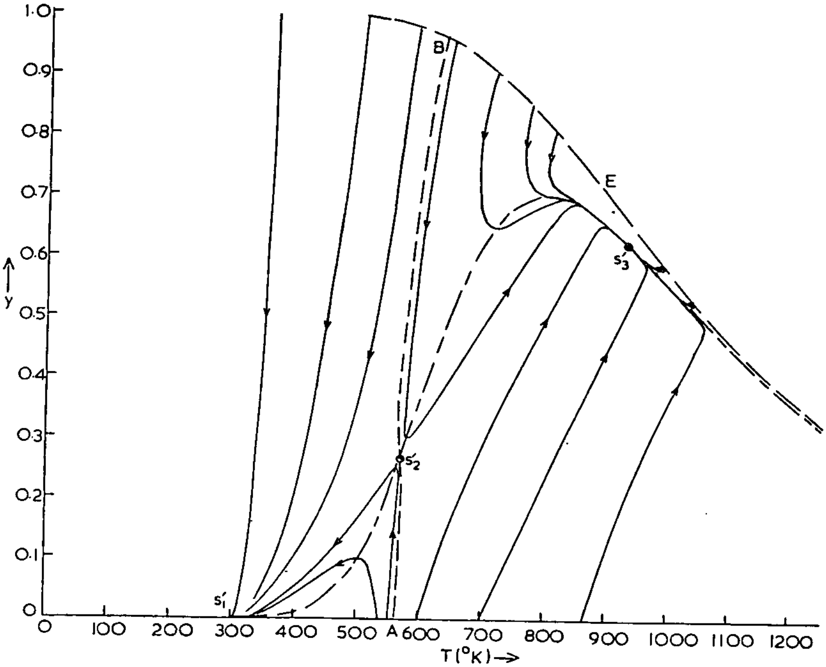


FIG. 5. Solution trajectories for $u = 0{\cdot}5, q = 0$.

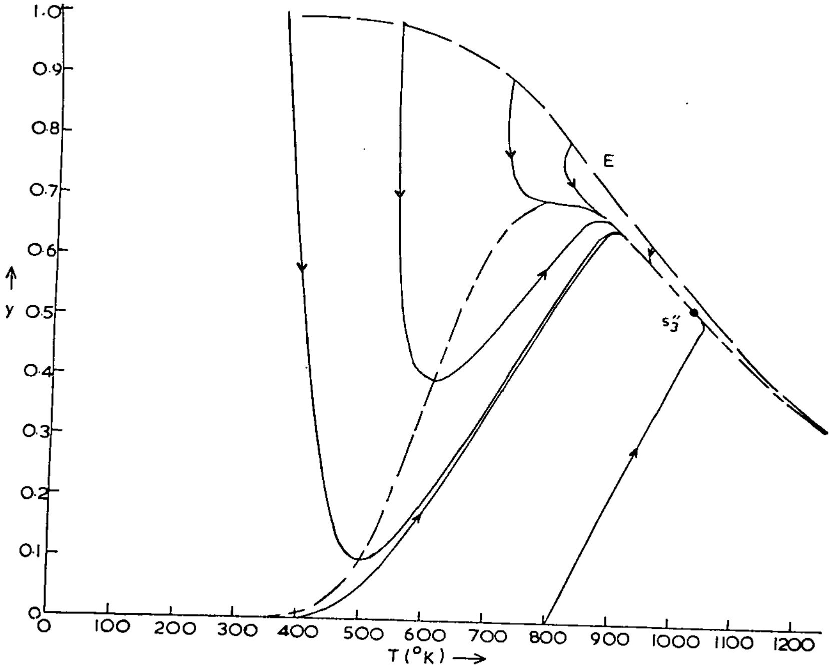


FIG. 6. Solution trajectories for $u = 0{\cdot}5, q = 100°C$.

be three steady states, labelled s_1', s_2' and s_3' to correspond to the notation of Fig. 2, and it is clear from the form of the trajectories that s_2' represents an unstable steady state, since all trajectories in its neighbourhood lead away from it. s_1' and s_3' on the other hand, represent the two stable steady states at low and high temperature respectively. The broken curve $A\,s_2'\,B$ represents a *separatrix* dividing the (y, T)-plane into two parts containing trajectories which converge to different steady states. Starting from any initial conditions to the left of this curve the system eventually settles in the state s_1', while from initial conditions to the right of the curve it converges to the autothermic state s_3'.

Figure 4 represents the solutions when there is no external heat supply and no regenerative heating, and it is seen that there is a single steady state s_1 with negligible reaction. This is the initial state to be considered in the startup problem. Figure 6 shows the solutions for conditions of maximum external and regenerative heating and once again there is a single steady state, in this case the autothermic state s_3''. The labelling of the steady states in Figs. 4 and 6 is chosen, once again, to correspond to the nomenclature of Fig. 2.

On all three diagrams the region of physical interest is bounded above by the broken curve E, on which the condition of chemical equilibrium $R(y,T) = 0$ is satisfied. Also of interest is the chain dotted curve $R(y, T) = y$, which represents the locus of points for which $dy/dt = 0$. All steady states, whatever the values of u and q, must lie on this curve, and it is clear from its shape that the steady state conversion passes through a maximum value $y = 0{\cdot}7$, so it follows that these conditions also maximise the steady state objective function

$$P_s = y_s - cq_s \tag{33}$$

and therefore represent the optimum conditions for steady operation.

We can now describe, in principle, the procedure to be followed in computing a solution to the optimum startup problem. It is convenient to start with the system in its final steady state $(y_s\, T_s)$ at $t = \theta$, and work backwards in time to the initial state (y_0, T_0) corresponding to point s_1 in Fig. 4. In the final state y and T take specified values y_s and T_s but λ_2 and λ are not determined by any boundary conditions. Nevertheless they are not free to be chosen arbitrarily and independently, since according to Eq. (23) the maximised value of the Hamiltonian must vanish at all times, including $t = \theta$. Thus only one variable, say λ_2, may have its value arbitrarily fixed. Having chosen a value for λ_2 Eqs. (6), (7), (27) and (28) are integrated backwards in time, with q and u determined throughout by the value of λ_3 through conditions (32). These ensure that the values of the control variables will change whenever λ_3 passes through one or other of the switching values 0 and γc which appear in conditions (32). In general the solution generated will not pass through the specified initial point (y_0, T_0), and it is necessary to adjust the value assumed for λ_2 at $t = \theta$ until a solution is obtained which does match the initial conditions.

The solution obtained in this way satisfies the necessary condition provided by the maximum principle for the minimisation of the objective function P. If it proves to be the only solution with this property it can be demonstrated that the condition is sufficient as well as necessary, so it provides the result we seek.

OPTIMUM STARTUP TO THE OPTIMUM STEADY STATE

We have already seen that the optimum stead state is given by $y_s = 0{\cdot}7$, $T_s = 800$°K, and the corresponding values of the control variables are $u = 0{\cdot}7, q = 0$. We now consider the optimum startup procedure, starting from the initial state s of Fig. 4 and terminating in this optimum state.

Following the procedure just outlined, we attemp to trace the optimum trajectory in the (y, T)-plan backwards in time from the final state $(y_s\, T_s)$. W also require that the trajectory should consis entirely of segments of the curves shown in Figs. 4–6 and this immediately gives rise to difficulty, since i each of these diagrams the trajectory passing throug the optimum state $(y_s\, T_s)$ leads upwards across th equilibrium curve E into the region of no physica interest when it is followed backwards in tim Thus there is no possibility of joining the final stat to the specified initial state by a trajectory compose entirely of segments from these three diagrams.

The resolution of this difficulty follows from th fact we have already noted, that conditions (32) are not completely exhaustive. It is possible that the maximum principle can be satisfied with u in the interior of its permitted interval if $\lambda_3 = 0$, for then the Hamiltonian (31) is independent of u. In the same way q may take a value within its permitted interval if $\lambda_3 = \gamma c$. The second of these possibilities does not appear to have any bearing on the problem of optimum startup but, as we shall show, the first is of vital importance.

Let us, then, consider the possible existence of a finite time interval during which u lies between its specified limits. Then λ_3 must vanish throughout this interval if the maximum principle is to be satisfied, and it follows that $d\lambda_3/dt$ must also vanish. Consequently Eqs. (27) and (28) reduce to

$$\frac{d\lambda_2}{dt} = -1 - \lambda_2(R_1 - 1) \tag{34}$$

and

$$\lambda_2 R_2 = 0 \tag{35}$$

Furthermore, the Hamiltonian must be maximised with respect to q, and when $\lambda_3 = 0$ this implies that q must vanish. Then according to equation (23) the corresponding maximum value of the Hamiltonian must also vanish, so we have

$$y - y_s + \lambda_2(R - y) = 0 \tag{36}$$

Equations (34), (35) and (36) must be satisfied throughout the interval for which u takes interior values. Considering first Eq. (35), this cannot be satisfied by taking $\lambda_2 = 0$ throughout the interval, for this would imply that $d\lambda_2/dt = 0$ and Eq. (34) would reduce to a contradiction. Thus it is necessary that $R_2(y, T) = 0$, and this determines a trajectory in the (y, T)-plane which must represent the required solution of Eqs. (6) and (7). It is indicated in Fig. 4, from which it is seen to pass through the specified final state at the highest point of the curve $R(y, T) = y$.

Not all parts of the curve $R_2 = 0$ can represent a solution of Eqs. (6) and (7) with $q = 0$ and u constrained to lie within its permitted interval; indeed this is only possible if there exists an interior value of u which makes the ratio of the right-hand sides of Eqs. (6) and (7) equal to the slope of $R_2 = 0$, so that

$$-\frac{R_{22}}{R_{21}} = \frac{dy}{dT} = \frac{R - y}{\Delta T_{ad}R - u(T - T_0)} \tag{37}$$

Equation (37) has a simple geometrical interpretation. It is satisfied whenever the direction of the curve $R_2 = 0$ lies between the directions of the trajectories of Figs. 4 and 5 respectively, which is seen to be the case on a short segment PQ below the final state P, as indicated in Fig. 4. This segment may therefore form part of an optimal trajectory, and on it u takes the values determined by Eq. (37). Actually the value of λ_2 is also determined at all points of this segment by Eq. (40), so there is no freedom of choice in the values of λ_2 or λ_3.

The segment PQ provides the desired means of escape from the point P in a manner which permits the state s_1 to be reached by trajectories which everywhere satisfy the maximum principle. Moving backwards in time from P along this segment towards Q, λ_3 vanishes at all points, so according to conditions (32) it is possible to switch to a trajectory from Fig. 4 or Fig. 5 at any point of PQ. Once this switch has been made the remainder of the startup procedure is determined by the "bang-bang" conditions (32) and can be generated in the manner described at the end of the previous section. The value of λ_2 at the start of the "bang-bang" segments is no longer available, since we have seen that Eq. (36) determines λ_2 at each point of PQ, but we have instead the freedom to leave PQ at any point and can choose our point of departure to ensure that the trajectory eventually passes through the specified initial state s_1.

Scrutiny of Figs. 4–6 reveals that we cannot leave PQ on a trajectory drawn from Fig. 5 if we are ever to reach s_1, so at the point of departure we must take $u = 1{\cdot}0$, $q = 0$ and proceed backwards in time along a trajectory from Fig. 4. It is then found that λ_3 falls below zero, passes through a minimum, then rises to change sign. At this point, according to conditions (32), we must change the control variables to $u = 0{\cdot}5$, $q = 0$ and proceed down a trajectory drawn from Fig. 5. λ_3 meanwhile continues to increase monotonically, and when it reaches the value γc we must switch the values of the control variables to $u = 0{\cdot}5$, $q = q_{\max}$ and continue along a trajectory from Fig. 6. No further switching conditions are encountered, so if it is to complete the solution this trajectory must pass through s_1. In general, of course, it does not, and in order to obtain the desired solution for given values of c and

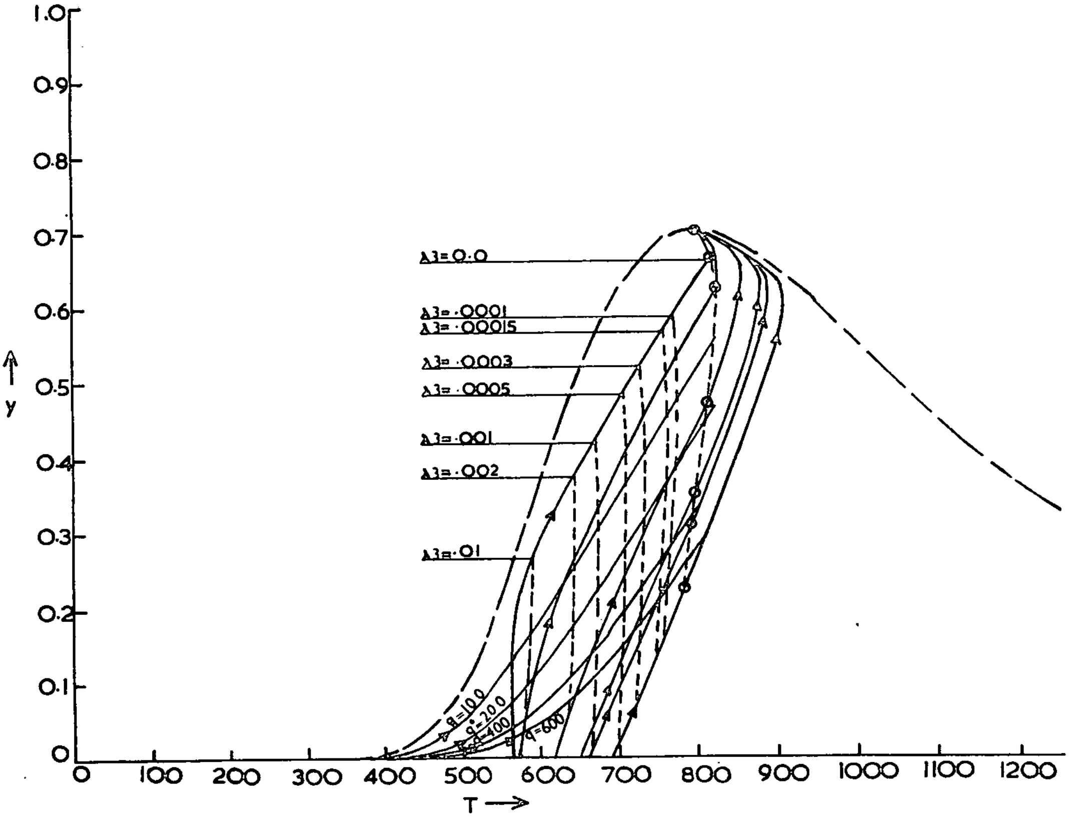


FIG. 7. Chart to determine optimum startup to the optimum steady state.

$q_{\max}$ it would be necessary to repeat the whole procedure from a different starting point on PQ until a trajectory was found which did pass through s_1.

In practice this iterative adjustment can be avoided if we seek solutions, not just for a single specified pair of values of c and $q_{\max}$, but for a range of values of each parameter. Suppose we ran a solution backwards in time from a point on PQ, as described above, as far as the segment with $u = 0{\cdot}5$, $q = 0$. We also run a trajectory with $u = 0{\cdot}5$, $q = q_{\max}$ forwards in time until it interesects this segment. Then the value reached by λ_3 in the backwards integration at the intersection point gives γc, for the complete solution obtained by joining these trajectories. By running solutions backwards from various points of PQ and forwards from s_1 with various values of $q_{\max}$, we can therefore generate complete solutions for various values of $q_{\max}$ and c, as required.

The results of numerical integrations to determine the trajectories are plotted in this way in Fig. 7, where the continuous curves represent solution trajectories. The broken curves joint points on these trajectories with equal values of λ_3 and the numerical value of λ_3 corresponding to each curve is indicated. In particular the curve $\lambda_3 = 0$ gives the locus of switching points between conditions $u = 1{\cdot}0, q = 0$ and $u = 0{\cdot}5$, $q = 0$. If we require the complete trajectory representing the optimum startup procedure for given values of $q_{\max}$ and c, we start at state s_1 and follow the appropriate trajectory $u = 0{\cdot}5, q = q_{\max}$ forwards in time until it intersects the curve $\lambda_3 = \gamma c$, with the specified value of c. We then follow the trajectory with $u = 0{\cdot}5$, $q = 0$ from this point to the curve $\lambda_3 = 0$, the consequent trajectory $u = 1{\cdot}0$, $q = 0$ to the segment PQ, and the segment PQ to the final steady state P.

If γc is large the solution trajectory passes close to the unstable steady state s'_2, and in the limit as

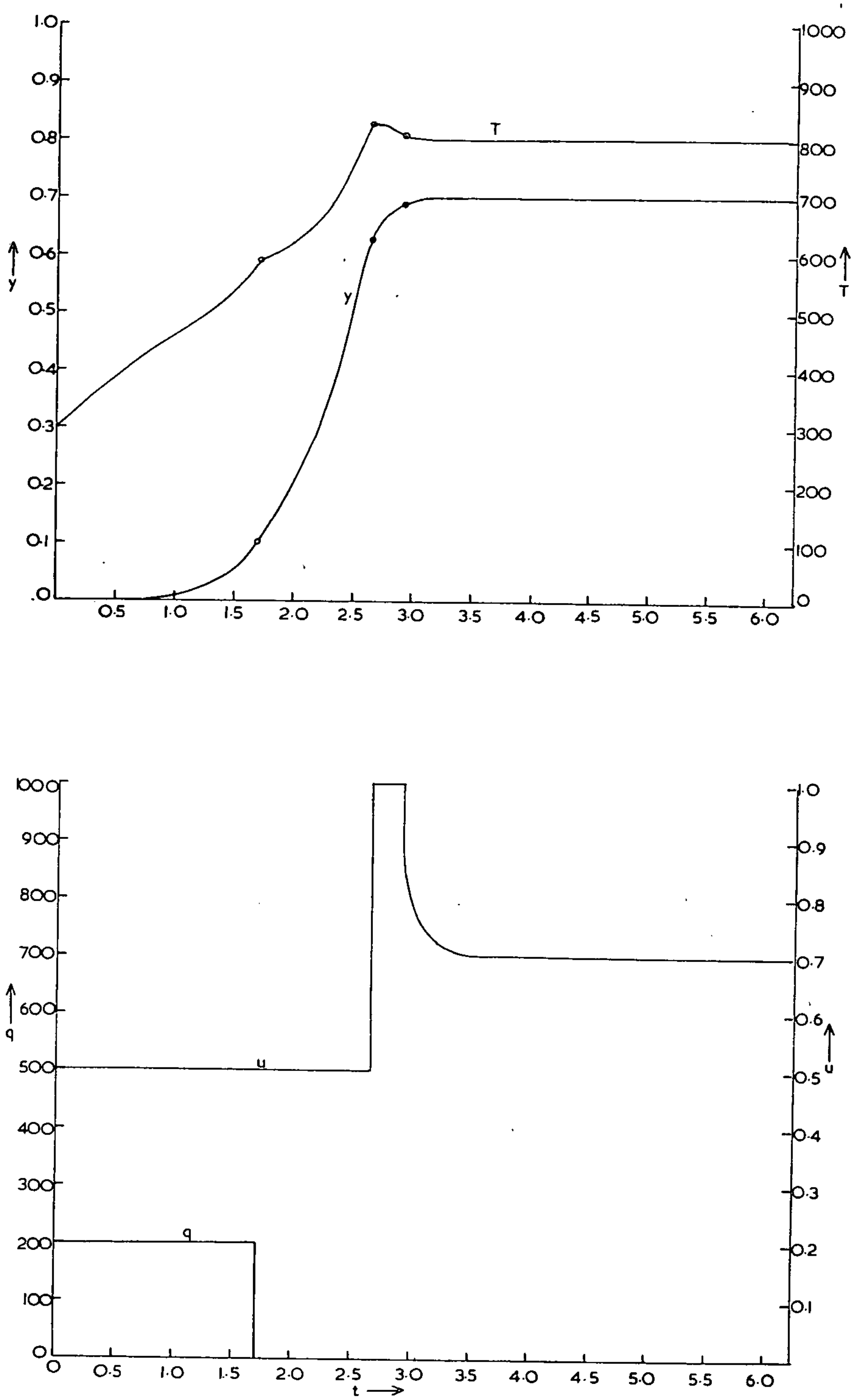


FIG. 8. Optimum startup procedure to the optimum steady state with $u_s = 0{\cdot}7$, $q_s = 0$ when $u_{\min} = 0{\cdot}5$, $q_{\max} = 200°\mathrm{C}$ and $c = 0{\cdot}008$.

$\gamma c \to \infty$ it actually passes through this state. This is physically reasonable since a large value of γc corresponds to a high cost of external heating, and the solution passing close to s'_2 cuts the external heating to the minimum necessary to initiate ignition to the autothermic state. When γc is small, on the other hand, the external heat supply is retained almost to the point where the trajectory $q = q_{\max}$, $u = 0{\cdot}5$ crosses the curve $\lambda_3 = 0$, and correspondingly the segment with $q = 0$, $u = 0{\cdot}5$ becomes very short. Once again this is reasonable. When external heating is cheap it is used to the fullest extent to promote rapid startup.

Instead of plotting the behaviour during startup parametrically in the (y, T)-plane, as in Fig. 7, the values of y, T and the control variables u and q can be plotted separately as functions of time. Fig. 8 shows the optimum startup procedure for $q_{\max} = 200°C$, $c = 0{\cdot}0081$, plotted in this way. The values of u and q are determined by conditions (32) everywhere except on the final segment PQ, where u is computed from Eq. (37). It is seen that the system is brought on line in its final steady state in a time a little greater than three times the mean residence time in the reactor.

It is interesting, and perhaps a little unexpected, to see that the optimum startup procedures include a short interval during which all heating, both external and regenerative, is cut off. During this interval y increases without very much increase in temperature as a result of cooling by the cold reactant stream entering the reactor. It seems probable that this feature of the optimum startup policy would be absent if the approximation of perfect mixing in the reactor were replaced by the other extreme assumption of no axial mixing.

OPTIMUM STARTUP TO A NON-OPTIMAL STEADY STATE

We have now seen how best to bring the system to its economic optimum state of steady operation, and it is not immediately obvious why one should be interested in any other steady state. The principal reason for operating in a non-optimal steady state is illustrated by Fig. 9, which shows trajectories describing the dynamical behaviour of the system in the (y, T)-plane for $u = 0{\cdot}7, q = 0$. These are the values of u and q required for steady operation in the optimum state P, so the trajectories illustrate what

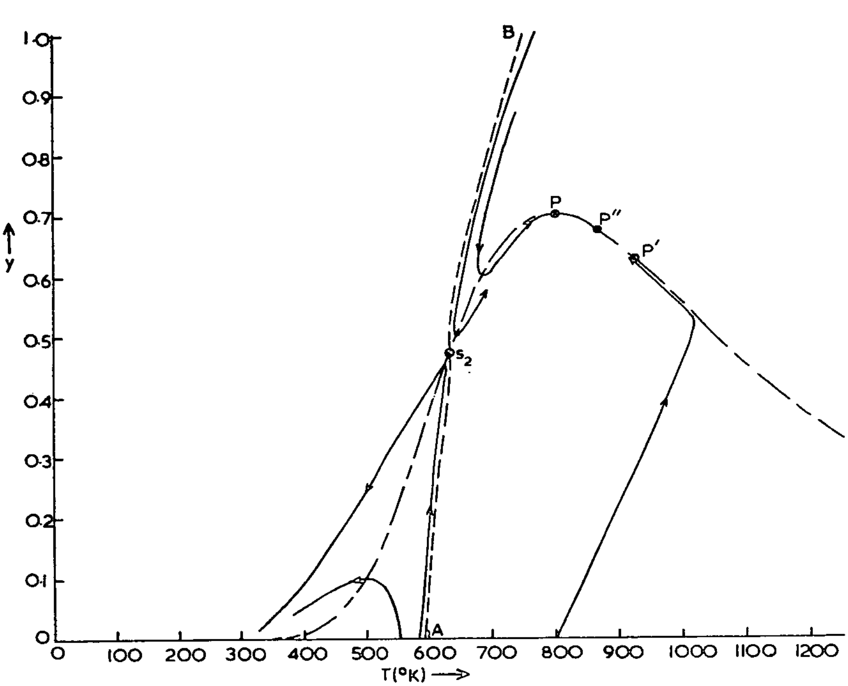


FIG. 9. Region of stability in relation to steady operating states.

will happen if the system is displaced from P by some disturbance and no correcting action is taken. For sufficiently small disturbances it is seen that the system will return to state P, but if the disturbance is large enough to carry the representative point across the separatrix AB the system will settle in the low temperature state s_1 and the reaction will be extinguished. Furthermore, AB passes fairly close to point P, and with different values for the constants in Eqs. (6) and (7) than those used here, it may pass even closer. Thus there is a certain danger that accidental disturbances affecting the system will cause the reaction to be extinguished if prompt corrective action is not taken. Of course the proper way of dealing with this difficulty, in principle, is to devise a control system which is capable of supplying the required prompt corrections, but in practice the margin of stability available can be increased substantially simply by operating at rather lower conversions in steady states represented by points to the right of P. The steady state can be moved to the right by reducing the value of u, eventually reaching the position P' when u reaches its minimum value, 0·5, but to go further than this a finite value of q would be needed, and this would be expensive. Thus the only steady states likely to be of interest, other than P, are those lying between P and P', and we shall illustrate the optimum startup procedure in such cases by considering the particular state P'' with $y_s = 0{\cdot}674$ and $T_s = 870°\mathrm{K}$, corresponding to $u_s = 0{\cdot}591, q_s = 0$.

The state P'' does not lie on the segment PQ of Fig. 4, on which u may take interior values, so in this case we would expect the startup procedure to terminate in the "bang-bang" mode, in contrast to the case already discussed. Following the procedure outlined earlier, we attempt to trace the optimal trajectory in the (y, T)-plane backwards in time from P'', so we must first decide which of the three types of trajectory distinguished by conditions (32) is to be used in leaving P''.

A glance at Fig. 4 reveals that the trajectory with $u = 1{\cdot}0, q = 0$ passing through P'' leads backwards across the equilibrium curve E into the region of no physical interest, and can therefore be eliminated from further consideration. If $q = q_{\max}$, $u = u_{\min}$ on leaving P'', corresponding to the trajectories of Fig. 6, the Hamiltonian (30) reduces to

$$H = -cq_{\max} + \lambda_3[q_{\max} - u_{\min}(T_s - T_0) + \Delta T_{\mathrm{ad}} R(y_s T_s)]/\gamma$$

at point P'', where $R(y, T) = y$ and $y = y_s$. Since the Hamiltonian must vanish, it then follows that

$$\lambda_3 = \frac{\gamma c q_{\max}}{q_{\max} - u_{\min}(T_s - T_0) + \Delta T_{\mathrm{ad}} R(y_s T_s)}$$

However, since P'' is a steady state with $u = u_s$, $q = 0$, it follows from Eq. (7) that $\Delta T_{\mathrm{ad}}\, R(y_s, T_s) = u_s(T_s - T_0)$ and the above reduces to

$$\lambda_3 = \frac{\gamma c}{1 + \dfrac{(T_s - T_0)(u_s - u_{\min})}{q_{\max}}} < \gamma c$$

But according to conditions (32), the values $q = q_{\max}$, $u = u_{\mathrm{m}}$ for the control variables only maximise the Hamiltonian when $\lambda_3 > \gamma c$, so it follows that H is not maximised and the maximum principle is not satisfied on the trajectory considered.

We are left, therefore, with a single acceptable possibility for leaving P'', namely the trajectory with $u = 0{\cdot}5$, $q = 0$. The value $q = 0$ is identical with the value q_s required for steady state operation at P'', so the Hamiltonian (30) now reduces to

$$H = \lambda_3[-0{\cdot}5(T_s - T_0) + \Delta T_{\mathrm{ad}} R(y_s T_s)]$$

at point P''. As we have already seen, $\Delta T_{\mathrm{ad}}\, R(y_s, T_s) = u_s(T_s - T_0) = 0{\cdot}591(T_s - T_0)$, so it follows that the content of the square bracket above is non-vanishing, and the condition that H should vanish implies that $\lambda_3 = 0$. The value of λ_2 at P'' remains available to be varied so as to ensure that the trajectory passes through the specified initial state represented by point s_1 of Fig. 4. According to conditions (32), if the values $u = 0{\cdot}5$, $q = 0$ are to satisfy the maximum principle immediately after leaving P'', λ_3 must become positive. Thus $d\lambda_3/dt$ must be positive at the initial point P'', and since $\lambda_3 = 0$ it follows from Eq. (32) that the value chosen for λ_2 must be negative.

On integrating Eqs. (6), (7), (27) and (28) with $u = 0{\cdot}5, q = 0$, backwards in time from P'', starting with a negative value for λ_2 and $\lambda_3 = 0$, it is found that λ_3 first increases to a maximum, then decreases to a minimum and finally increases monotonically.

If the initial value of λ_2 is sufficiently large and negative, λ_3 remains positive throughout these variations, and the conditions $u = 0{\cdot}5$, $q = 0$ continue to satisfy the maximum principle until λ_3 reaches the value γc, when a switch must be made to a trajectory with $u = 0{\cdot}5$, $q = q_{\max}$. If this fails to pass through the specified initial state s_1, the value of λ_2 used to start the integration is available to be adjusted. Thus the optimum startup trajectory consists simply of two segments joining s_1 and P''. When the value of λ_2 used to start the integration is smaller in magnitude, however, the situation is more complicated, since the minimum value through which λ_3 passes becomes negative. Thus λ_3 passes through a maximum after leaving P'', then falls to zero, at which point it is necessary to switch to a trajectory with $u = 1{\cdot}0, q = 0$. Pursuing this backwards in time λ_3 decreases, passes through a minimum, then increases once again and changes sign. At this second change of sign we must revert to a trajectory with $u = 0{\cdot}5$, $q = 0$, and this is followed until λ_3 reaches the value γc, when a final switch is made to a trajectory with $u = 0{\cdot}5$, $q = q_{\max}$. Once again the value of λ_2 used to start the integration is available for adjustment to ensure that the trajectory finally passes through state s_1. The complete optimal trajectory joining s_1 and P'' now consists of four separate segments.

As discussed earlier, the iterative adjustments of λ_2 can be avoided if we generate, not one solution, but solutions for a range of values of $q_{\max}$ and c. The method is the same as that described previously. Optimal trajectories are run backwards in time from P'' to meet trajectories with $u = 0{\cdot}5$, $q = q_{\max}$ run forwards in time from s_1. The value of λ_3 obtained by backwards integration at the intersection of any such pair of trajectories gives the value of γc corresponding to the complete optimal startup procedure represented by the trajectories taken together. Thus the complete optimal startup pro-

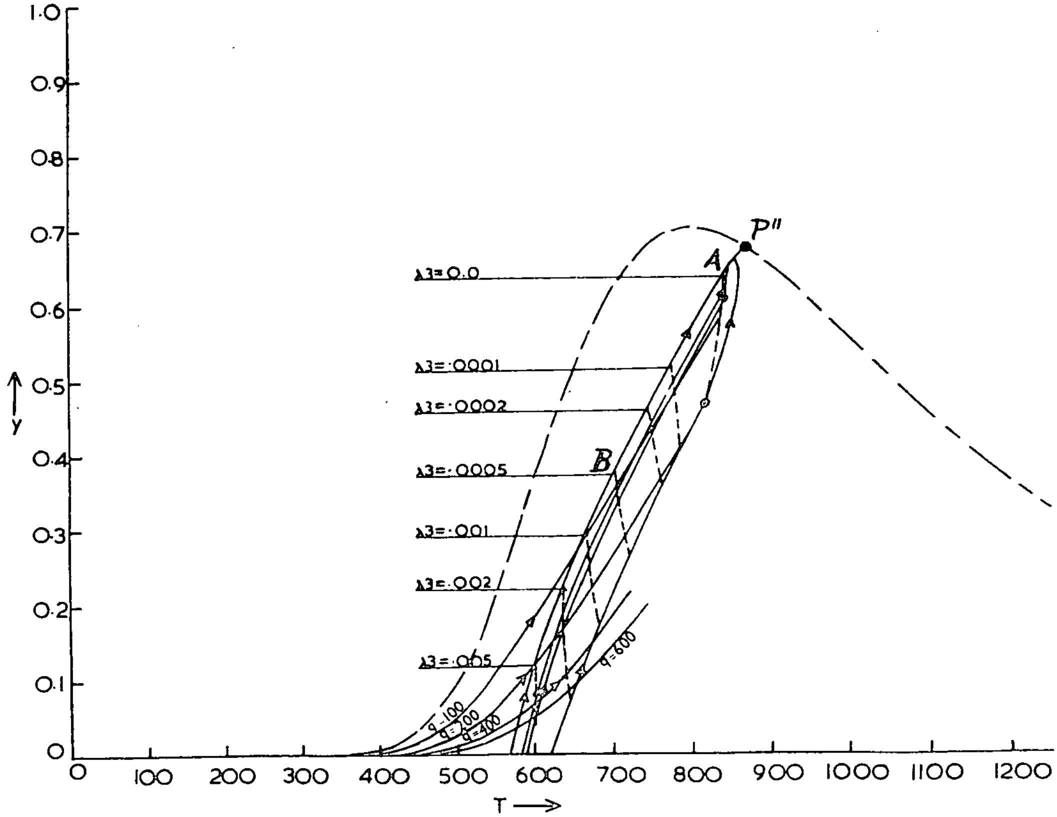


FIG. 10. Chart to determine optimum startup to a non-optimum steady state.

cedure for any pair of values of $q_{\max}$ and γc can be picked out if we superimpose on the trajectory map in the (y, T)-plane a set of curves joining points on different trajectories with equal values of λ_3. The results are presented in this way in Fig. 10. Here $P'' AB$ is the trajectory with $u = 0{\cdot}5$, $q = 0$ passing through P''.

For sufficiently small negative values of $\lambda_2(P'')$, the optimal paths leave this trajectory, as indicated between P'' and A, along segments with $u = 1{\cdot}0$, $q = 0$, and these in turn are relinquished for other trajectories with $u = 0{\cdot}5$, $q = 0$, as discussed above. The broken curves join points on the different trajectories with common values of λ_3, and the value of λ_3 corresponding to each curve is indicated.

The trajectory $P'' AB$ differs from the others in that λ_3 is not uniquely defined along it. This arises from the fact, noted above, that decreasing $\lambda_2(P'')$ below a certain value has no effect on the optimal trajectory, which coincides with $P'' AB$, but it does change the values of λ_3 at points along this trajectory. Effectively, we may take it that all the curves $\lambda_3 =$ const. change direction where they meet $P'' AB$, and follow this trajectory upwards towards P''. Thus they become superimposed.

To pick out the trajectory representing the complete optimum startup procedure for given values of $q_{\max}$ and c, the procedure is similar to that described earlier. The trajectory $q = q_{\max}$ is followed forwards in time from the specified initial states s_1 until it meets the curve $\lambda_3 = \gamma c$ which may occur either on the part of this curve indicated by a broken line, or on the part lying along $P'' AB$. If the intersection occurs on the part of $\lambda_3 = \gamma c$ which is superimposed on $P'' AB$, the optimal trajectory is completed by following $P'' AB$ upwards to the final state P'', and correspondingly the control variables are switched to $u = 0{\cdot}5$, $q = 0$, which values they retain up to P''. If the intersection occurs on the part of $\lambda_3 = \gamma c$ indicated by a broken curve, on the other hand, the trajectory springing from this intersection point is followed through two further switching points to P''.

For given values of $q_{\max}$ and c the results may alternatively be plotted as graphs of y, T, u and q against time, and Fig. 11 presents them in this way for $q_{\max} = 200$°C, $c = 0{\cdot}0018$. It should be compared with Fig. 8, which gives the same information for startup to the optimum steady state.

The optimum startup procedure to the optimum steady state always includes an interval during which all heating, both external and regenerative, is withdrawn, as can be seen from Fig. 7. In the present case, however, such an interval occurs only if c is sufficiently small. With larger values of c there is only one switching point in the optimum trajectory, and regenerative heating is always employed to the full. This is physically reasonable, since it serves to economise in the use of external heating at the expense of a rather slower startup when external heat is expensive.

It should not be assumed that the form of the optimum trajectories is as indicated by Fig. 10 for all positions of the final steady state P''. When P'' lies sufficiently close to the optimum steady state P, the optimum startup procedure makes use of part of the segment PQ in Fig. 4 on which u may take interior values, so the trajectory may not represent exclusively "bang-bang" operation. The form of the optimum startup procedure in such cases can be obtained by reasoning entirely analogous to that already presented here.

Conclusion

It has now been shown how the optimal startup procedure can be found for a specified final state of steady operation, and the result is quite complex, involving a number of changes of conditions, all of which must be correctly timed. It is therefore of interest to consider a very simple startup procedure and see how far it falls short of the optimal performance. The system can be brought to its final steady state very simply by first setting $q = q_{\max}$ and $u = u_{\min}$, thereby providing the maximum possible heating effect, then switching to the values $q = q_s$, $u = u_s$ required for steady operation in the specified final state. Provided the heating phase is of sufficiently long duration, the system will then settle eventually to the required steady state (y_s, T_s). (If heating is not continued long enough to induce "ignition" of course, it will revert to a state with negligible reaction.) The value of the objective function P can be computed as a function of the duration of the first (heating) phase, t_0, and one would expect it to be minimised for some value of

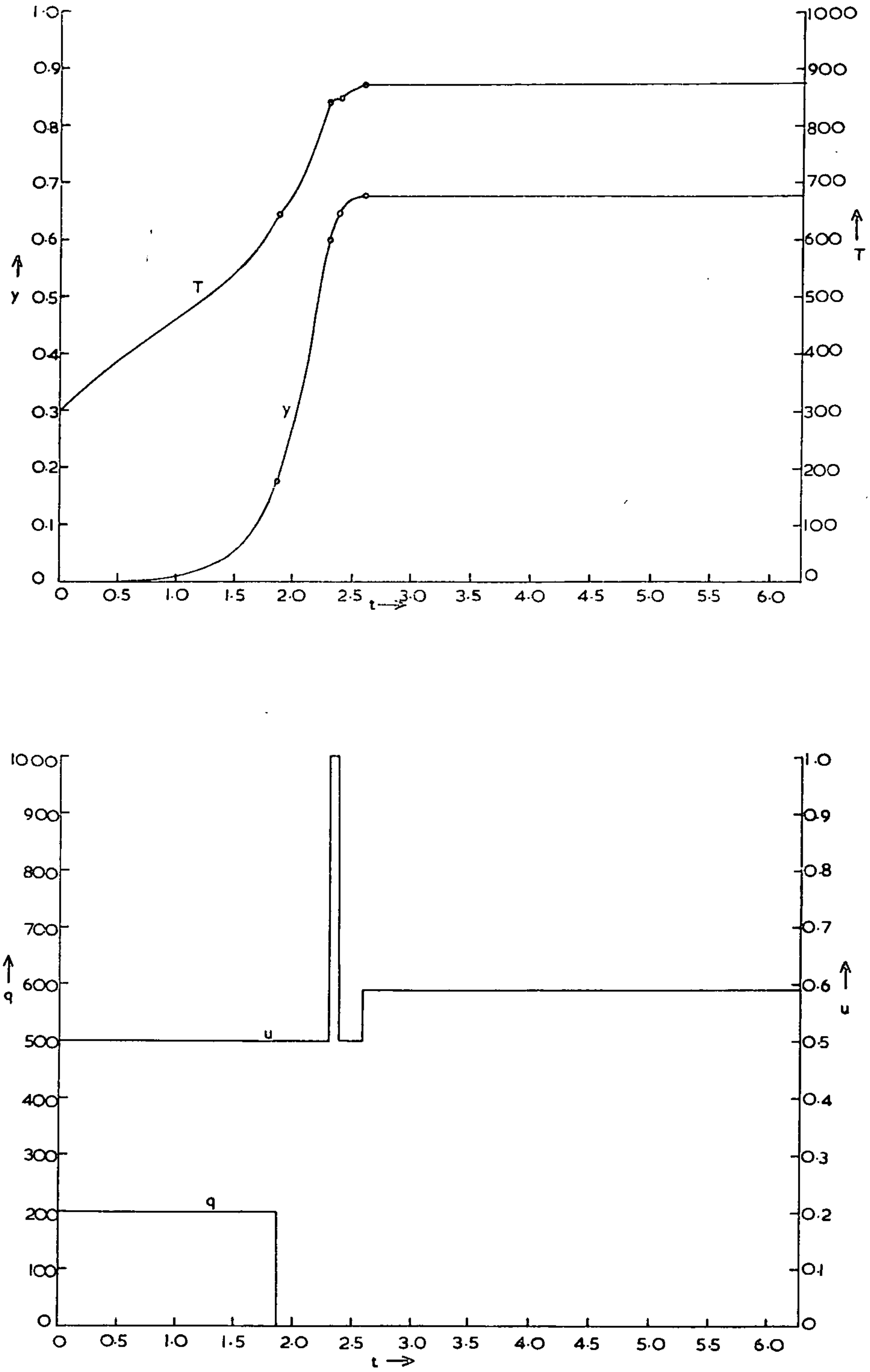


FIG. 11. Optimum startup procedure to the non-optimum steady state $u_s = 0{\cdot}674$, $q_s = 0$ when $u_{\min} = 0{\cdot}5$, $q_{\max} = 200°\text{C}$ and $c = 0{\cdot}0018$.

Table 1

$q_{\max}$	c	P_1	P_2
100	0·0992	47·5	50·0
200	0·0081	4·33	4·45
600	0·000195	0·616	0·653

t_0, which would then determine the optimum startup procedure within this limited class of possibilities.

The value of the objective function P for startup to the optimum steady state ($y_s = 0{\cdot}7$, $T_s = 800°\text{K}$) has been computed for various values of the parameters $q_{\max}$ and c, and the results are given in Table 1. P_1 is the value of P corresponding to the optimum startup procedure deduced from the maximum principle as described earlier in this paper, while P_2 is the lowest value of P obtainable by adjusting the value of t_0 in a simple startup procedure of the type just described.

P_1 is necessarily smaller than P_2, but in no case is the difference very large. Of course these results refer only to specific cases, and it would be unwise to generalise on this basis, but they at least suggest that it is worthwhile, in specific problems, to examine the quality of a simple startup procedure in comparison with the absolute optimum. If the difference is small, as it is in the examples just considered, an almost optimum startup can be achieved with a single change of conditions, and the optimum time t_0 for this change could be found by experimental trial.

NOTATION

A	Heat transfer surface area in exchanger
c	Ratio p_2/p_1F
C	Thermal capacity of reactor per unit molar capacity
C_g	Molar specific heat of reaction mixture
e_1, e_2	Activation energies of forward and reverse reactions, divided by the universal gas constant
f	Molar flow rate through hot side of exchanger
f_i	Right-hand sides of differential equations in the general statement of the maximum principle
F	Molar flow rate through reactor
h	Mean overall heat transfer coefficient in exchanger
H	Pontryagin's Hamiltonian
ΔH	Heat of reaction
k_{01}, k_{02}	Frequency factors in velocity constants of forward and reverse reactions
k_1, k_2	Quantities $\tau k_{01}\exp(-e_1/T)$ and $\tau k_{02}\exp(-e_2/T)$ respectively. Proportional to velocity constants of forward and reverse reactions
p_1, p_2	Costs for product and externally supplied heat respectively
P	Objective function to be minimised
P'	Loss of profit due to startup
P_a	Net profit during startup period
P_v	Net profit for steady operation during startup
P_s	Value of P in steady state
q	Temperature rise in heater
q_s	Value of q in steady state
$q_{\max}$	Upper bound for q
$r(y, T)$	Reaction rate
$R(y, T)$	Scaled reaction rate, equal to $\tau r(y,T)$
R_1	Partial derivative $\partial R/\partial y$
R_2	Partial derivative $\partial R/\partial T$
R_{21}	Second derivative $\partial^2 R/\partial y \partial T$
R_{22}	Second derivative $\partial^2 R/\partial T^2$
t	Dimensionless time. $t = t'/\tau$
t'	Time
T	Temperature at reactor exit
T_0	Temperature at entry to cold side of exchanger
T_1	Temperature at exit from cold side of exchanger
T_1'	Temperature at entry to reactor
T_s	Value of T in steady state
ΔT_{ad}	Adiabatic temperature rise for reaction
u	Control variable for exchanger. Related to f by Eq. (4)
u_s	Value of u in steady state
$u_{\min}$	Lower bound for u
V	Molar capacity of reactor
w_p	Control variables in general statement of maximum principle
x_i	Dependent variables in general statement of maximum principle
x_{0i}	Initial values of the x_i
x_{fi}	Terminal values of the x_i
y	Mole fraction of product in mixture leaving reactor.
y_s	Steady state value of y
z	Auxiliary variable, defined by Eqs. (24) and (25)
α	Constant characteristic of exchanger. $\alpha = hA/C_gF$
α_i	Constants defining objective function in general statement of maximum principle
γ	Ratio C/C_g
θ	Value of t at which the steady state is attained
λ_1	Variable adjoint to z
λ_2	Variable adjoint to y
λ_3	Variable adjoint to T
λ_i	Variables adjoint to the x_i
τ	Mean residence time in reactor.

REFERENCES

[1] VAN HEERDEN C., *Chem. Engng Sci.* 1958 **8** 133.

[2] ROZONOER L. I., *Autom. remote Control.* 1959 **20** 1288, 1405, 1517.

[3] SIEBENTHAL C. D. and ARIS R., *Chem. Engng Sci.* 1964 **19** 729, 747.

[4] PIAGGIO H. T. H. *Differential Equations*, Chap. 1. Bell, London 1948.

Résumé—Certaines réactions exothermiques d'une grande importance commerciale, utilisent un échange thermique régénérateur entre les courants de réactant et de produit, de telle sorte que lorsqu'on atteint le régime d'équilibre, la chaleur fournie par la réaction compense les besoins thermiques du procédé. Cependant leurs démarrages exigent un apport de chaleur aux réactants par une source extérieure, bien que celle-ci puisse être éloignée dès que l'"ignition" est terminée.

Ce fait donne de l'importance au problème de la détermination d'un régime correct de démarrage.

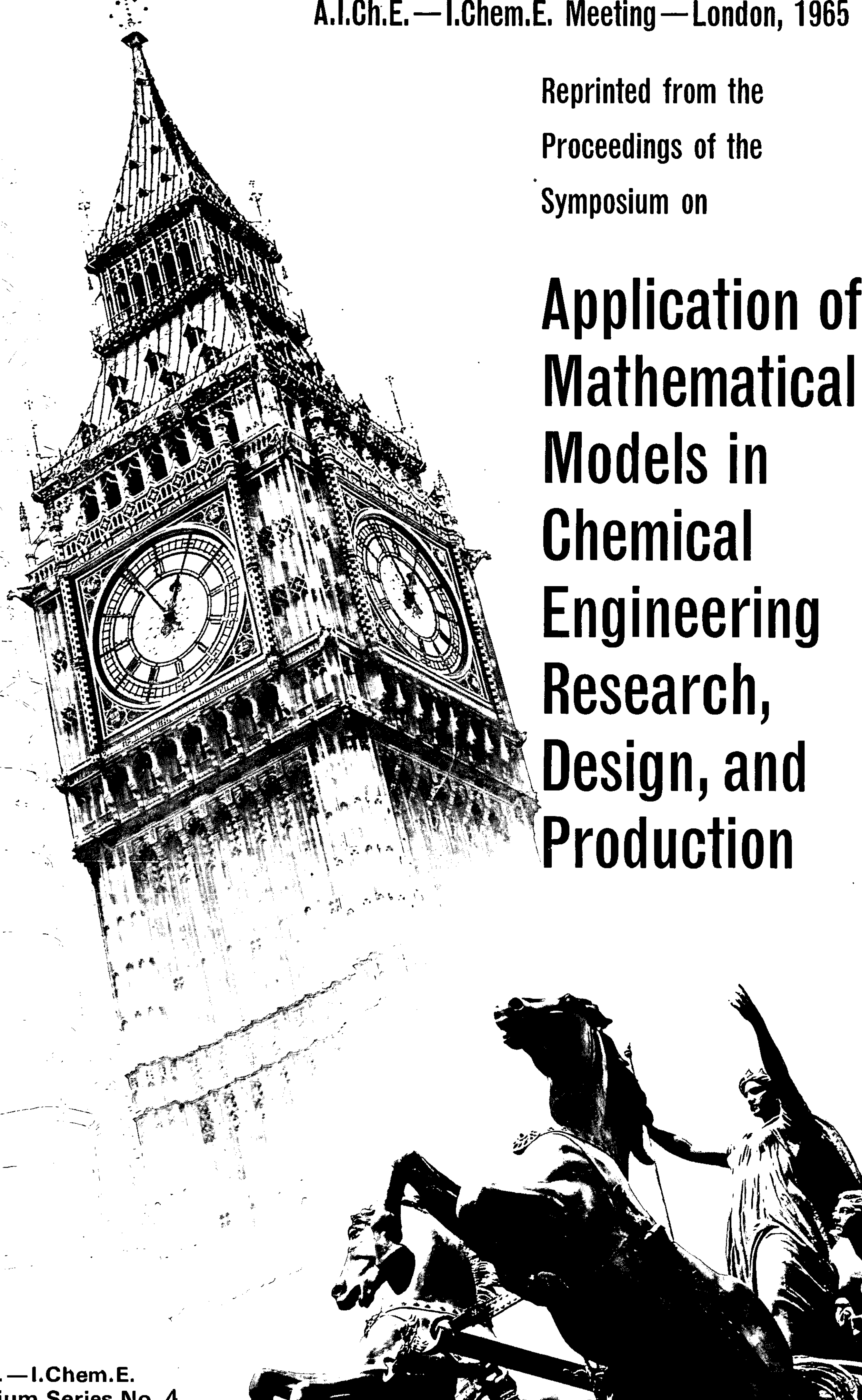
Dans cet article l'auteur montre comment on peut donner une formule quantitative précise de l'idée d'un régime transitoire optimum, et utilise le principe du maximum de Pontryagin pour déterminer ce régime optimum.

Zusammenfassung—Bei einigen technisch interessierenden exothermen Reaktionen wird ein regenerativer Wärmeaustausch zwischen den Reaktanden und den Produkten so durchgeführt, daß der Prozeß im stationären Zustand thermisch stabil ist. Jedoch benötigt ein solcher Prozeß im Anfahrzustand eine äußere Wärmezufuhr, die erst später abgeschaltet werden kann. Hier soll nun gezeigt werden, daß man die optimalen Anfahrbedingungen mit Hilfe des Pontryaginschen Maximum-Prinzips quantitativ formulieren kann.

A.I.Ch.E.—I.Chem.E. Meeting—London, 1965

Reprinted from the
Proceedings of the
Symposium on

Application of Mathematical Models in Chemical Engineering Research, Design, and Production



A.I.Ch.E.—I.Chem.E.
Symposium Series No. 4

1965 A.I.Ch.E.—I.Chem.E. Symposium Series

The Proceedings of the A.I.Ch.E—I.Chem.E. Joint Meeting held in London, 13–17 June 1965 are being published in the 1965 A.I.Ch.E–I.Chem.E. Symposium Series, each symposium forming a separate volume.

1 Advances in Separation Techniques

2 Chemical Engineering Under Extreme Conditions

3 Reaction Kinetics in Product and Process Design

4 Application of Mathematical Models in Chemical Engineering Research, Design, and Production

5 Materials Associated With Direct Energy Conversion

6 Transport Phenomena

7 Management Oriented Topics

8 Chemical Engineering Education (Published in *The Chemical Engineer,* September 1965)

9 New Chemical Engineering Problems in the Utilisation of Water

10 Mixing—Theory Related to Practice

Published by The Institution of Chemical Engineers, 16 Belgrave Square, London S.W.1
Price £3 ($9).

OPTIMUM TEMPERATURE GRADIENTS IN TUBULAR REACTORS WITH DECAYING CATALYST

By R. JACKSON, M.A.[*]

SYNOPSIS

The problem of determining the variation of temperature along the length of a tubular reactor so as to maximise the yield of a specified product is well known and complete solutions have been obtained in a number of cases. In practice, however, tubular reactors often contain a catalyst which decays with time. Since the decay is a result of a side reaction involving the catalyst, it does not occur at the same speed everywhere in the reactor and, in particular, the pattern of decay is dependent on the temperature policy adopted.

The present paper considers the problem of determining the optimum temperature policy, as a function of both time and position in the reactor, to maximise the total yield of a specified product in a given time interval. An optimising algorithm is derived, based on the concept of the " gradient in function space ", and the results of some preliminary computations are reported.

Introduction

There has recently been considerable interest in variational methods of solving optimisation problems in chemical plants. A variational treatment of plants consisting of sequences of discrete units was first given by Horn[1] and was later extended by the present writer[2,3] to deal with non-sequential systems involving recycle loops and other complex configurations. Horn[4] has also treated the continuous problem of determining optimum temperature gradients in tubular reactors, introducing the concept of the gradient in function space.

The work so far cited has all been concerned with optimisation problems in the steady state. Recently the present writer[5] has shown how variational methods can also be used to optimise the behaviour in time-varying situations such as those encountered at start up or in the presence of time-dependent perturbations. The method was developed specifically for the case in which the parameters available for adjustment are associated with separately distinguishable plant units, but there are also systems of interest in which the adjustable parameters are functions of a continuous position variable. In this paper we shall be concerned with one such problem, namely the determination of the best way to vary the temperature profile as a function of time in a tubular reactor with a decaying catalyst.

In general catalyst decay results from some side reaction involving the catalyst and consequently the instantaneous rate of decay depends on the temperature and composition of the reaction mixture and is not the same at all points in the reactor. It follows that the instantaneous rate of decay will depend on the choice of reactor temperature at each point, and the pattern of decay of the catalyst at any time will depend on the complete previous history of the temperature profile of the reactor. This leads to an interesting optimisation problem in which the current temperature profile influences the whole future course of the reaction by leaving its imprint on the pattern of catalyst decay. Mathematically, we are faced with the problem of optimisation with respect to a function of two continuous variables, namely the reactor temperature as a function of time and position. The object of the present work is to develop a variational method of attacking this problem.

The General Problem

We shall consider a catalyst-packed tubular reactor in which R independent chemical reactions take place. If we neglect axial diffusion, the composition of the reaction mixture at any point is determined by its composition on entering the reactor and the stoichiometric extents of reaction ζ_r $(r = 1, 2, \ldots, R)$.[†] The rate of each reaction at any point is a function of the local values of the composition and temperature of the reaction mixture and of the catalyst activity. We shall assume that the rate of decay of the catalyst is slow compared with the time required by the reactor to respond to changes in conditions so that the state of the reactor differs only very slightly from a steady state at all times. We shall also assume that we can neglect dynamical effects in the response of the reactor to the catalyst changes. The usual mass balance equations for the reactor then take the form:

$$\frac{\partial \zeta_r}{\partial x} = \mathrm{f}_r(\zeta_s, \theta, \phi) \quad (r = 1, 2, \ldots, R) \quad . \quad (1)$$

where x is the distance along the reactor from the entry, θ is the temperature, ϕ is the catalyst activity, and the form of the functions f_r is determined by the kinetic scheme of the reactions.

The catalyst acivity at any point decays at a rate determined by the local values of the temperature, the composition of the reaction mixture, and the catalyst activity itself: thus we can write:

$$\frac{\partial \phi}{\partial t} = \mathrm{g}(\zeta_s, \theta, \phi) \quad . \quad . \quad (2)$$

giving the rate of the side reaction responsible for the catalyst decay.

When the temperature $\theta(x, t)$ is specified in the domain of interest $0 \leqslant x \leqslant X$, $0 \leqslant t \leqslant T$, equations (1) and (2) can be solved subject to the boundary conditions:

$$\zeta_r = 0 \text{ when } x = 0 \text{ (all } 0 \leqslant t \leqslant T) \ (r = 1, 2, \ldots, R) \quad (3)$$

$$\text{and: } \phi = \phi_0 \text{ when } t = 0 \text{ (all } 0 \leqslant x \leqslant X) \quad (4)$$

where ϕ_0 is the uniform initial catalyst activity, X is the total length of the reactor and $0 \rightarrow T$ is the time interval of interest.

[*] University of Edinburgh and Heriot-Watt College Chemical Engineering Laboratories, Chambers Street, Edinburgh 1.

[†] *Symbols have the meanings given them on page* **4 : 38**.

The behaviour of the reactor is thus completely determined.

We now consider the problem of choosing $\theta(x, t)$ so as to maximise the total yield of a specified product during the time interval $0 \rightarrow T$. The concentration of any substance in the mixture leaving the reactor may be expressed as a linear combination of the extents of the separate reactions, so the objective function to be maximised takes the form:

$$P=\int_0^T \sum_{r=1}^{R} \alpha_r \zeta_r(X, t)\,\mathrm{d}t \quad . \qquad (5)$$

where the α_r are given constants.

The first step of any variational method is to express a small change δP in the dependent variable P in terms of the corresponding small change $\delta\theta(x, t)$ in the independently adjustable variable. A change in temperature from θ to $\theta+\delta\theta$ induces consequent changes $\delta\zeta_r$ in the extents of reaction and $\delta\phi$ in the catalyst activity, and these are related to first order by the incremental forms of equations (1) and (2), namely:

$$\frac{\partial}{\partial x}(\delta\zeta_r)=\sum_s \frac{\partial f_r}{\partial \zeta_s}\delta\zeta_s+\frac{\partial f_r}{\partial \phi}\delta\phi+\frac{\partial f_r}{\partial \theta}\delta\theta \quad . \qquad (6)$$

and:

$$\frac{\partial}{\partial t}(\delta\phi)=\frac{\partial g}{\partial \phi}\delta\phi+\sum_s \frac{\partial g}{\partial \zeta_s}\delta\zeta_s+\frac{\partial g}{\partial \theta}\delta\theta \quad . \qquad (7)$$

with the boundary conditions:

$$\delta\zeta_r = 0 \text{ when } x = 0 \text{ (all } 0\leqslant t\leqslant T) \; (r = 1, 2, \ldots, R) \; . \qquad (8)$$

and:

$$\delta\phi = 0 \text{ when } t = 0 \text{ (all } 0\leqslant x\leqslant X) \; . \qquad (9)$$

We now introduce a new set of variables λ_r adjoint to the $\delta\zeta_r$ and μ adjoint to $\delta\phi$. These are defined by the differential equations they satisfy and the associated boundary conditions, namely:

$$\frac{\partial \lambda_r}{\partial x}=-\sum_s \frac{\partial f_s}{\partial \zeta_r}\lambda_s-\frac{\partial g}{\partial \zeta_r}\mu \quad . \qquad (10)$$

and:

$$\frac{\partial \mu}{\partial t}=-\frac{\partial g}{\partial \phi}\mu-\sum_s \frac{\partial f_s}{\partial \phi}\lambda_s \quad . \qquad (11)$$

with boundary conditions:

$$\lambda_r = \alpha_r \text{ when } x = X \text{ (all } 0\leqslant t\leqslant T) \; (r = 1, 2, \ldots, R) \qquad (12)$$

and:

$$\mu = 0 \text{ when } t = T \text{ (all } 0\leqslant x\leqslant X) \quad . \qquad (13)$$

Now consider:

$$\frac{\partial}{\partial x}\Big(\sum_r \lambda_r\delta\zeta_r\Big)=\sum_r \lambda_r\frac{\partial}{\partial x}(\delta\zeta_r)+\sum_r \delta\zeta_r\frac{\partial \lambda_r}{\partial x} \quad . \qquad (14)$$

Substituting from the incremental equations (6) and the adjoint equations (10) into the right hand side of equation (14) gives, after simplification:

$$\frac{\partial}{\partial x}\Big(\sum_r \lambda_r\delta\zeta_r\Big)=\sum_r \lambda_r\frac{\partial f_r}{\partial \phi}\delta\phi-\sum_r \mu\frac{\partial g}{\partial \zeta_r}\delta\zeta_r+\sum_r \lambda_r\frac{\partial f_r}{\partial \theta}\delta\theta \qquad (15)$$

Consider also:

$$\frac{\partial}{\partial t}(\mu\delta\phi)=\mu\frac{\partial}{\partial t}(\delta\phi)+\delta\phi\frac{\partial \mu}{\partial t} \quad . \qquad (16)$$

Substituting from the incremental equation (7) and the adjoint equation (11) into the right hand side of equation (16) gives:

$$\frac{\partial}{\partial t}(\mu\delta\phi) = -\sum_r \delta\phi\frac{\partial f_r}{\partial \phi}\lambda_r+\sum_r \mu\frac{\partial g}{\partial \zeta_r}\delta\zeta_r+\mu\frac{\partial g}{\partial \theta}\delta\theta \quad . \qquad (17)$$

Adding equations (15) and (17) and simplifying then gives:

$$\frac{\partial}{\partial t}(\mu\delta\phi)+\frac{\partial}{\partial x}\Big(\sum_r \lambda_r\delta\zeta_r\Big)=\mu\frac{\partial g}{\partial \theta}\delta\theta+\sum_r \lambda_r\frac{\partial f_r}{\partial \theta}\delta\theta \quad . \qquad (18)$$

We now integrate both sides of equation (18) over the rectangular domain $0\leqslant x\leqslant X$, $0\leqslant t\leqslant T$. Considering the two terms on the left hand side separately, the first gives:

$$\int_0^X\int_0^T \frac{\partial}{\partial t}(\mu\,\delta\phi)\,\mathrm{d}x\,\mathrm{d}t=\int_0^X \big|\mu\,\delta\phi\big|_0^T\,\mathrm{d}x=0 \quad . \qquad (19)$$

making use of the boundary conditions (9) and (13) on $\delta\phi$ and μ, while the second term gives:

$$\int_0^X\int_0^T \frac{\partial}{\partial x}\Big(\sum_r \lambda_r\delta\zeta_r\Big)\,\mathrm{d}x\,\mathrm{d}t=\int_0^T \Big|\sum_r \lambda_r\delta\zeta_r\Big|_0^X\,\mathrm{d}t$$

$$=\int_0^T \sum_r \alpha_r\delta\zeta_r(X, t)\,\mathrm{d}t \qquad (20)$$

making use of the boundary conditions (8) and (12) on $\delta\zeta_r$ and λ_r. The result of integrating equation (18) is therefore:

$$\int_0^T \sum_r \alpha_r\delta\zeta_r(X, t)\,\mathrm{d}t=\int_0^X\int_0^T\Big\{\mu\frac{\partial g}{\partial \theta}+\sum_r \lambda_r\frac{\partial f_r}{\partial \theta}\Big\}\delta\theta\,\mathrm{d}x\,\mathrm{d}t.$$

But reference to equation (5) shows that the left hand side of this is simply the variation δP in the objective function, so we may write:

$$\delta P=\int_0^X\int_0^T\Big\{\mu\frac{\partial g}{\partial \theta}+\sum_r \lambda_r\frac{\partial f_r}{\partial \theta}\Big\}\delta\theta\,\mathrm{d}x\,\mathrm{d}t \quad . \qquad (21)$$

and we have achieved our objective of expressing δP in terms of the small variation $\delta\theta(x, t)$ in the temperature policy.

Probably the best way of using this result is to regard:

$$P_\theta=\mu\frac{\partial g}{\partial \theta}+\sum_r \lambda_r\frac{\partial f_r}{\partial \theta} \quad . \qquad (22)$$

as the gradient of P in the function space of the function $\theta(x, t)$. The idea of a gradient in function space was originally suggested by Courant[6] and was introduced into chemical engineering by Horn.[4] Let $\theta_0(x, t)$ be an initial guess at the temperature policy. We consider small variations $\delta\theta(x, t)$ from θ_0 such that the integral:

$$\int_0^X\int_0^T (\delta\theta)^2\,\mathrm{d}x\,\mathrm{d}t$$

takes the same value in all cases. Then, by an obvious analogy with spaces of finite dimensionality, we say that all these variations are of the same magnitude. It can then be shown, (Leitman,[7] Chapter 6), that the largest increase in P results from the member of this set of variations whose value is proportional to P_θ for all x and t; in other words:

$$\delta\theta(x, t)=l\Big\{\mu\frac{\partial g}{\partial \theta}+\sum_r \lambda_r\frac{\partial f_r}{\partial \theta}\Big\} \quad . \qquad (23)$$

where l is a constant and may be regarded as a displacement of θ along the direction of steepest ascent through the point θ_0 in the θ–space, and correspondingly P_θ may be referred to as the gradient of P in the function space of θ.

Equation (23) provides the basis of a computational procedure to maximise P. Starting with an initial guess $\theta_0(x, t)$ at the temperature policy, the direction of steepest ascent in the function space of θ may be determined from equation (22). θ_0 may then be modified by increments given by equation (23) with successively increasing values of l, thus moving up the steepest ascent line through θ_0 and computing the value of P at each stage. At some suitably determined point on this line the direction of steepest ascent can be redetermined from equation (22) and the ascent continued along the new steepest ascent line, continuing this procedure until P no longer changes significantly. In spaces of finite dimensionality many ways of using the gradient more efficiently than the simple steepest ascent procedure have been described[7,8,9] and there is no difficulty in principle in generalising these to apply to the function spaces encountered in the present type of problem. The most effective procedure will clearly have to be found by trial in each particular case.

A First-order, Reversible, Exothermic Reaction with Temperature-dependent Catalyst Decay

As an example with which to develop a practical computational procedure we shall consider a single, first-order, reversible, exothermic reaction. If y_0 is the mole fraction of the reaction product in the feed and ζ is the stoichiometric extent of reaction, the single equation corresponding to equations (1) of the general case is:

$$\frac{\partial \zeta}{\partial x}=\phi\,.\,\tau k_{20}\exp\,(-A_2/\theta)$$
$$\times\,[(1-y_0-\zeta)\,K_0\exp\,(Q/\theta)-(y_0+\zeta)]\equiv \mathrm{f}(\zeta,\theta,\phi) \tag{24}$$

Here $K_0 \exp (Q/\theta)$ is the equilibrium constant, with Q equal to the ratio of the heat of reaction and the gas constant, A_2 is the ratio of the activation energy of the reverse reaction to the gas constant, k_{20} is the frequency factor for the velocity constant of the reverse reaction, and τ is the total residence time in the reactor. The independent variable x represents distance along the reactor expressed as a fraction of the total length X, and this choice of variable reduces the whole equation to a conveniently dimensionless form, since the group τk_{20} is dimensionless. The decay of the catalyst activity ϕ will be assumed to be influenced only by the temperature, and will be described by an equation of the form:

$$\frac{\partial \phi}{\partial t}=-\frac{\theta}{\theta_c}\phi \qquad . \qquad . \quad (25)$$

The time t is conveniently expressed in dimensionless form as a fraction of the interval T of interest, and it then follows that the constant θ_c has the dimensions of temperature. It may be regarded as a characteristic temperature which determines the extent of decay in the time interval considered.

Some simplification can be obtained by a change of variable in equation (25). If ϕ is replaced by $\psi = \log_e \phi$, equation (25) becomes:

$$\frac{\partial \psi}{\partial t}=-\frac{\theta}{\theta_c}\equiv \mathrm{g}(\theta) \qquad . \qquad . \quad (26)$$

and the boundary condition $\phi = 1$ at $t = 0$ is replaced by $\psi = 0$ at $t = 0$. With the catalyst decay equation in the form (26) its right hand side is independent of ψ and the general adjoint equation (11) reduces to:

$$\frac{\partial \mu}{\partial t}=-\frac{\partial \mathrm{f}}{\partial \psi}\lambda \qquad . \qquad . \quad (27)$$

(There is only one variable λ corresponding to the single reaction variable ζ, so the suffix $_r$ used to distinguish reactions in the general case can be omitted.) Taking account of the boundary condition (13) on μ, equation (27) integrates to give:

$$\mu(t)=\int_t^1 \frac{\partial \mathrm{f}(u)}{\partial \psi}\lambda(u)\,\mathrm{d}u$$

where the upper limit is unity in view of the way t is defined. Using this result the general expression (22) for the gradient in function space reduces to:

$$P_\theta=\lambda\frac{\partial \mathrm{f}}{\partial \theta}+\frac{\mathrm{d}g}{\mathrm{d}\theta}\int_t^1\frac{\partial \mathrm{f}(u)}{\partial \psi}\lambda(u)\,\mathrm{d}u \qquad . \quad (28)$$

This expresses P_θ in terms of the solution λ of a single adjoint equation, namely:

$$\frac{\partial \lambda}{\partial x}=-\lambda\frac{\partial \mathrm{f}}{\partial \zeta} \qquad . \qquad . \qquad . \quad (29)$$

with the boundary condition:

$$\lambda = 1 \text{ when } x = 1 \text{ (all } 0\leqslant t\leqslant 1) \tag{30}$$

Equations (29) and (30) are the appropriate special forms of the general equations (10) and (12).

Having established the form of the relevant equations, the problem of maximising P resolves itself into two parts: firstly we must obtain a numerical method of computing the gradient P_θ, and secondly we must devise a numerical scheme which makes use of the computed values of the gradient to maximise P. These will be discussed in turn.

Given an estimate $\theta_0(x, t)$ of the temperature policy, the following four steps are involved in the computation of the corresponding gradient P_θ.

(i) With the assumed value θ_0 for θ, equation (26) is integrated forward in time at each value of x from the initial condition $\psi = 0$ at $t = 0$, thus generating the function $\psi\,(x, t)$.

(ii) Making use of the above result to determine ϕ, equation (24) is integrated forward in x at each value of t from the initial condition $\zeta = 0$ at $x = 0$, thus generating the function $\zeta\,(x, t)$.

(iii) Knowing ϕ and ζ, $\partial \mathrm{f}/\partial \zeta$ can be evaluated at any point and the adjoint equation (29) can be integrated backwards in x at each value of t from the terminal condition $\lambda = 1$ at $x = 1$, thus generating the function $\lambda(x, t)$.

(iv) Differentiation of the function f gives explicit expressions for $\partial \mathrm{f}/\partial \theta$ and $\partial \mathrm{f}/\partial \psi$ and these can be evaluated at any point using the values of ψ, ζ and λ found in steps (i), (ii) and (iii) above. The integral on the right hand side of equation (28) may then be evaluated numerically, and hence P_θ may be computed for all x and t.

The integration of differential equations required by steps (i), (ii), and (iii) and the evaluation of the definite integrals in step (iv) all have to be carried out by finite difference methods.

The demands on storage and computing time of a digital computer will depend on the number of steps into which the basic intervals in the x and t directions are divided, but

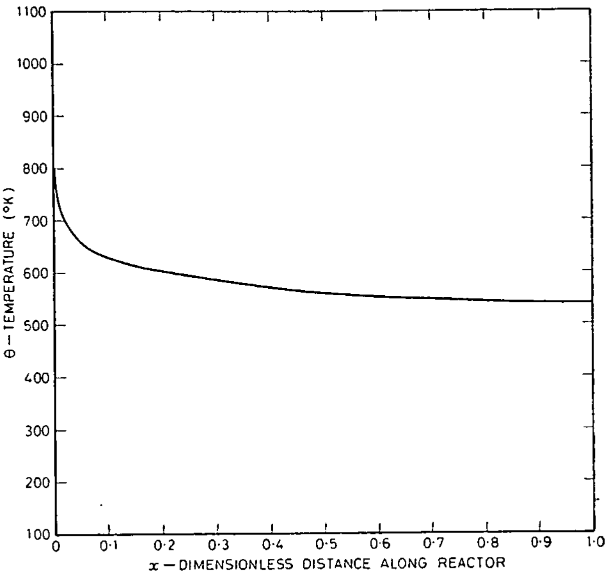


Fig. 1.—*Optimum temperature profile in the absence of catalyst decay*

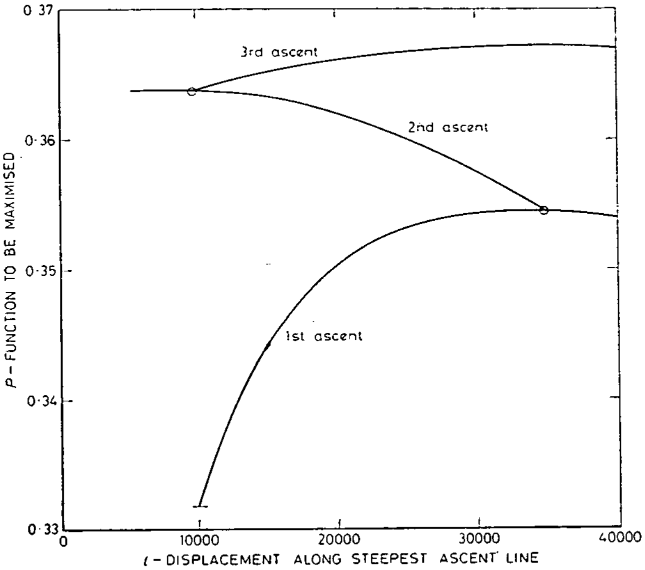


Fig. 2.—*Variation of P on successive ascents*

clearly these demands will both be heavy if reasonable accuracy is to be attained. The exploratory calculations described in this paper used an Atlas digital computer, which is both very large and very fast. Nevertheless it would have been inappropriate to make excessive demands in computing P_θ until an efficient procedure for using P_θ in the maximisation of P had been evolved. In the absence of any catalyst decay the optimum temperature profile in the reactor can be obtained by well-known methods and is known to be a curve which decreases with increasing x and is concave upwards. This suggests that the finite difference intervals should be short at small values of x but may be longer further along the reactor. In the light of this consideration it was decided to divide the interval $0 \leqslant x \leqslant 0{\cdot}1$ into 20 equal sub-intervals, and the interval $0{\cdot}1 \leqslant x \leqslant 1{\cdot}0$ into 18 equal sub-intervals. In the t-direction, on the other hand, the interval $0 \leqslant t \leqslant 1{\cdot}0$ was divided into 25 equal sub-intervals. In all, therefore, each function of two variables such as $\zeta(x, t)$ had to be computed and stored at 1014 points. The numerical integrations made use of a process of iterative adjustment at each step of the

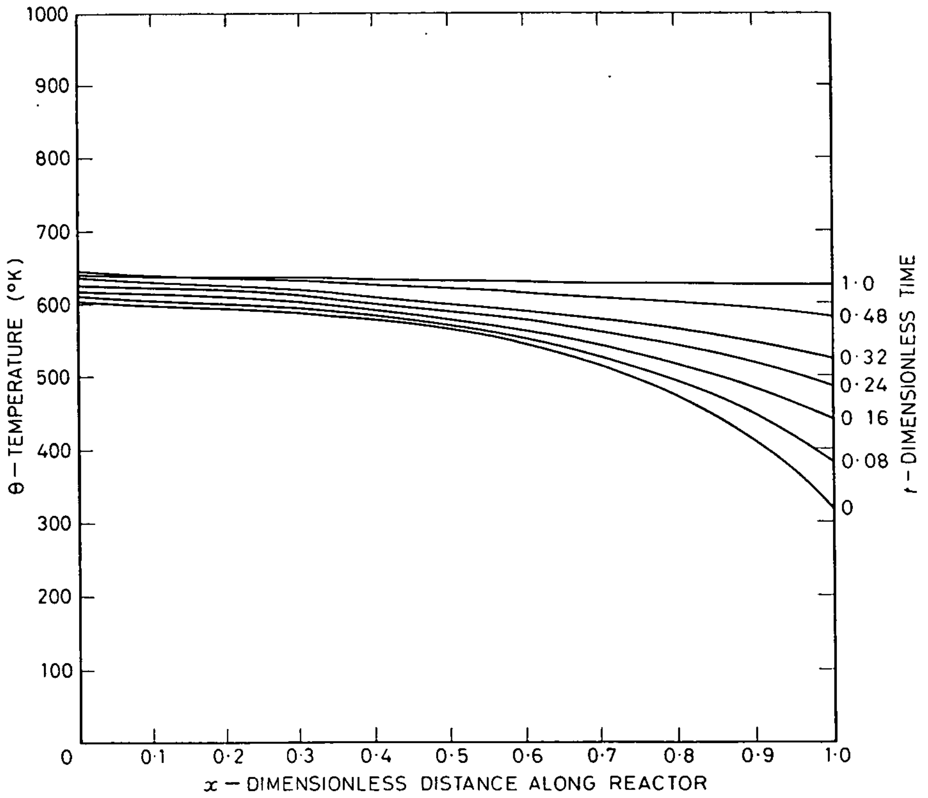


Fig. 3.—*Temperature profiles at the end of the first ascent*

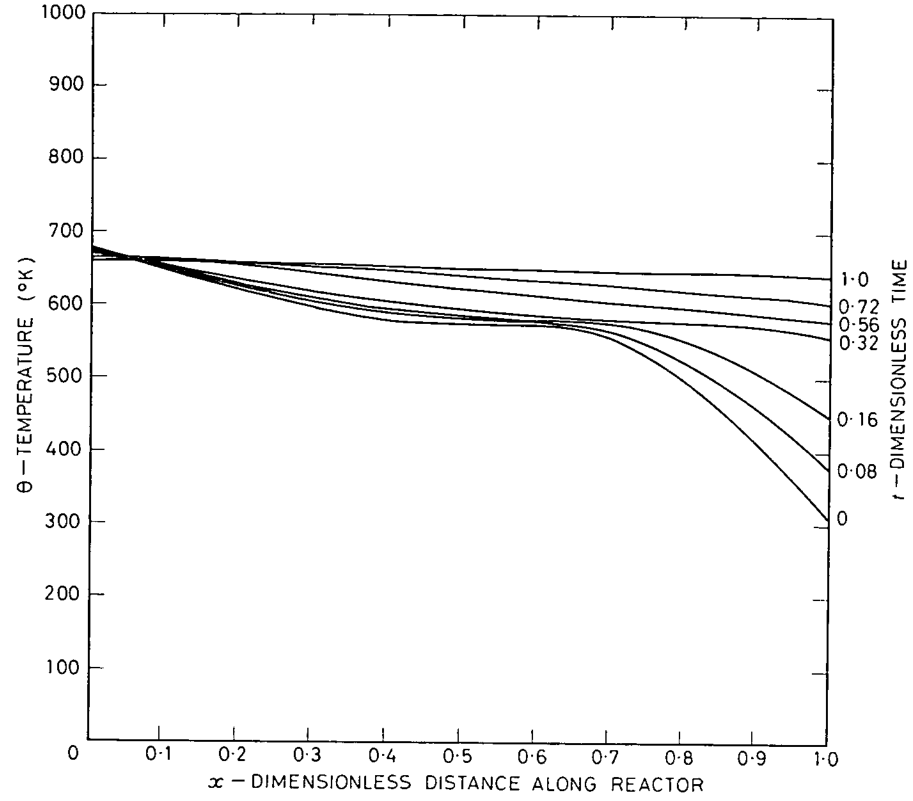


Fig. 4.—*Temperature profiles at the end of the second ascent*

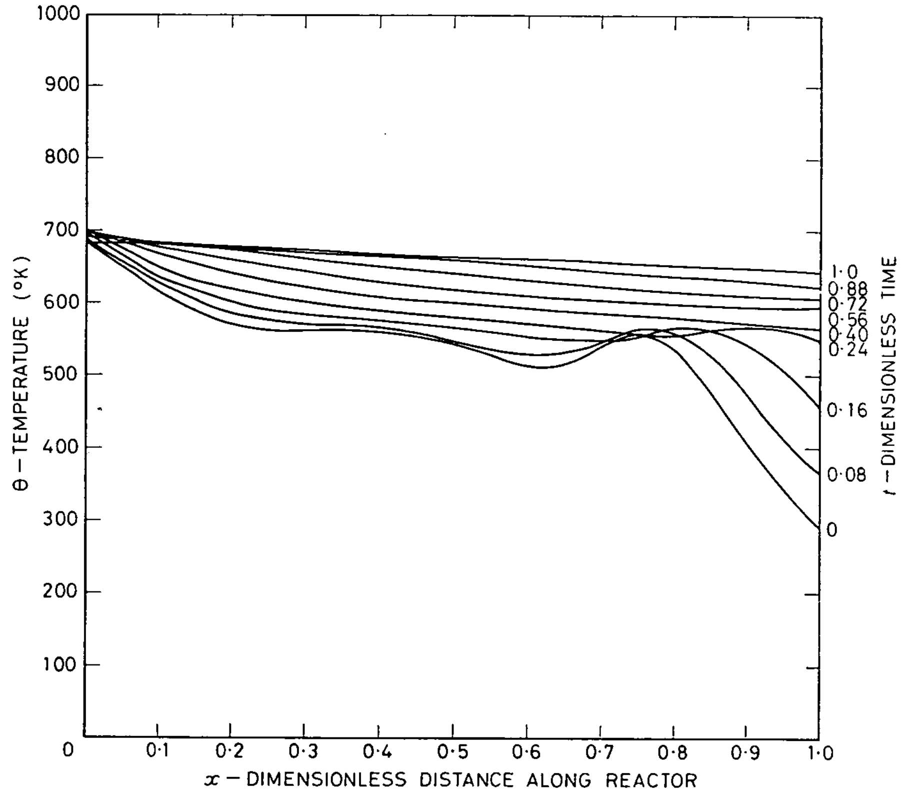


Fig. 5.—*Temperature profiles at the end of the third ascent*

forward integration so that the slope of a chord joining the values of the solution at the ends of one sub-interval was equal to the gradient, as calculated from the right hand side of the differential equation, at the mid point of the sub-interval.

The computed values of P_θ can be used in schemes of varying sophistication to maximise P. In this initial exploration it was decided to adopt a simple stepwise steepest-ascent procedure in θ-space, and to examine the results at the end of each successive ascent. From an initial estimate $\theta_0(x, t)$ of the temperature policy, a sequence of improved policies $\theta(x, t)$ were obtained from the formula:

$$\theta(x, t) = \theta_0(x, t)+l P_\theta(x, t)$$

using successively increasing values of l. For each value of l the corresponding value of the objective function P was computed and the value $l = l_m$ which maximised P was located approximately by interpolation. The new temperature policy, $\theta = \theta_0+l_m P_\theta$, was then taken as the starting point of another similar ascent, and so on for successive ascents. It is known that this is not a very efficient maximisation method; in particular, in spaces of finite dimensionality, progress becomes very slow once the current point has ascended to the neighbourhood of a ridge. However, it is a perfectly adequate method to use until it is certain that the computation of P_θ itself is in order, and it has the advantage that progress is divided into more or less separate and independent stages, namely the successive ascents. Thus it is possible to print out the temperature policy for inspection at the end of each ascent before reading it back into the computer as the starting point of the next ascent, and no other information need be carried forward from ascent to ascent.

Results and Discussion

For the preliminary calculations, the following values were taken for the constants appearing in equations (24) and (25):

$$\tau k_{20} = 3\times 10^7$$
$$y_0 = 0{\cdot}06$$
$$A_2 = 10\,000°\mathrm{K}$$
$$K_0 = 0{\cdot}000\,23$$
$$Q = 5000°\mathrm{K}$$
$$\theta_c = 250°\mathrm{K}$$

and, as an initial guess at the temperature policy, θ was assumed to take the value 600°K for all x and t. Fig. 1 shows the optimum temperature profile in the absence of any catalyst decay. This is obtained, as is well known, by choosing θ to maximise the rate of reaction at each point, and with the above values of the constants this problem can actually be solved in closed form.

The stepwise steepest-ascent procedure described in the previous section was then pursued through three successive ascents from the initial temperature profile. The variation of P with displacement, l, along the steepest ascent lines is shown in Fig. 2, and the behaviour is seen to be as expected. P increases with l initially, then passes through a maximum whose location determines the starting point of the next ascent. Figs 3, 4, and 5 show the temperature policies at the end of the first, second, and third ascents respectively, plotting θ against x at various values of t. The complexity of the problem is such that it is very difficult to make an advance guess at the form of the solution on physical grounds, but the general features of Figs 3, 4, and 5 are not unreasonable. At small values of t, the improved profiles of θ against x are characterised by an initial fairly-rapid fall in temperature, followed by a section in which the temperature falls relatively slowly, and terminated by a second region of rapidly falling temperature. Such a profile has the effect of preserving the catalyst near the end of the reactor in a state of high activity. At later times, the temperature is increased generally and the profile flattens out markedly thus raising the temperature near the end of the reactor substantially and making use of the relatively fresh catalyst there. More difficult to explain are the curious waves appearing in the profiles of θ against x shown in Fig. 5 at small values of t. It seems unlikely that these are genuine features of the optimum policy and it is not yet known whether they arise from truncation errors in integrating the differential equations or whether they are an inherent feature of the stepwise steepest-ascent procedure in function space.

It seems clear from the results given that the method of determining P_θ is sound in principle, though some improvement in the finite difference approximations will probably be necessary to reduce truncation errors. It also seems likely that the simple stepwise steepest ascent procedure is not a very suitable method for determining the optimum policy accurately, since the first three ascents show no very clear evidence of convergence. Further work will be directed to developing an improved numerical procedure to evaluate P_θ and to replacing the stepwise steepest-ascent procedure by one of the more sophisticated methods of using the information contained in P_θ.[8, 9]

Finally, it must be pointed out that the procedure described in this paper tackles only one aspect of the problem of a decaying catalyst. It has been assumed throughout that the total time, T, is given, and the method permits the total yield of a given product or, more generally, the total operating profits, to be maximised. In practice one is also faced with the problem of determining the value of T at which the catalyst should be replaced in order to give maximum profit over a long period, taking account of the cost of catalyst replacement. It is possible to give a variational equation to determine the optimum value of T, but in practice this turns out to have no advantage over the straightforward procedure of solving the problem stated in the present paper for various values of T and determining the value which maximises the difference between the operating profits and the costs of catalyst replacement.

Symbols Used

A_2 = ratio of the activation energy of the reverse reaction to the universal gas constant.

f_r = functions on the right hand sides of equation (1).

g = function on the right hand side of equation (2).

K_0 = temperature independent factor in the equilibrium constant.

k_{20} = frequency factor in the velocity constant of the reverse reaction.

l = displacement along a steepest ascent line.

l_m = value of l which maximises P.

P = objective function to be maximised.

P_θ = gradient of P in the space of the function $\theta(x, t)$.

Q = ratio of the heat of reaction to the gas constant.

R = total number of independent reactions.

T = length of the time interval of interest.

t = dimensionless measure of time.

X = total length of the reactor.

x = dimensionless measure of distance along the reactor.

y_0 = mole fraction of reaction product present in the feed mixture.

α_r = constants defining the objective function.

ζ = stoichiometric extent of reaction for single reversible reaction.

ζ_r = stoichiometric extents of reaction for independent reactions.

ζ_s = steady-state value of ζ_r.

θ = absolute temperature.

θ_c = characteristic temperature determining the rate of catalyst decay.

θ_0 = initial approximate temperature policy.

λ = variable adjoint to ζ.

λ_r = variables adjoint to the ζ_r.

μ = variable adjoint to ϕ.

τ = total residence time in the reactor.

ϕ = catalyst activity.

ϕ_0 = initial catalyst activity.

ψ = related by change of variable to the catalyst activity.

The above quantities may be expressed in any set of consistent units in which force and mass are not defined independently.

References

[1] Horn, F. *Chemical Engineering Science*, 1961, **15**, 176.

[2] Jackson, R. *Chemical Engineering Science*, 1964, **19**, 9.

[3] Jackson, R. *Chemical Engineering Science*, 1964, **19**, 253.

[4] Horn, F. and Troltenier, U. *Chem.-Ing.-Tech.*, 1960, **32**, 382.

[5] Jackson, R. *Chemical Engineering Science*. To be published.

[6] Courant, R. and Hilbert, D. " *Methods of Mathematical Physics* ", 1953, Vol. I (New York: Interscience Publications).

[7] Leitman, G. (ed.). " *Optimisation Techniques* ", 1962 (New York: Academic Press Inc.).

[8] Fletcher, R. and Powell, M. J. D. *Computer Journal*, 1963, **6**, 163.

[9] Fletcher, R. and Reeves, C. M. *Computer Journal*, 1964, **7**, 149.

The manuscript of this paper was received on 27 *October*, 1964.

DISCUSSION OF PAPERS PRESENTED AT THE FIRST SESSION

Prof. Donald L. Katz said that during some thirty-five years in the general field of chemical engineering he had seen a gradual transition from the physical system to mathematical models. It was indeed wonderful that there had been such a transition to the field of mathematics, and that a great deal was being learned. However, it might be as well as the discussions proceeded on the topics to relate what was currently being done back to the physical situation which was being described.

Prof. A. B. Metzner said that Bird had produced an excellent survey of his own activities and those of his students, but attention should also be drawn to the analyses of engineering problems of several of Bird's contemporaries, particularly Drs. Pearson and Walters in the United Kingdom, Dr. Giesekus in Germany, Dr. Astarita in Italy, and of several currently active Americans. Together they appeared to give a rather clear insight into the kinds of mathematical models which must be used on various occasions. Some lead to considerations as complex as those which Bird had discussed, but others were very simple and, as they might suffice for the analysis of certain classes of problems, could not be neglected. The recent extensive studies of the Coleman-Noll theory of " Simple Fluids " fell into the latter category.

With regard to the flow patterns in non-Newtonian fluids, perhaps the most graphical illustration was the work of Walters and Giesekus, who have pointed out that visco-elastic fluids could be mixed by pumping material through an open pipe of non-circular cross-section, or, equivalently, through a helical coil of round tube. The secondary flows which arose as a consequence of visco-elastic properties caused intense cellular mixing patterns to develop in both cases.

Prof. Bird said that the principal new feature of the work presented in the present paper was the use of the Rouse molecular theory to reduce drastically the number of constants in a rheological model and thereby obtain a model containing a small number of constants. Such a procedure was believed to be new and had not been used by the researchers cited by Prof. Metzner.

The works of Coleman and Noll have been appropriately cited in Refs 36–38 of the paper. One comment needs to be made relative to their " Second order fluid ", however, which was inadequate for describing experimental results. In Fig. 5 of the paper the Coleman–Noll second-order fluid described only the region less than $\lambda\omega$ or $\lambda\gamma$ about 0·1 (*i.e.*, where η and ζ were both constants) and failed to describe the " power law region " above $\lambda\omega$ or $\lambda\gamma$ about 1, which was generally the region of engineering interest.

Prof. Bird also emphasised the importance of comparing the time-constant of the fluid with a time-constant for the flow system, and reference was made to seven experiments which had been performed at the University of Wisconsin. That comparison of time-constants often provided a simple means for determining when viscoelastic effects were important (see Ref. 46 of the paper).

[*Note Added* 4 *November* 1965: At the time of preparing the manuscript for this paper, the author was unaware of an interesting model proposed by White and Metzner.[1] This model is a non-linear Maxwell model containing one adjustable function (the non-Newtonian viscosity) and one constant. This model may prove particularly useful for those engineering problems in which it is important to describe non-Newtonian viscosity accurately, but unimportant to have an accurate portrayal of oscillatory response, stress relaxation, and secondary normal stress.]

Dr. J. Wei referred to the paper by Kelsall and Reid and said that it was gratifying to learn that a subject as difficult as grinding was amenable to exact analysis. The problem was very similar to that in the analysis of a complex system of first-order chemical reactions: one would thus expect that the techniques developed in one field might be beneficial to the other.

Because the effect of residence time distribution was absent, a batch (rather than a continuous) experiment might be of value to the investigation of Kelsall and Reid. For example, the drop in breakage rate at particle size 2362 microns was probably an effect of residence time, and not a true effect.

Mr. P. J. Hoftyzer said that during the symposium it appeared that the use of mathematical models in chemical engineering had already developed into a number of specialised fields. They corresponded to a number of different purposes, for which a mathematical model was constructed, for instance:

(1) fundamental process studies,

(2) design of apparatus,

(3) process control, and

(4) plant optimisation.

The complexity of the mathematical model for a given process decreased markedly in the above-mentioned order. At the same time, an increasing number of other factors had to be incorporated into the calculating programme in which the model was used.

The first four papers of the session of the symposium were examples of the use of a mathematical model for the first-mentioned purpose. They dealt with quite different fields of chemical engineering—rheology, comminution, packed columns, thermodynamics. But all of them resulted from physical–chemical process studies, and contained a number of constants, to be determined experimentally as a function of several process parameters. Yet in the derivation of the models a number of simplifying assumptions had been made.

So mathematical models of that type showed a tendency towards increasing complexity. The possibilities of calculations with complicated models had grown considerably by the development of rapid computers. It should be stressed, however, that those possibilities were certainly not unlimited. That would probably often result in a limit for the complexity of the model beyond which it becomes inefficient.

1965 A.I.Ch.E.—I.Chem.E. Symposium Series

The Proceedings of the A.I.Ch.E–I.Chem.E. Joint Meeting held in London, 13–17 June 1965 are being published in the 1965 A.I.Ch.E–I.Chem.E. Symposium Series, each symposium forming a separate volume.

1 Advances in Separation Techniques

2 Chemical Engineering Under Extreme Conditions

3 Reaction Kinetics in Product and Process Design

4 Application of Mathematical Models in Chemical Engineering Research, Design, and Production

5 Materials Associated With Direct Energy Conversion

6 Transport Phenomena

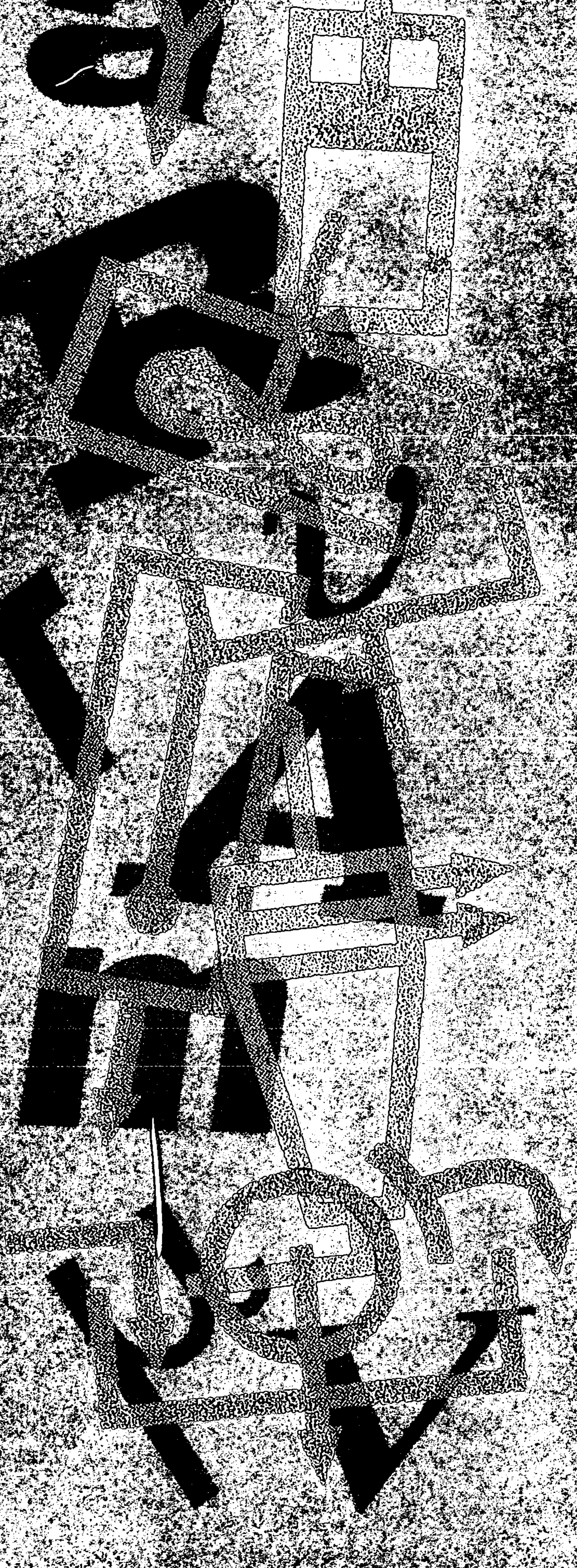
7 Management Oriented Topics

8 Chemical Engineering Education (Published in *The Chemical Engineer*, September 1965)

9 New Chemical Engineering Problems in the Utilisation of Water

10 Mixing—Theory Related to Practice

Published by The Institution of Chemical Engineers, 16 Belgrave Square, London S.W.1
Price £3 ($9).



Reprinted from:

THE TRANSACTIONS OF THE INSTITUTION OF CHEMICAL ENGINEERS

Vol. 45, No. 4
May 1967

THE CHEMICAL ENGINEER & THE TRANSACTIONS

Together, these two journals are the most authoritative source for all that is new and progressive in the field of chemical engineering. Because of the lasting value of these publications a special binding service is available to all subscribers.

These official journals of The Institution of Chemical Engineers are entirely written and edited for chemical engineers. Their aim is to keep the profession informed of the latest developments and technology in the chemical and allied industries.

Papers and articles are devoted to any of the aspects of chemical engineering. They contain accounts of original research or of the application of the results of research in the design and construction of plant. They deal with the development and operation of new processes, or with improvement and modifications to existing processes, in so far as the treatment relates to the underlying principles and their quantitative application, and is not merely descriptive. Knowledge which has been published elsewhere is not published in the Institution's journals unless it forms part of the collective treatment of a broad field or is used to illustrate the applications of a principle or as a basis of the discussion of further advances.

The main purpose of *The Transactions* is to publish papers which develop the science of chemical engineering in all its aspects. The matter provides quantitative information leading to the establishment of principles. Should the matter be descriptive rather than quantitative, then the value of any paper published in *The Transactions* lies in the critical nature of the treatment and the perspective achieved.

The main purpose of *The Chemical Engineer* is to provide information of current interest to chemical engineers. Consequently, the matter is either directly applicable to the problems with which such engineers are faced or of interest to them by virtue of its reflection on their own problems.

Joint Subscription £7 per volume ($20)
(monthly except January and August)

A special service to subscribers

If you are a subscriber to *The Chemical Engineer* and *The Transactions* you can receive in addition complete bound volumes of both journals at the end of each year These volumes of 10 issues are full-bound in cloth on board covers gold blocked on the spine. The two bound volumes will be sent to you in addition to your subscription copies. This special service will only cost an additional £1 ($3) for each bound volume of *The Chemical Engineer* or *The Transactions* that you require.

The Transactions (Bound Volume) £1 ($3.00) per volume

The Chemical Engineer (Bound Volume) £1 ($3.00) per volume

(only available to subscribers)

AN APPROACH TO THE NUMERICAL SOLUTION OF TIME-DEPENDENT OPTIMISATION PROBLEMS IN TWO-PHASE CONTACTING DEVICES

By R. JACKSON, M.A. (ASSOCIATE MEMBER)*

SYNOPSIS

In distillation, gas absorption, and liquid extraction applications, and in tubular catalytic reactors, two phases in relative motion interact with each other by the transfer of matter and heat. In the approximation of no axial diffusion within each phase, the time-dependent behaviour of all these systems is governed by similar sets of first-order partial differential equations which provide constraining conditions for problems of optimum start-up and control. This paper gives a common mathematical formulation of all such optimisation problems and examines the practicability of solving them numerically with reference to a particular problem, namely that of the optimum temperature policy in a tubular reactor with decaying catalyst.

Introduction

Many chemical engineering operations involve thermal interaction, transfer of materials, and other mutual influences of two (or more) flowing streams of fluid. One might instance gas absorption apparatus which makes use of mass transfer between liquid and gas streams (often flowing in opposite directions), packed distillation columns in which mass transfer is accompanied by a substantial heat transfer, and latent heat effects are involved, and to introduce a further complication chemical reactors in which reaction on a stationary catalyst phase is accompanied by mass and heat transfer between the catalyst and a flowing reactant stream. Within each stream mixing in the direction of flow always takes place to a greater or less extent but it is frequently a reasonable approximation to neglect this altogether and to make the assumption of ideal plug flow in each separate stream. In this approximation all the above types of system may be represented mathematically by equations of similar form, namely:†

$$\gamma\frac{\partial \mathbf{u}}{\partial t} + \alpha\frac{\partial \mathbf{u}}{\partial x} = \mathbf{f}(\mathbf{u}, \mathbf{v}, \boldsymbol{\theta}) \qquad . \qquad . \qquad (1)$$

$$\epsilon\frac{\partial \mathbf{v}}{\partial t} + \beta\frac{\partial \mathbf{v}}{\partial x} = \mathbf{g}(\mathbf{u}, \mathbf{v}, \boldsymbol{\theta}) \qquad . \qquad . \qquad (2)$$

where t represents time, x distance along the common axis of flow, **u** and **v** are sets of variables, or state column vectors, representing physical quantities such as the composition of the two streams, and **θ** represents a column vector of functions of x whose form is available to be varied; for example the temperature or rate of heat removal from the system.

The detailed forms of the vector functions f and g depend on the particular process considered and may be influenced by mass transfer coefficients, the detailed kinetics of chemical reactions, and similar physical considerations. The coefficient ratios α/γ and β/ϵ on the left hand sides of the equations are velocities of the flowing streams relative to the stationary apparatus; they will have the same sign in systems with co-current flow and opposite signs in the case of counter-current flow, and we shall assume that the coefficients α, β, γ, and ϵ are independent of x and t. Boundary conditions for equations (1) and (2) comprise initial conditions, namely the values of **u** (x) and **v** (x) at $t = 0$, together with entry conditions for the separate streams, both specified at the same end of the apparatus (say $x = 0$) in the case of co-current flow, or at opposite ends of the apparatus (say $x = 0$ and $x = x_1$) in the case of countercurrent flow. Thus **u** (t) at $x = 0$ and **v** (t) at $x = x_1$ would complete the set of boundary conditions for a counter-current apparatus, where **u** $(t, 0)$ and **v** (t, x_1) may either be given functions of time or may contain components which are under the control of the operator and may, therefore, like $\theta\ (x, t)$, be regarded as available to regulate the operation of the system.

In particular cases where the coefficient of the x or t derivative in equations (1) or (2) vanishes, rather less complete boundary conditions may be specified. For example, if $\alpha = 0$ only initial conditions **u** (**x**, 0) are specified, while if $\gamma = 0$ only boundary conditions **u** $(0, t)$ or **u** $(\mathbf{x}_1, t)$ may be specified. Similar considerations apply to **v** and a case of this type will be the subject of numerical study later in the paper.

If the operating period extends from $t = 0$ to $t = T$, one is then interested in the solution of equations (1) and (2) in the rectangular domain:

$$0 \leqslant x \leqslant x_1;\ 0 \leqslant t \leqslant T \qquad . \qquad . \qquad (3)$$

and, in particular, along the edges of this rectangle where boundary conditions are not specified, since these correspond to the process streams leaving the apparatus, and the state of the contents at the final time $t = T$.

Mathematically equations (1) and (2) are hyperbolic first-order partial differential equations. By a linear transformation of independent variables of the form:

$$X = at + bx;\ Y = ct + dx$$

it is always possible to reduce them to the canonical form:

$$\frac{\partial \mathbf{u}}{\partial X} = \mathbf{f}(\mathbf{u}, \mathbf{v}, \boldsymbol{\theta}) \qquad . \qquad . \qquad . \qquad (4)$$

$$\frac{\partial v}{\partial Y} = \mathbf{g}(\mathbf{u}, \mathbf{v}, \boldsymbol{\theta}) \qquad . \qquad . \qquad . \qquad (5)$$

at the cost of distorting the rectangular domain of interest (3) into a parallelogram with one corner at the origin. Attention may therefore be limited to equations of the form (4) and (5) without any real loss in generality.

* The University of Edinburgh and Heriot-Watt University, Chambers Street, Edinburgh 1.

† *Symbols have the meanings given them on p.* T169.

Optimisation Problems

Let Γ denote the boundary of the domain of interest, namely the rectangle in the original problem, or, more generally, the parallelogram in the problem transformed into canonical form. Let Γ_u, Γ_v denote those parts of Γ on which **u**, **v** respectively are either specified or available as a control variable, and let $\Gamma - \Gamma_u$ and $\Gamma - \Gamma_v$ denote the remaining parts of the boundary. Then optimisation problems of interest can frequently be formulated in terms of an objective function of the form:

$$P = \int_{\Gamma-\Gamma_u} \mathbf{l} \,.\, \mathbf{u} \mathrm{d}s + \int_{\Gamma-\Gamma_v} \mathbf{m} \,.\, \mathbf{v} \mathrm{d}s \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \qquad (6)$$

where **l** and **m** are specified vector functions of position on the segments of Γ indicated and s denotes distance along Γ. Very often, for example, one is concerned with the problem of maximising or minimising the time integral of some property of one or both exit streams from the unit (the total yield of a desired product in the time interval $0 \rightarrow T$ is of this type), in which case finite contributions to P arise only from the two sides $x = 0, x = x_1$ $(0 \leqslant t \leqslant T)$ of the boundary. The side $t = T (0 \leqslant x \leqslant x_1)$ gives a finite contribution only if the final hold-up of the apparatus is economically significant.

The starting point for many optimising computations is a first-order relationship between small changes in the adjustable variables and the consequent small change δP in P, and there is no difficulty in deriving such a relation for problems of the present class. Indeed, this has already been done in particular cases by Volin and Ostrovskii,[1, 2] Denn, Gray, and Ferron,[3] and the present writer.[4] For small variations the linearised form of equations (4) and (5), correct to first order in small quantities, is:

$$\frac{\partial}{\partial X} \delta \mathbf{u} = \frac{\partial \mathbf{f}}{\partial \mathbf{u}} \cdot \delta \mathbf{u} + \frac{\partial \mathbf{f}}{\partial \mathbf{v}} \cdot \delta \mathbf{v} + \frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}} \cdot \delta \boldsymbol{\theta} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \qquad (7)$$

$$\frac{\partial}{\partial Y} \delta \mathbf{v} = \frac{\partial \mathbf{g}}{\partial \mathbf{u}} \cdot \delta \mathbf{u} + \frac{\partial \mathbf{g}}{\partial \mathbf{v}} \cdot \delta \mathbf{v} + \frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}} \cdot \delta \boldsymbol{\theta} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \qquad (8)$$

where $\partial \mathbf{f}/\partial \mathbf{u}$, $\partial \mathbf{g}/\partial \mathbf{v}$ *etc.* represent matrices of partial derivatives* and the terms on the right hand sides are products of these matrices with the column vectors $\boldsymbol{\delta u}$, $\boldsymbol{\delta v}$, $\boldsymbol{\delta \theta}$.

Introducing row vectors $\boldsymbol{\psi}$ and $\boldsymbol{\chi}$ adjoint to **u** and **v** and satisfying the differential equations:

$$\frac{\partial \boldsymbol{\psi}}{\partial X} = - \boldsymbol{\psi} \cdot \frac{\partial \mathbf{f}}{\partial \mathbf{u}} - \boldsymbol{\chi} \cdot \frac{\partial \mathbf{g}}{\partial \mathbf{u}} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \qquad (9)$$

$$\frac{\partial \boldsymbol{\chi}}{\partial Y} = - \boldsymbol{\psi} \cdot \frac{\partial \mathbf{f}}{\partial \mathbf{v}} - \boldsymbol{\chi} \cdot \frac{\partial \mathbf{g}}{\partial \mathbf{v}} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \qquad (10)$$

it follows that:

$$\frac{\partial}{\partial X}(\boldsymbol{\psi} \,.\, \delta \mathbf{u}) + \frac{\partial}{\partial Y}(\boldsymbol{\chi} \,.\, \delta \mathbf{v}) = \left(\boldsymbol{\psi} \cdot \frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}} + \boldsymbol{\chi} \cdot \frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}\right) . \, \delta \boldsymbol{\theta}$$

Integrating both sides of this over the interior of the parallelogram and applying Gauss' theorem gives:

$$\oint_{\Gamma} (\tau_2 \boldsymbol{\psi} \,.\, \delta \mathbf{u} - \tau_1 \boldsymbol{\chi} \,.\, \delta \mathbf{v}) \mathrm{d}s = \int\!\!\int_{\Sigma} \left(\boldsymbol{\psi} \cdot \frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}} + \boldsymbol{\chi} \cdot \frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}\right) . \, \delta \boldsymbol{\theta} \, \mathrm{d}X \mathrm{d}Y$$

where Σ denotes the interior of the parallelogram with boundary Γ and τ_1 and τ_2 are components of the unit tangent to Γ.

Now if boundary conditions for $\boldsymbol{\psi}$ and $\boldsymbol{\chi}$ are chosen such that:

$$\tau_2 \boldsymbol{\psi} = \mathbf{l} \quad \text{on } \Gamma - \Gamma_u \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \qquad (11)$$

$$-\tau_1 \boldsymbol{\chi} = \mathbf{m} \quad \text{on } \Gamma - \Gamma_v \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \qquad (12)$$

the above may be written:

$$\int_{\Gamma-\Gamma_u} \mathbf{l} \,.\, \delta \mathbf{u} \, \mathrm{d}s + \int_{\Gamma-\Gamma_v} \mathbf{m} \,.\, \delta \mathbf{v} \, \mathrm{d}s = \int\!\!\int_{\Sigma} \left(\boldsymbol{\psi} \cdot \frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}} + \boldsymbol{\chi} \cdot \frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}\right) . \, \delta \boldsymbol{\theta} \, \mathrm{d}X \mathrm{d}Y - \int_{\Gamma_u} \tau_2 \, \boldsymbol{\psi} \,.\, \delta \mathbf{u} \, \mathrm{d}s + \int_{\Gamma_v} \tau_1 \, \boldsymbol{\chi} \,.\, \delta \mathbf{v} \, \mathrm{d}s$$

or:

$$\delta P = \int\!\!\int_{\Sigma} \left(\boldsymbol{\psi} \cdot \frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}} + \boldsymbol{\chi} \cdot \frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}}\right) . \, \delta \boldsymbol{\theta} \, \mathrm{d}X \mathrm{d}Y - \int_{\Gamma_u} \tau_2 \, \boldsymbol{\psi} \,.\, \delta \mathbf{u} \, \mathrm{d}s + \int_{\Gamma_v} \tau_1 \, \boldsymbol{\chi} \,.\, \delta \mathbf{v} \, \mathrm{d}s \quad . \quad . \quad . \qquad (13)$$

Equation (13) relates a small variation in the objective function P to small variations in the adjustable variables, namely $\boldsymbol{\theta}$ within Σ, **u** on Γ_u, and **v** on Γ_v. If **u** on Γ_u and **v** on Γ_v are *specified* inlet conditions rather than adjustable variables, $\delta \mathbf{u}$ and $\delta \mathbf{v}$ vanish on these segments and only the first term remains on the right hand side of equation (13).

* $\left(\frac{\partial \mathbf{f}}{\partial \mathbf{u}}\right)_{ij} = \frac{\partial f_i}{\partial u_j}$; $\left(\frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}}\right)_{ir} = \frac{\partial f_i}{\partial \theta_r}$, *etc.*

Equation (13) can be used as the basis of a number of computational schemes for maximisation or minimisation of P and one of the principal interests in studies of this type is the development of effective algorithms for numerical computation of optimum conditions. Confining attention, for the moment, to the problem with given values of $\mathbf{u}$ and $\mathbf{v}$ on Γ_u and Γ_v respectively, it is seen from equation (13) that a necessary condition for a stationary value of P is that:

$$\boldsymbol{\psi} \cdot \frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}} + \boldsymbol{\chi} \cdot \frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}} = 0 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (14)$$

throughout Σ. This provides a set of algebraic equations for the determination of $\boldsymbol{\theta}$ in terms of $\mathbf{u}$, $\mathbf{v}$, $\boldsymbol{\psi}$, and $\boldsymbol{\chi}$, and Volin and Ostrovskii[1] suggest that these should be used to eliminate $\boldsymbol{\theta}$ from the partial differential equations (4), (5), (9), and (10), which can then be integrated with the given boundary conditions. The solutions then determine $\boldsymbol{\theta}$ through equations (14). However, these authors do not report any numerical work based on this algorithm, and the present writer believes that it would be almost impracticably cumbersome, except in unrealistically simple cases where equations (14) could be solved to give $\boldsymbol{\theta}$ in closed form as a function of the physical and adjoint variables. A more promising approach is to use equation (13) to suggest means of improving an initial guess $\boldsymbol{\theta}_0$ at the optimum policy. For problems in which the second and third terms on the right hand side are absent, equation (13) reduces to:

$$\delta P = \iint_{\Sigma} \left(\boldsymbol{\psi} \cdot \frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}} + \boldsymbol{\chi} \cdot \frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}} \right) . \, \delta\boldsymbol{\theta} \, \mathrm{d}X \, \mathrm{d}Y$$

and any choice of $\delta\boldsymbol{\theta}$ such that:

$$\left(\boldsymbol{\psi} \cdot \frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}} + \boldsymbol{\chi} \cdot \frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}} \right) . \, \delta\boldsymbol{\theta} \geqslant 0 \text{ throughout } \Sigma \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (15)$$

will lead to an increase in the value of P, so that $\boldsymbol{\theta}_0 + \delta\boldsymbol{\theta}$ will be an improvement on the control policy $\boldsymbol{\theta}_0$. One particular variation satisfying equation (15) is given by:

$$\delta\boldsymbol{\theta} = \delta l \left(\boldsymbol{\psi} \cdot \frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}} + \boldsymbol{\chi} \cdot \frac{\partial \mathbf{g}}{\partial \boldsymbol{\theta}} \right)^T \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (16)$$

where δl is a small positive constant and the superscript ^{T} indicates the transpose. The change $\delta\boldsymbol{\theta}$ is then said to be in the direction of the *gradient* of P in the function space $\boldsymbol{\theta}$ (X, Y), and it is not difficult to show[5] that, among all variations satisfying:

$$\iint_{\Sigma} (\delta\boldsymbol{\theta})^2 \, \mathrm{d}X \, \mathrm{d}Y = \text{const}$$

(16) gives the largest increase in P. Successive modification of $\boldsymbol{\theta}$ by increments of the form (16) was first used to solve a chemical engineering problem by Horn and Troltenier,[6] though it has also been used in work on aerospace problems.[5] However, the work mentioned is all concerned with cases in which $\boldsymbol{\theta}$ is a function of one independent variable. To the present writer's knowledge there are, at present, no reports of the feasibility of using this technique in cases where $\boldsymbol{\theta}$ is a function of two independent variables, as in the present class of problems, though Volin and Ostrovskii[2] suggested that a gradient technique might prove useful without, however, carrying out any calculation or even formulating a specific computational algorithm. The principle object of the present paper, therefore, is to carry out a numerical exploration of the technique for a particular problem of this type, namely the optimum operation of a tubular catalytic reactor whose catalyst activity decays at a rate which varies from point to point depending on local conditions. A preliminary report of this work was presented[3] at the A.I.Ch.E.—I. Chem.E. Joint Meeting in London, 1965, and the remainder of the present paper reports the completion of this work and the resolution of many uncertainties remaining at the time of the preliminary report.

Before leaving the general problem, however, attention should be drawn to a particular situation in which difficulties arise. The validity of equation (13) depends on the choice of boundary conditions (11) and (12) for $\boldsymbol{\psi}$ and $\boldsymbol{\chi}$, and this choice is always acceptable unless the arc $\Gamma - \Gamma_u$ contains one or more finite segments on which $\tau_2 = 0$ and $\mathbf{l} \neq 0$ or the arc $\Gamma - \Gamma_v$ contains one or more finite segments on which $\tau_1 = 0$ and $\mathbf{m} \neq 0$. Difficulties therefore arise if there are finite contributions of the form $\mathbf{l} \, . \, \mathbf{u}$ to P from horizontal segments of $\Gamma - \Gamma_u$, or finite contributions of the form $\mathbf{m} \, . \, \mathbf{v}$ from vertical segments of $\Gamma - \Gamma_v$. Mathematically this is a result of the fact that these segments are parallel to characteristics of the hyperbolic differential equations (4) and (5) and in such a case δP can no longer be represented in the simple form of equation (13). This situation is discussed fully elsewhere[7] but does not arise in the numerical problem treated here.

A first-order Reversible, Exothermic Reaction with Temperature-dependent Catalyst Decay

As in the earlier report,[4] we shall consider a catalyst-packed tubular reactor with a single first-order reversible reaction whose effective velocity constants are proportional to a variable ϕ measuring the catalyst activity. It is assumed that the catalyst decays sufficiently slowly that conditions in the reactor approximate closely to a steady state throughout. If y_0 is the mole fraction of reaction product in the feed and ζ represents the stoichiometric extent of reaction, a function of distance x along the reactor, we then have:

$$\begin{aligned} \frac{\partial \zeta}{\partial x} &= \phi f(\zeta, \theta) = \phi \tau k_{20} \exp(-A_2/\theta)[(1 - y_0 - \zeta) K_0 \exp(Q/\theta) - (y_0 + \zeta)] \\ &= \phi[U(\theta) - V(\theta) \, . \, \zeta] \text{ (say)} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (17) \end{aligned}$$

Here $K_0 \exp(Q/\theta)$ is the equilibrium constant, with Q equal to the ratio of the heat of reaction and the gas constant, A_2 is the ratio of the activation energy of the reverse reaction to the gas constant, k_{20} is the frequency factor for the velocity constant of the reverse reaction, and τ is the total residence time in the reactor. The symbol x represents the fractional distance along the reactor, θ the temperature, and ϕ the catalyst activity, which is assumed to decay with increasing time according to the law:

$$\frac{\partial\phi}{\partial t} = -\frac{\theta}{\theta_c}\phi \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \tag{18}$$

where θ_c is a constant with the dimensions of temperature which determines the rate of decay and time t is expressed in dimensionless form as a fraction of the total interval of operation before catalyst renewal. The work can be simplified slightly by a change of variables from ϕ to $\lambda = \log_e \phi$, when equation (18) is replaced by:

$$\frac{\partial\lambda}{\partial t} = -\theta/\theta_c = g(\theta) \text{ (say)} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \tag{19}$$

The boundary conditions for equations (17) and (18) are then:

$$\zeta = 0 \text{ at } x = 0 \text{ (all } 0 \leqslant t \leqslant 1) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \tag{20}$$

and:

$$\lambda = 0 \text{ at } t = 0 \text{ (all } 0 \leqslant x \leqslant 1) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \tag{21}$$

The total production during the catalyst life is proportional to:

$$P = \int_0^1 \zeta(1, t)\,\mathrm{d}t \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \tag{22}$$

and it is required to choose the temperature policy $\theta(x, t)$ to maximise this quantity. The numerical values of the constants are the same as those taken previously,[4] namely:

$$\begin{aligned} \tau k_{20} &= 3 \times 10^7 \\ y_0 &= 0{\cdot}06 \\ A_2 &= 10\,000^\circ \text{ K} \\ K_0 &= 0{\cdot}00023 \\ Q &= 5000^\circ \text{ K} \\ \theta_c &= 250^\circ \text{ K} \end{aligned}$$

In the present problem it is seen that equations (17) and (19) are already in the canonical form referred to earlier so no change of independent variables is required and the adjoint equations corresponding to equations (9) and (10) in the general case are:

$$\frac{\partial\psi}{\partial x} = -e^{\lambda}\frac{\partial f}{\partial\zeta}\cdot\psi \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \tag{23}$$

and:

$$\frac{\partial\chi}{\partial t} = -e^{\lambda} f\psi \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \tag{24}$$

with boundary conditions:

$$\psi = 1 \text{ at } x = 1 \text{ (all } 0 \leqslant t \leqslant 1) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \tag{25}$$

$$\chi = 0 \text{ at } t = 1 \text{ (all } 0 \leqslant x \leqslant 1) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \tag{26}$$

respectively.

In the preliminary report[4] of this work equations (17) and (19) and their adjoints (23) and (24) were integrated numerically by a "marching" type finite difference method, and it was suspected that certain curious features of the final results might be attributable to a numerical instability in the integration procedure. In the present work this uncertainty has been eliminated by reducing the problem of solving the differential equations to one of evaluating definite integrals. The reduction to quadratures, or definite integration, is accomplished as follows: firstly equation (19) with boundary condition (21) can be integrated immediately to give:

$$\lambda(x, t) = -\frac{1}{\theta_c}\int_0^t \theta(x, t')\,\mathrm{d}t' \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \tag{27}$$

Equation (17) is a linear, first order, inhomogeneous equation in ζ with variable coefficients, so it may be integrated by the well known elementary device of multiplying by an integrating factor of the form $\exp\left(\int \phi V\,\mathrm{d}x\right)$, after which use of the boundary condition (20) leads to the result:

$$\zeta(x, t) = \int_0^x \exp\,[\lambda(x', t)]U(x', t)\,W(x, x', t)\,\mathrm{d}x' \quad . \quad . \quad . \quad . \quad . \quad . \quad . \tag{28}$$

with:

$$W(x, x', t) = \exp\left\{-\int_{x'}^{x} \exp\,[\lambda(x'', t)]V(x'', t)\,\mathrm{d}x''\right\} \quad . \quad . \quad . \quad . \quad . \quad . \tag{29}$$

so two definite integrals must be evaluated to give ζ for any particular x and t.

The adjoint equation (23) may be integrated directly after dividing through by ψ, giving:

$$\log_e \psi(x, t) = -\int_x^1 \exp\,[\lambda(x', t)]\,V(x', t)\,\mathrm{d}x' \quad . \quad . \quad . \quad . \quad . \quad . \quad . \tag{30}$$

and finally equation (24) may be integrated to give:

$$\chi(x, t) = \int_t^1 \exp\,[\lambda(x, t')]\{\,U\,(x, t') - V\,(x,t')\zeta(x, t')\}\psi(x, t')\,\mathrm{d}t' \qquad (31)$$

By performing the integrations in the orders (27), (29), (28), (30), (31), quantities required in the later stages are always evaluated earlier. The method of integration adopted was the trapezium rule with step lengths $\delta x = \delta t = 0{\cdot}04$, giving 25 integration steps in each variable and a total of 676 values of each function to be stored to provide tabulation at all points of the $x - t$ grid. To check the importance of truncation errors, a number of calculations were also carried out with δx reduced to $0{\cdot}02$, and in no case did this change the resulting temperature policy $\theta\,(x, t)$ by more than 0·2 degK in temperatures of the order 600° K. Once the integration was complete and the variables stored, the gradient of P in the function space $\theta\,(x, t)$ could be calculated at each grid point from the form of equation (13) appropriate to this problem, namely:

$$\delta P = \int_0^1 \int_0^1 P_\theta(x, t)\delta\theta(x, t)\,\mathrm{d}x\,\mathrm{d}t \qquad (32)$$

with

$$P_\theta = -\frac{\chi}{\theta_c} + \psi\exp(\lambda)\left(\frac{\partial U}{\partial\theta} - \zeta\frac{\partial V}{\partial\theta}\right) \qquad (33)$$

The values of P_θ were used in a simple stepwise steepest ascent procedure in θ-space. From an initial estimate $\theta_0\,(x, t)$ of the temperature policy, a corresponding function $P_\theta\,(x, t)$ was computed in the manner just described, and a sequence of improved policies $\theta\,(x, t)$ was obtained from the formula:

$$\theta(x, t) = \theta_0(x, t) + lP_\theta(x, t) \qquad (34)$$

using successively increasing values of l. For each value of l the corresponding value of the objective function was computed from equation (22) after using equations (27) and (28) to determine ζ. The sequence of values of P thus obtained was printed out, and the value $l = l_m$ which maximised P was located. The new temperature policy:

$$\theta_1 = \theta_0 + l_m P_\theta \qquad (35)$$

then replaced θ_0 as the starting policy for a further iteration of the process.

In spaces of finite dimensionality this simple stepwise steepest ascent procedure is known to converge very poorly in many cases, so its use initially in the present work was purely exploratory, and it was expected that replacement by some more sophisticated technique might be necessary to obtain useful results.

Results of Numerical Work

The first set of computations started from a uniform and constant temperature policy $\theta\,(x, t) = 600°$ K, which is not very far from the optimum policy of this very restricted class. Figs 1A to 1G show this policy and the policies at the end of successive ascents in the function space of $\theta\,(x, t)$, performed as described in the last section, plotted as functions of x for $t = 0$ and for $t = 1$. The corresponding curves for other values of t are intermediate in nature between those plotted, and are omitted in the interest of clarity. Fig. 1H shows the corresponding sequence of values of the objective function and this appears to be converging in a satisfactory maner by the end of the sixth ascent. The temperature policy at the end of the third ascent (Fig. 1D) shows oscillations which were previously noted in the preliminary report of this work,[4] but if these are disregarded they become damped out in the succeeding ascents until the temperature policy at the end of the sixth ascent retains only a trace of oscillation in the neighbourhood of its large change of slope.

There is no guarantee, of course, that the temperature policy given in Fig. 1G approximates to the true optimum rather than a secondary maximum of P, but one's confidence would be increased by a second calculation starting from completely different initial conditions. Accordingly such a calculation was carried out starting from the policy shown in Fig. 2A, chosen to differ markedly from both Fig. 1A and Fig. 1G. The resulting policies at the ends of the second, fourth, sixth, eight, tenth, and twelfth ascents are shown in Figs 2B to 2G and the corresponding sequence of values of P is given in Fig. 2H. To avoid numerical difficulty in evaluating the velocity constants accurately a lower bound of 500° K was placed on the temperature throughout. If the steepest ascent procedure prescribed a temperature below 500° K at any stage of the calculations, this value was simply replaced by the lower bound. The final result of Fig. 2G is a policy with temperatures everywhere above 500° K, so the imposed lower bound is of no importance ultimately; indeed it could probably have been taken lower without difficulty, since the calculations of Figs 1A to 1G make use of temperatures well below this level.

The most striking feature of these results is the very large oscillation which develops in the temperature policy, so that after the eighth ascent there seems to be very little prospect of convergence to a reasonable final result. Nevertheless, four further ascents prove sufficient to smooth out this oscillation almost completely, and after the twelfth ascent the value of the objective function is converging well and indeed is a little higher than the final value in the previous calculation. Although the final values of the objective function in the two calculations agree quite closely, the sharp drops near $x = 0$ and $x = 1$ in the initial profile of Fig. 1G are absent from Fig. 2G. Thus the policies giving rise to near optimal values of P differ quite significantly from each other. In other words, the maximum of P in the function space is very flat, and one might expect to find a variety of functional forms for $\theta\,(x, t)$ giving performances very near to the optimum. This is not very surprising, and is in any case valuable information, since it may be possible to find quite simple policies which are almost optimal. With this in view a third set of calculations was carried out, limiting the search to the class of policies

$$\theta(x, t) = A + Bx + Ct \qquad (36)$$

The search method employed was that of Powell,[8] which does not make use of gradients. In this case $\theta\,(x, t)$ was simply regarded as a function of the three variables A, B, and C on which a search was initiated from two different starting points. The initial values of A, B, C, and P and their values at the end of each successive iteration of Powell's

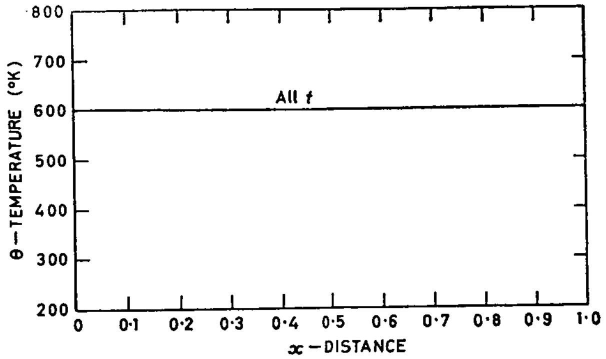


Fig. 1A.—*Uniform temperature policy used to start the steepest ascent procedure*

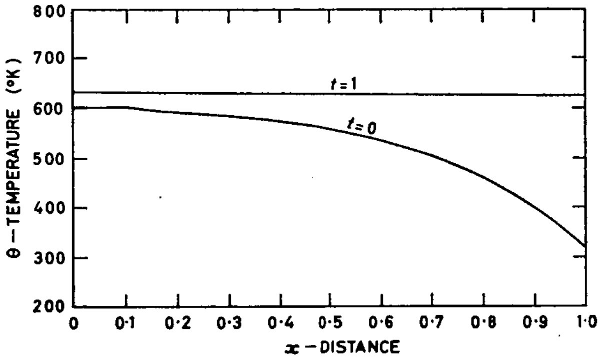


Fig. 1B.—*Temperature policy at the end of the first ascent*

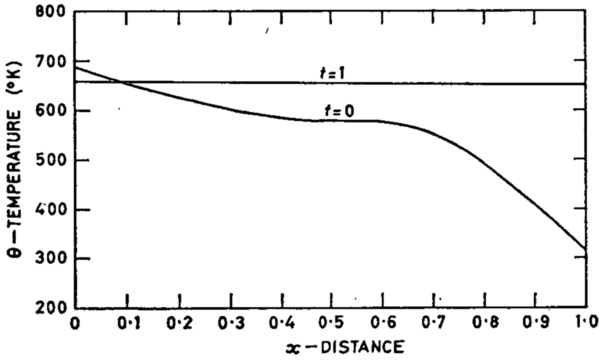


Fig. 1C.—*Temperature policy at the end of the second ascent*

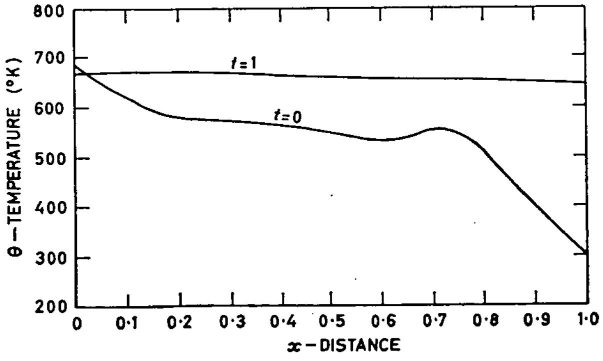


Fig. 1D.—*Temperature policy at the end of the third ascent*

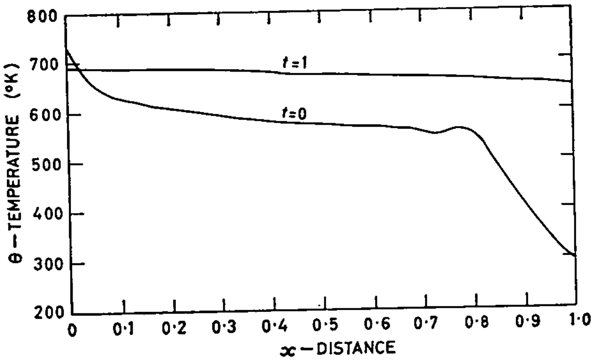


Fig. 1E.—*Temperature policy at the end of the fourth ascent*

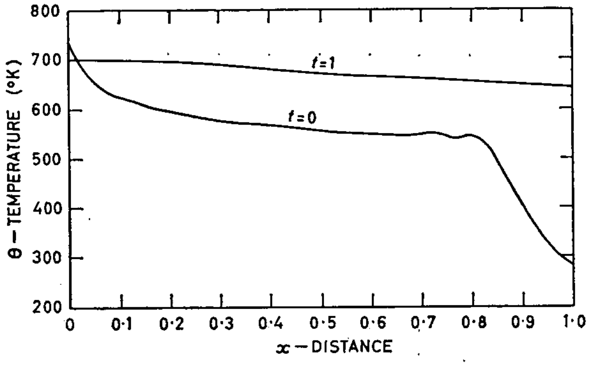


Fig. 1F.—*Temperature policy at the end of fifth ascent*

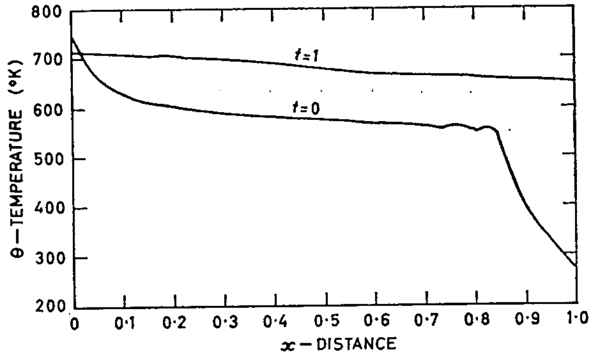


Fig. 1G.—*Temperature policy at the end of the sixth ascent*

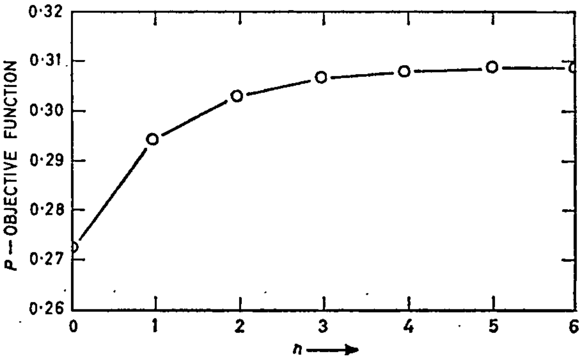


Fig. 1H.—*Values of the objective function after successive ascents*

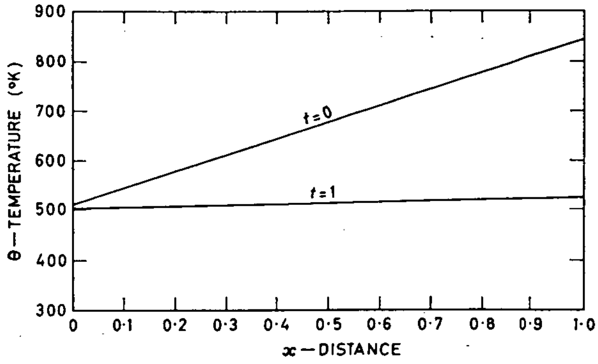


Fig. 2A.—*Non-uniform temperature policy used to start the steepest ascent procedure*

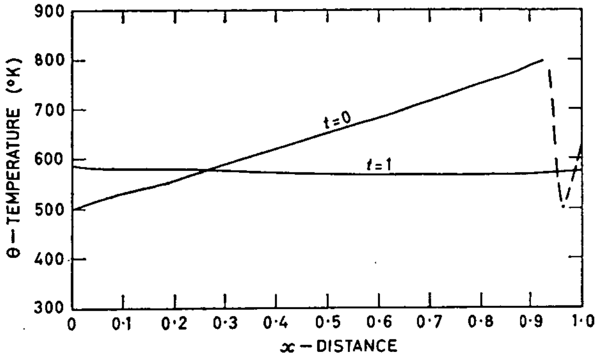


Fig. 2B.—*Temperature policy at the end of the second ascent*

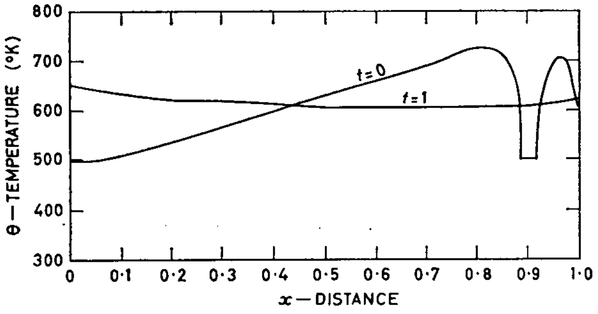


Fig. 2C.—*Temperature policy at the end of the fourth ascent*

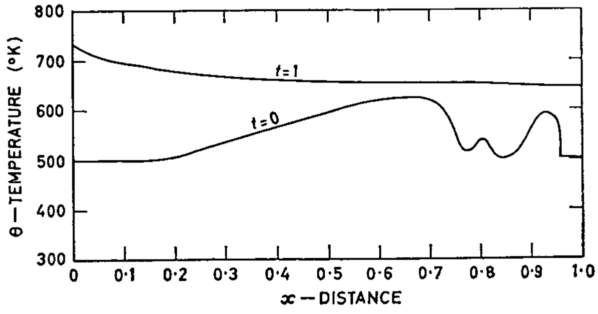


Fig. 2D.—*Temperature policy at the end of the sixth ascent*

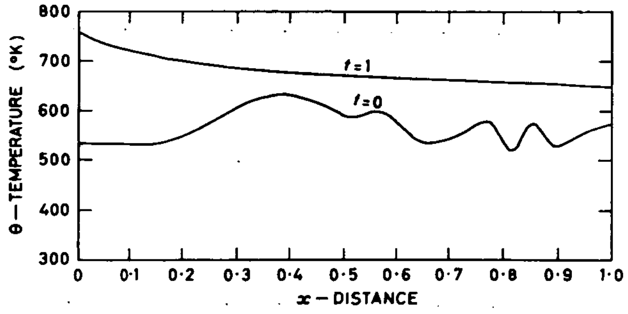


Fig. 2E.—*Temperature policy at the end of the eighth ascent*

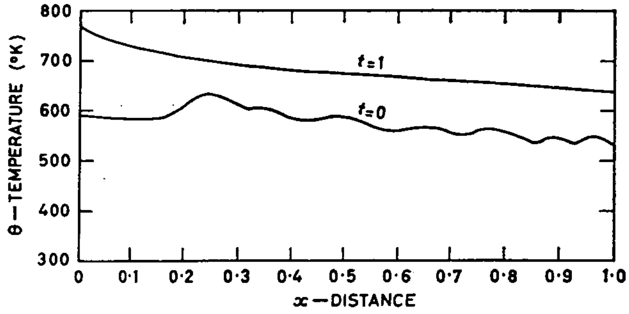


Fig. 2F.—*Temperature policy at the end of the tenth ascent*

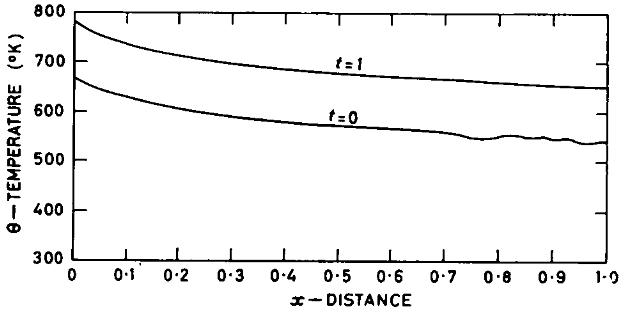


Fig. 2G.—*Temperature policy at the end of the twelfth ascent*

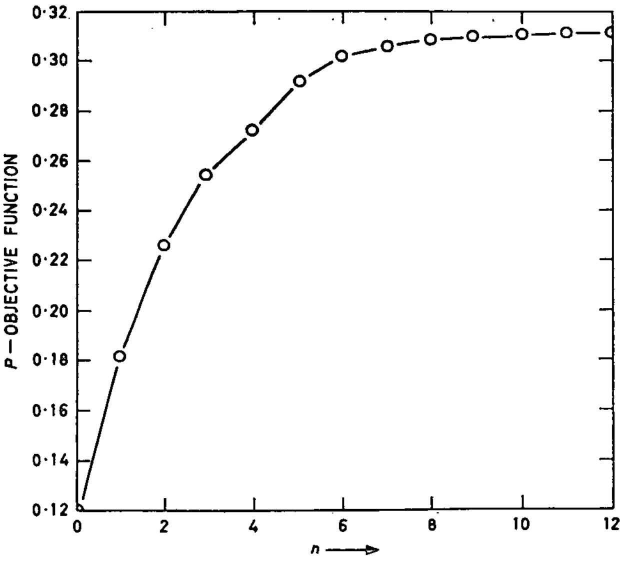


Fig. 2H.—*Values of the objective function after successive pairs of ascents*

basic procedure, are given in Table I from which it is seen that P converges to 0·307646 to six decimal places after only four iterations from either starting point. This value differs from the highest value given in Figs 1H or 2H only in the third significant figure, so it is possible, as suspected, to obtain almost optimal results within a class of policies as limited as (36) above. It should be noted, however, that the economy in computation obtainable by restricting attention to a class of policies such as (36) is not as large as might be imagined. In spite of the curious oscillations apparent in the intermediate stages of Figs 1 and 2, convergence to a near optimal value of P is remarkably rapid; indeed it is much better than could be expected in view of the usually poor performance of a simple stepwise steepest ascent procedure in a space of finite dimensionality. Thus, contrary to expectation, the results indicate that hill climbing in a function space—even a function of two independent variables—is quite a practical proposition provided that one is interested primarily in the value of the objective function rather than the form of the optimising policy. Although experience of actual computations by this method is still quite limited, this conclusion has also been reached by other workers[6, 9] on the basis of a limited number of computations.

TABLE I.—*Values of Parameters Before and After Iterations according to Powell's Basic Procedure*

Starting Conditions	$A = 400$	$A = 600$
	$B = 0$	$B = 0$
	$C = 0$	$C = 0$
	$P = 0·011950$	$P = 0·271967$
1st iteration	$A = 595·75$	$A = 605·104$
	$B = -15·02$	$B = -20·017$
	$C = 50·688$	$C = 67·557$
	$P = 0·290900$	$P = 0·289899$
2nd iteration	$A = 587·597$	$A = 606·033$
	$B = -48·817$	$B = -85·995$
	$C = 120·622$	$C = 118·662$
	$P = 0·302245$	$P = 0·306588$
3rd iteration	$A = 615·554$	$A = 607·980$
	$B = -102·97$	$B = -87·237$
	$C = 111·898$	$C = 111·611$
	$P = 0·307084$	$P = 0·306739$
4th iteration	$A = 623·285$	$A = 623·319$
	$B = -106·376$	$B = -106·706$
	$C = 107·528$	$C = 107·634$
	$P = 0·307646$	$P = 0·307646$
5th iteration	$A = 623·319$	$A = 623·318$
	$B = -106·344$	$B = -106·464$
	$C = 107·392$	$C = 107·391$
	$P = 0·307646$	$P = 0·307646$
6th iteration	$A = 623·334$	$A = 623·318$
	$B = -106·369$	$B = -106·343$
	$C = 107·386$	$C = 107·392$
	$P = 0·307646$	$P = 0·307646$

The source of the "self-healing" oscillations, which appear only in the earlier stages of the steepest ascent procedure, is still obscure. They must be clearly distinguished from the well-known oscillations in the values of the independent variables after successive ascents, which almost invariably accompany this type of hill-climbing procedure when used in spaces of finite dimensionality. The present oscillations are exhibited when θ is plotted as a function of the dimension index x which, in this case, is a continuous variable. In spaces of finite dimensionality N the dimension index takes only integer values $1 \rightarrow N$, so when N is small, as in most published applications of hill climbing, such a plot consists of just a small number of isolated points and there is no discernible corresponding phenomenon.

Finally, it should be noted that Chou, Ray, and Aris[10] have very recently obtained some results in closed form for a catalyst decay problem which is a simplified version of the one discussed here. By considering an irreversible reaction, whose rate depends on the value of only one velocity constant, they are able to show that the optimal operating policy must be such that the overall conversion remains constant throughout the catalyst life.

Acknowledgments

This work was carried out while the author was Visting Professor at Rice University in the session 1965–66, and was supported by the National Science Foundation grant for Systems Research, GU-1153. He would like to thank Mr. Mao-Jen Tsai who carried out the computations reported in Table I. Thanks are also due to Professor R. Aris for the opportunity to see Ref. ([10]) before publication.

Symbols Used

A = constant used in defining a simple temperature policy through equation (36).

A_2 = ratio of the activation energy of the reverse reaction to the universal gas constant.

a = coefficient in the transformation from (t, x) to (X, Y).

B = constant used in defining a simple temperature policy through equation (36).

b = coefficient in the transformation from (t, x) to (X, Y).

C = constant used in defining a simple temperature policy through equation (36).

c = coefficient in the transformation from (t, x) to (X, Y).

d = coefficient in the transformation from (t, x) to (X, Y).

$\mathbf{f}$ = vector function of $\mathbf{u}$, $\mathbf{v}$, $\boldsymbol{\theta}$ in equation (1).

$\mathbf{g}$ = vector function of $\mathbf{u}$, $\mathbf{v}$, $\boldsymbol{\theta}$ in equation (2).

$g(\theta)$ = Function $-\theta/\theta_c$.

K_0 = temperature independent factor in the equilibrium constant.

k_{20} = pre-exponential factor in the velocity constant of the reverse reaction.

l = displacement along the steepest ascent line.

$\mathbf{l}$ = vector given on Γ and defining the objective function through equation (6).

l_m = value of l which maximises P.

$\mathbf{m}$ = vector given on Γ and defining the objective function through equation (6).

P = objective function to be maximised.

P_θ = gradient of P in θ-space.

Q = ratio of the heat of reaction to the universal gas constant.

s = distance, measured along Γ.

T = length of time interval of interest.

t = time.

t' = variable of integration in equation (31).

U = function of θ defined by equation (17).

$\mathbf{u}$ = column vector of physical variables associated with the first phase.

V = function of θ defined by equation (17).

$\mathbf{v}$ = column vector of physical variables associated with the second phase.

X = variable related to t and x by the transformation $X = at + bx$.

x = distance measured in the direction of flow.

x_1 = total length of the apparatus.

x' = variable of integration in equations (29) and (30).

Y = variable related to t and x by the transformation $Y = ct + dx$.
y_0 = mole fraction of the reaction product in the feed mixture.
α = proportional to velocity of the first phase.
β = proportional to velocity of the second phase.
Γ = boundary of the region of interest.
Γ_u = part of Γ on which $\mathbf{u}$ is specified or available for control.
Γ_v = part of Γ on which $\mathbf{v}$ is specified or available for control.
ζ = stoichiometric extent of reaction.
θ = absolute temperature.
$\boldsymbol{\theta}$ = column vector of variables available for control.
θ_c = characteristic temperature determining the rate of catalyst decay.
θ_0 = value of $\theta\,(x, t)$ at the beginning of an ascent in θ-space.
λ = defined by $\lambda = \log_e \phi$.
Σ = region contained within Γ.
τ = residence time in the reactor.
τ_1, τ_2 = components of the unit tangent vector to Γ.
ϕ = catalyst activity.
ψ = variable adjoint to ζ.
$\boldsymbol{\psi}$ = row vector of variables adjoint to $\mathbf{u}$.
χ = variable adjoint to ϕ.
$\boldsymbol{\chi}$ = row vector of variables adjoint to $\mathbf{v}$.

The above quantities may be expressed in any set of consistent units in which force and mass are not defined independently.

References

[1] Volin, Yu. M., and Ostrovskii, G. M. *Avtomatika Telemekh.*, 1964, **25**, 1414.

[2] Volin, Yu. M., and Ostrovskii, G. M. *Avtomatika Telemekh.*, 1965, **26**, 1197.

[3] Denn, M. M., Gray, R. D., and Ferron, J. R. *Ind. Engng Chem., Fundamentals*, 1966, **5**, 59.

[4] Jackson, R., in Pirie, J. M. (Ed.). "*The Application of Mathematical Models in Chemical Engineering Research, Design and Production*", 1965, p. 33. (London: The Institution of Chemical Engineers).

[5] Leitman, G. (Ed.). "*Optimisation Techniques*", 1962, Chapter 6. (New York: Academic Press Inc.).

[6] Horn, F., and Troltenier, U. *Chemie-Ingr-Tech.*, 1960, **32**, 382.

[7] Jackson, R. *International Journal of Control*, T.B.P.

[8] Powell, M. J. D. *Comput. J.*, 1964, **7**, 155.

[9] Horn, F. *Private Communication*, 1965

[10] Chou, A., Ray, W. H., and Aris, R. *Trans. Instn chem. Engrs*, 1967, **45**, 153.

The manuscript of this paper was received on 28 November, 1966.

C5

Int. J. Control, 1966, Vol. 4, No. 2, 127–136

Optimization Problems in a Class of Systems described by Hyperbolic Partial Differential Equations

Part I. Variational Theory†

By R. Jackson

University of Edinburgh and Heriot-Watt University,
Chemical Engineering Laboratories, Chambers Street, Edinburgh 1

[Received October 4, 1966]

Abstract

Distributed systems such as chemical reactors, absorption columns and other similar apparatus used in the chemical industry have time-varying behaviour approximately describable by hyperbolic partial differential equations. Questions of optimum start-up and control therefore find mathematical expression as variational problems in two independent variables with these differential equations as side conditions. Such problems have interesting mathematical features when the integral to be extremized is taken round a closed curve in the plane of the independent variables, and this curve includes finite straight segments parallel to the characteristics of the differential equations.

§ 1. Introduction

A wide class of optimization and optimum control problems in chemical engineering can be formulated mathematically as variational problems with sets of first-order hyperbolic partial differential equations as side conditions. It has been noted elsewhere (Jackson 1966) that such problems have interesting mathematical features when the boundary of the domain of interest includes finite segments parallel to characteristics of the constraining differential equations. It is the purpose of the present paper to pursue this point for a class of problems which is a slight generalization of those encountered in chemical engineering systems.

Consider functions u and v satisfying differential equations of the form:

$$\frac{\partial u}{\partial x}=f(u,v,\theta), \tag{1}$$

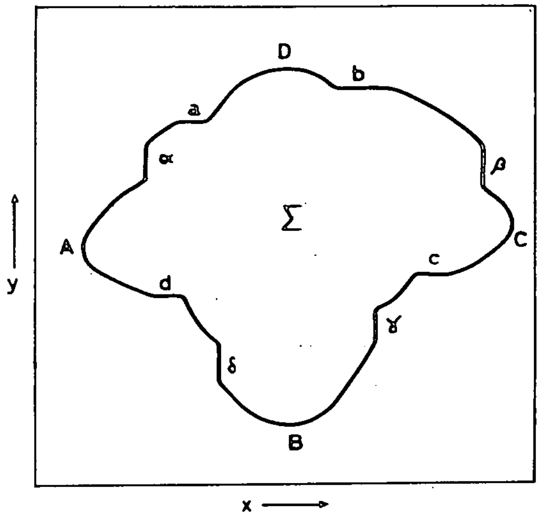
$$\frac{\partial v}{\partial y}=g(u,v,\theta), \tag{2}$$

where f and g are given functions and $\theta(x,y)$ is available to be varied. The domain of interest, $\sum$, in the x,y plane is bounded by a closed curve Γ

† Communicated by Dr. A. T. Fuller.

which may include finite straight segments parallel to the coordinate axes, as in fig. 1, where the segments parallel to the x-axis are denoted by a, b, c, d, and those parallel to the y-axis by $\alpha, \beta, \gamma, \delta, \ldots$. In general, of course, there may be any number of such segments of each kind. However, it will be assumed that no two segments parallel to the x-axis have the same ordinate and no two segments parallel to the y-axis have the same abscissa. C and A are the points of Γ with largest and smallest abscissa respectively, while D and B are the points with largest and smallest ordinate.

Fig. 1



Boundary conditions for u are specified by giving its value at all points of the arc DAB not lying on horizontal segments, and also at the left-hand end points of these segments. The set of points so defined will be denoted by Γ_u, so u is specified on the sub-set Γ_u of Γ. Similarly boundary conditions for v are specified by giving its value at all points of the arc ABC not lying on vertical segments, and also at the lower end points of these segments. The set of points so defined will be denoted by Γ_v. These boundary conditions, together with eqns. (1) and (2) are sufficient to determine u and v throughout $\sum$ and, in particular, at all points of Γ. It is then required to find that function $\theta(x,y)$, in $\sum$, which maximizes an integral of the form:

$$I = \oint_{\Gamma} (lu + mv)\, ds, \tag{3}$$

where ds is the magnitude of a small displacement along Γ, and l and m are specified functions of position on Γ.

The problem, as just stated, differs from problems of importance in chemical engineering in two respects; firstly eqns. (1) and (2) are normally replaced by the more general forms:

$$\alpha\frac{\partial u}{\partial x}+\beta\frac{\partial u}{\partial y}=f(u,v,\theta), \tag{1'}$$

$$\gamma\frac{\partial v}{\partial x}+\delta\frac{\partial v}{\partial y}=g(u,v,\theta), \tag{2'}$$

in practical problems, and secondly the boundary Γ is most frequently a simple rectangle. The first difference is unimportant, since a linear transformation of the independent variables can always be used to transform (1′), (2′) into (1), (2). This transformation will distort the rectangular region of interest into a parallelogram, so the introduction of the more general boundary Γ genuinely widens the class of problems treated.

Certain variational problems with hyperbolic partial differential equations as side conditions have been considered by Egorov (1964), but the interesting features of the present class of problems do not appear in Egorov's work.

§ 2. The Effect of the Characteristic Segments

Variational problems are usually dealt with by expressing a small variation δI to first order as an integral over the region of interest of the corresponding small variation $\delta\theta$ in the available variable. The resulting expression can either be employed as it stands in one of the direct methods of the calculus of variations (Leitmann 1962), or used to provide necessary conditions for a stationary value of I in the form of Euler–Lagrange equations. Thus the central mathematical problem is the expansion of δI to first order in $\delta\theta$, and in order to accomplish this in the present case consider the linearized form of eqns. (1) and (2) relating small variations in u, v and θ, namely

$$\frac{\partial}{\partial x}(\delta u)=\frac{\partial f}{\partial u}\delta u+\frac{\partial f}{\partial v}\delta v+\frac{\partial f}{\partial \theta}\delta\theta \tag{4}$$

and

$$\frac{\partial}{\partial y}(\delta v)=\frac{\partial g}{\partial u}\delta u+\frac{\partial g}{\partial v}\delta v+\frac{\partial g}{\partial \theta}\delta\theta. \tag{5}$$

Introduce variables ψ and χ, adjoint to u and v respectively and satisfying the differential equations:

$$\frac{\partial\psi}{\partial x}=-\psi\frac{\partial f}{\partial u}-\chi\frac{\partial g}{\partial u} \tag{6}$$

and

$$\frac{\partial\chi}{\partial y}=-\psi\frac{\partial f}{\partial v}-\chi\frac{\partial g}{\partial v}\cdot \tag{7}$$

Then

$$\frac{\partial}{\partial x}(\psi\delta u)+\frac{\partial}{\partial y}(\chi\delta v)=\left(\psi\frac{\partial f}{\partial\theta}+\chi\frac{\partial g}{\partial\theta}\right)\delta\theta.$$

Integrating over $\sum$ and using Gauss's theorem, this gives:

$$\oint_{\Gamma}(\tau_2\psi\delta u-\tau_1\chi\delta v)\,ds=\iint_{\Sigma}\left(\psi\frac{\partial f}{\partial\theta}+\chi\frac{\partial g}{\partial\theta}\right)\delta\theta\,dx\,dy, \tag{8}$$

where (τ_1,τ_2) are the components of the unit tangent to Γ in the direction of integration.

Now $\delta u=0$ on Γ_u and $\delta v=0$ on Γ_v, so if we could choose

$$\tau_2\psi=l \qquad \text{at all points of } \Gamma-\Gamma_u, \tag{9}$$

and

$$-\tau_1\chi=m \qquad \text{at all points of } \Gamma-\Gamma_v, \tag{10}$$

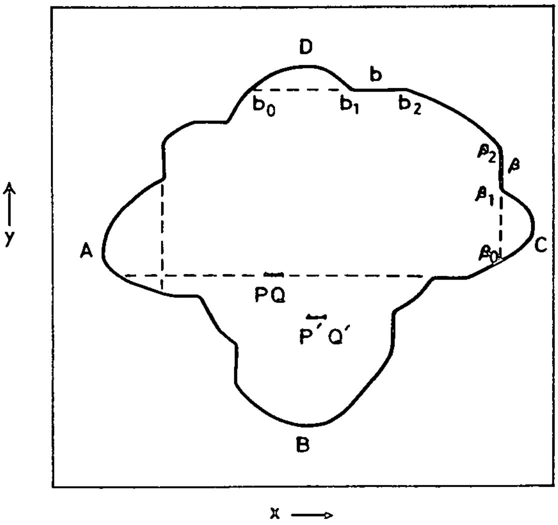
the left-hand side of (8) would be equal to δI and the desired expansion would be accomplished.

However, (9) cannot be satisfied on the horizontal segments of $\Gamma-\Gamma_u$ (unless $l=0$ there), since $\tau_2=0$ at all points of these segments, and similarly (10) cannot be satisfied on the vertical segments of $\Gamma-\Gamma_v$ (unless $m=0$ there), since $\tau_1=0$ at all points of these segments. Thus (8) cannot be reduced to an expansion of δI by choice of the boundary conditions for the adjoint variables when Γ contains finite horizontal and vertical segments which make contributions to both terms in the integrand of I. The remainder of the paper will show how this difficulty can be overcome.

§ 3. The First Variation

Referring to fig. 1, produce into the interior of $\sum$ all horizontal segments of the arc BCD and all vertical segments of the arc CDA, thus obtaining fig. 2.

Fig. 2



If b is a typical horizontal segment, denote its end points by b_1 and b_2, as indicated, and the point where its projection meets Γ again by b_0. For horizontal segments of the arc DAB, b_0 may be regarded as coincident with b_1. The corresponding points of a typical vertical segment β are denoted by β_0, β_1 and β_2, and if the segment belongs to the arc ABC, β_0 may be regarded as coincident with β_1.

Now introduce a set of functions $\lambda_b(x)$, defined on the lines b_0, b_1, b_2, where they satisfy the differential equations:

$$\frac{d\lambda_b}{dx} = -\lambda_b \frac{\partial f}{\partial u} - l_b(x), \tag{11}$$

with

$$\lambda_b(b_2) = 0, \tag{12}$$

and

$$\begin{aligned} l_b(x) &\equiv l(x) \qquad \text{on } b_1 \to b_2 \\ &= 0 \qquad \text{on } b_0 \to b_1. \end{aligned} \tag{13}$$

Then from (4) and (11):

$$\frac{d}{dx}(\lambda_b \delta u) + l_b \delta u = \lambda_b \frac{\partial f}{\partial \theta}\delta\theta + \lambda_b \frac{\partial f}{\partial v}\delta v$$

on $b_0 \to b_2$, and integrating both sides of this between b_0 and b_2 gives:

$$\int_{b_1}^{b_2} l\delta u\, dx = \int_{b_0}^{b_2} \lambda_b \frac{\partial f}{\partial \theta}\delta\theta\, dx + \int_{b_0}^{b_2} \lambda_b \frac{\partial f}{\partial v}\delta v\, dx,$$

since $\lambda_b = 0$ at b_2 and $\delta u = 0$ at b_0. This may alternatively be written:

$$\int_{b_1}^{b_2} \left(l\delta u - \lambda_b \frac{\partial f}{\partial v}\delta v \right) dx = \int_{b_0}^{b_2} \lambda_b \frac{\partial f}{\partial \theta}\delta\theta\, dx + \int_{b_0}^{b_2} \lambda_b \frac{\partial f}{\partial v}\delta v\, dx. \tag{14}$$

A set of functions $\mu_\beta(y)$ may similarly be introduced on the vertical segments where they satisfy the differential equations:

$$\frac{d\mu_\beta}{dy} = -\mu_\beta \frac{\partial g}{\partial v} - m_\beta(y) \tag{15}$$

with

$$\mu_\beta(\beta_2) = 0 \tag{16}$$

and

$$\begin{aligned} m_\beta(y) &\equiv m(y) \text{ on } \beta_1 \to \beta_2 \\ &= 0 \text{ on } \beta_0 \to \beta_1. \end{aligned} \tag{17}$$

Then similar reasoning to the above shows that

$$\int_{\beta_1}^{\beta_2} (m\delta v - \mu_\beta \frac{\partial g}{\partial u}\delta u)\, dy = \int_{\beta_0}^{\beta_2} \mu_\beta \frac{\partial g}{\partial \theta}\delta\theta\, dy + \int_{\beta_0}^{\beta_2} \mu_\beta \frac{\partial g}{\partial u}\delta u\, dy. \tag{18}$$

The lines obtained by producing horizontal and vertical segments of Γ divide $\sum$ into a number of sub-regions. Denote by Γ_i the closed boundary of the ith sub-region $\sum_i$ and in each region introduce two functions ψ and χ satisfying the differential eqns. (6) and (7). Then, applying Gauss's theorem to the sub-region $\sum_i$ it follows, in the same way as (8) was obtained from (6) and (7), that

$$\oint_{\Gamma_i} (\tau_2 \psi \delta u - \tau_1 \chi \delta v)\, ds = \iint_{\Sigma_i} \left(\psi \frac{\partial f}{\partial \theta} + \chi \frac{\partial g}{\partial \theta}\right) \delta\theta \, dx\, dy. \tag{19}$$

Provided the integrals on the left-hand side are all taken round the bounding curves in the same sense (say, anti-clockwise), they may be added to give an integral round Γ together with contributions from each of the lines obtained above by producing horizontal and vertical segments into $\sum$. Thus, on adding eqns. (19) for all i:

$$\oint_{\Gamma} (\tau_2 \psi \delta u - \tau_1 \chi \delta v)\, ds + \sum_b \int_{b_0}^{b_1} (\chi_l - \chi_u)\delta v\, dx$$
$$+ \sum_\beta \int_{\beta_0}^{\beta_1} (\psi_l - \psi_r)\, \delta u\, dy = \iint_{\Sigma} \left(\psi \frac{\partial f}{\partial \theta} + \chi \frac{\partial g}{\partial \theta}\right) \delta\theta\, dx\, dy, \tag{20}$$

where χ_l denotes the value taken by χ on approaching the projection of a horizontal segment from below and χ_u the value taken on approaching this line from above, and similarly ψ_l and ψ_r denote the values taken by ψ on approaching the projection of a vertical segment from the left and right respectively.

Now denote by $\tilde{\Gamma}$ that part of Γ which belongs neither to horizontal nor vertical segments, so that the left-hand side of eqn. (20) may be written:

$$\int_{\Gamma} (\tau_2 \psi \delta u - \tau_1 \chi \delta v)\, ds = \int_{\tilde{\Gamma}} (\tau_2 \psi \delta u - \tau_1 \chi \delta v)\, ds + \sum_b \int_b (-\tau_1 \chi \delta v)\, ds$$
$$+ \sum_\beta \int_\beta (\tau_2 \psi \delta u)\, ds. \tag{21}$$

Here

$$\int_b (-\tau_1 \chi \delta v)\, ds$$

denotes the integral along the horizontal boundary segment b in the sense of traversal of Γ, and will therefore be

$$\int_{b_1}^{b_2} (-\chi \delta v)\, dx$$

in some cases and

$$\int_{b_2}^{b_1} (-\chi \delta v)\, dx$$

in others. Using the same notation, the left-hand sides of eqns. (14) and (18) may be written as:

$$\int_b \left(l\delta u - \lambda_b \frac{\partial f}{\partial v}\delta v\right) ds \quad \text{and} \quad \int_\beta \left(m\delta v - \mu_b \frac{\partial g}{\partial u}\delta u\right) ds$$

respectively.

By adding all eqns. (14) and (18) to eqn. (20), using the form (21) for the left-hand side of (20), we may obtain:

$$\int_{\tilde{\Gamma}} (\tau_2\psi\delta u - \tau_1\chi\delta v)\, ds + \sum_b \int_b \left[l\delta u - \left(\lambda_b \frac{\partial f}{\partial v} + \tau_1\chi\right)\delta v\right] ds$$
$$+ \sum_\beta \int_\beta \left[m\delta v - \left(\mu_\beta \frac{\partial g}{\partial u} - \tau_2\psi\right)\delta u\right] ds + \sum_b \int_{b_0}^{b_1} \left(\chi_l - \chi_u - \lambda_b \frac{\partial f}{\partial v}\right)\delta v\, dx$$
$$+ \sum_\beta \int_{\beta_0}^{\beta_1} \left(\psi_l - \psi_r - \mu_\beta \frac{\partial g}{\partial u}\right)\delta u\, dy = \sum_b \int_{b_0}^{b_2} \lambda_b \frac{\partial f}{\partial \theta}\delta\theta\, dx + \sum_\beta \int_{\beta_0}^{\beta_2} \mu_\beta \frac{\partial g}{\partial \theta}\delta\theta\, dy$$
$$+ \iint_\Sigma \left(\psi \frac{\partial f}{\partial \theta} + \chi \frac{\partial g}{\partial \theta}\right)\delta\theta\, dx\, dy. \tag{22}$$

Now δu is finite on $\Gamma - \Gamma_u$, where u is not specified as a boundary condition. Let us therefore choose ψ so that

$$\tau_2\psi = l, \tag{23}$$

at points common to $\tilde{\Gamma}$ and $\Gamma - \Gamma_u$. Referring to fig. 1, this specifies a boundary condition for ψ at points of the arc BCD not lying on horizontal or vertical segments. Similarly, since δv is finite on $\Gamma - \Gamma_v$, we may choose

$$-\tau_1\chi = m, \tag{24}$$

at points common to $\tilde{\Gamma}$ and $\Gamma - \Gamma_v$, and this specifies a boundary condition for χ at points of the arc CDA not lying on horizontal or vertical segments. Then from eqns. (23) and (24):

$$\int_{\tilde{\Gamma}} (\tau_2\psi\delta u - \tau_1\chi\delta v)\, ds = \int_{\tilde{\Gamma}} (l\delta u + m\delta v)\, ds. \tag{25}$$

On horizontal segments belonging to the arc ABC where v is specified, $\delta v = 0$ and

$$\int_b \left[l\delta u - \left(\lambda_b \frac{\partial f}{\partial v} + \tau_1\chi\right)\delta v\right] ds = \int_b l\delta u\, ds = \int_b (l\delta u + m\delta v)\, ds. \tag{26}$$

On horizontal segments belonging to the arc CDA, however, $\delta v \neq 0$, but we may choose χ so that

$$\tau_1\chi + \lambda_b \frac{\partial f}{\partial v} = -m, \tag{27}$$

thus ensuring that eqn. (26) is satisfied for these segments also. Equations (24) and (27) together determine χ at all points of the arc CDA not belonging to vertical segments and therefore provide appropriate boundary conditions on Γ for eqn. (7).

On vertical segments belonging to the arc DAB where u is specified, $\delta u=0$ and

$$\int_\beta \left[m\delta v-\left(\mu_\beta\frac{\partial g}{\partial u}-\tau_2\psi\right)\delta u\right]ds=\int_\beta m\delta v\,ds=\int_\beta (l\delta u+m\delta v)\,ds. \quad (28)$$

On vertical segments belonging to the arc BCD, however, $\delta u\neq 0$, but we may choose ψ so that

$$\tau_2\psi-\mu_\beta\frac{\partial g}{\partial u}=l, \quad (29)$$

when eqn. (28) is also satisfied for these segments. Equations (23) and (29) together determine ψ at all points of the arc BCD not belonging to horizontal segments, and therefore provide appropriate boundary conditions on Γ for eqn. (6).

Finally, choose the discontinuities of ψ and χ on the projections of vertical and horizontal segments respectively as follows:

$$\psi_l=\psi_r+\mu_\beta\frac{\partial g}{\partial u} \quad \text{on} \quad \beta_0\to\beta_1 \quad (30)$$

and

$$\chi_l=\chi_u+\lambda_b\frac{\partial f}{\partial v} \quad \text{on} \quad b_0\to b_1, \quad (31)$$

so that the last two terms on the left-hand side of eqn. (22) vanish.

Using eqns. (25) to (31), eqn. (22) may now be reduced to:

$$\int_{\tilde{\Gamma}}(l\delta u+m\delta v)\,ds+\sum_b\int_b(l\delta u+m\delta v)\,ds+\sum_\beta\int_\beta(l\delta u+m\delta v)\,ds$$
$$=\sum_b\int_{b_0}^{b_1}\lambda_b\frac{\partial f}{\partial \theta}\delta\theta\,dx+\sum_\beta\int_{\beta_0}^{\beta_1}\mu_\beta\frac{\partial g}{\partial\theta}\delta\theta\,dy+\iint_\Sigma\left(\psi\frac{\partial f}{\partial\theta}+\chi\frac{\partial g}{\partial\theta}\right)\delta\theta\,dx\,dy.$$

But the left-hand side of this is simply:

$$\oint_\Gamma(l\delta u+m\delta v)\,ds \quad \text{or} \quad \delta I,$$

so

$$\delta I=\sum_b\int_{b_0}^{b_1}\lambda_b\frac{\partial f}{\partial \theta}\delta\theta\,dx+\sum_\beta\int_{\beta_0}^{\beta_1}\mu_\beta\frac{\partial g}{\partial\theta}\delta\theta\,dy+\iint_\Sigma\left(\psi\frac{\partial f}{\partial\theta}+\chi\frac{\partial g}{\partial\theta}\right)\delta\theta\,dx\,dy, \quad (32)$$

which is the required expansion of δI to first order in $\delta\theta$. Note that eqns. (23), (24), (27), (29), (30) and (31) provide just the boundary conditions required to determine a solution of eqns. (6) and (7) in each of the subregions $\sum_i$ into which $\sum$ is divided by the projections of horizontal and vertical segments.

§ 4. Discussion

Comparing eqn. (32) with the form it assumes in the absence of finite horizontal and vertical boundary segments, namely

$$\delta I = \int\int_\Sigma \left(\psi \frac{\partial f}{\partial \theta} + \chi \frac{\partial g}{\partial \theta}\right) \delta\theta \, dx \, dy. \tag{33}$$

It is seen to contain additional contributions in the form of line integrals along the singular segments of the boundary, and along the projection of some of these segments into the interior of $\sum$. This means that a perturbation of θ confined to a small horizontal line element, such as PQ or P′Q′ in fig. 2, will make a first-order contribution to δI if the element lies on the projection of a horizontal segment of Γ, as in the position PQ, but will contribute only to higher order if the element does not lie on such a projection, as in the position P′Q′. This is perhaps hardly surprising since disturbances propagate along the characteristics of eqns. (1) and (2). The lines of discontinuity of ψ and χ in the interior of $\sum$ are also a result of the presence of the singular segments in Γ.

Any perturbation $\delta\theta(x,y)$ which has the same sign as $\lambda_b \partial f/\partial\theta$ at all points on $b_0 \to b_2$ (all b), the same sign as $\mu_\beta \partial g/\partial\theta$ at all points on $\beta_0 \to \beta_2$ (all β), and the same sign as $\psi(\partial f/\partial\theta) + \chi(\partial g/\partial\theta)$ at all other points, will increase the value of I, so eqn. (32) provides a means of selecting small changes in θ which will increase the value of the objective function I and permit a 'hill climbing' procedure to be devised for its maximization. The functions $\theta(x,y)$ generated in this way need not, of course, be continuous, and may in general have discontinuities wherever ψ or χ is discontinuous.

Alternatively eqn. (32) yields necessary conditions for a stationary value of I in the form:

$$\lambda_b \frac{\partial f}{\partial \theta} = 0 \quad \text{on} \quad b_0 \to b_2 \text{ (all } b\text{)}, \tag{34}$$

$$\mu_\beta \frac{\partial g}{\partial \theta} = 0 \quad \text{on} \quad \beta_0 \to \beta_2 \text{ (all } \beta\text{)} \tag{35}$$

and

$$\psi \frac{\partial f}{\partial \theta} + \chi \frac{\partial g}{\partial \theta} = 0 \quad \text{(all other points of } \textstyle\sum\text{)}. \tag{36}$$

Once again, a solution of these equations will, in general, have discontinuities wherever ψ and χ are discontinuous.

Acknowledgments

This work was carried out while the author was Visiting Professor at Rice University in the session 1965–66 and was supported by the National Science Foundation grant for Systems Research GU–1153. He would like to acknowledge his debt to his colleagues, Professors F. Horn and Angelo Miele for many valuable discussions of variational problems in general and the present class of problems in particular.

References

Egorov, A. I., 1964, *Automn remote Control,* **25,** No. 5, May 1964.
Jackson, R., 1966, *Trans. Instn chem. Engrs* (in the press).
Leitmann, G., 1962, *Optimisation Techniques* (New York: Academic Press Inc.).

C6

Int. J. Control, 1966, Vol. 4, No. 6, 585–598

Optimization Problems in a Class of Systems Described by Hyperbolic Partial Differential Equations

Part II. A Maximum Principle†

By R. Jackson

University of Edinburgh and Heriot-Watt University

[Received December 21, 1966]

Abstract

Questions of optimum start-up and control of certain types of chemical plant find mathematical expression as variational problems in two independent variables with hyperbolic partial differential equations as side-conditions. It was shown in Part I of this work that such problems have interesting features when the integral to be extremized is taken round a closed curve in the plane of the independent variables, and this curve includes finite straight segments parallel to the characteristics of the differential equations.

In the present paper the first-order variational theory described in Part I will be extended to obtain a result analogous to Pontryagin's maximum principle.

§ 1. Brief Statement of Problem

The problem has been formulated in Part I of this work (Jackson 1966), and the formulation will be reiterated briefly here for convenience.

We are interested in two functions u and v of two independent variables x and y, which satisfy the differential requations:

$$\frac{\partial u}{\partial x} = f(u, v, \theta), \tag{1}$$

$$\frac{\partial v}{\partial y} = g(u, v, \theta), \tag{2}$$

in the region Σ enclosed by the curve Γ, which may include finite straight segments parallel to the coordinate axes as indicated in fig. 1, where segments parallel to the x-axis are denoted by $a, b, c, \ldots$ and segments parallel to the y-axis by $\alpha, \beta, \gamma, \ldots$. No two vertical segments have the same abscissa and no two horizontal segments have the same ordinate. The points of Γ with largest and smallest abscissa are denoted by C and A, and the points with largest and smallest ordinate by D and B respectively.

u is specified at all points of the arc DAB not on horizontal segments, and at the left-hand end-points of these segments, while v is specified at all points of the arc ABC not on vertical segments, and at the lower end points of these segments. Denote the sub-sets of Γ on which u and v are specified by Γ_u and Γ_v respectively. These boundary conditions, together

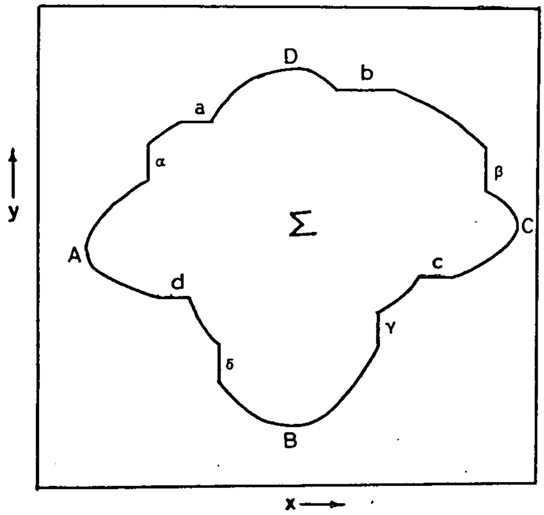
† Communicated by Dr. A. T. Fuller.

with eqns. (1) and (2), determine u and v at all points of Σ and Γ when the function $\theta(x,y)$ is given. The problem is then to find necessary conditions for the function $\theta(x,y)$ to maximize an integral of the form:

$$I = \oint_{\Gamma} (lu+mv)\,ds, \tag{3}$$

where ds is the magnitude of a small displacement along Γ and l and m are given functions of position on Γ.

Fig. 1



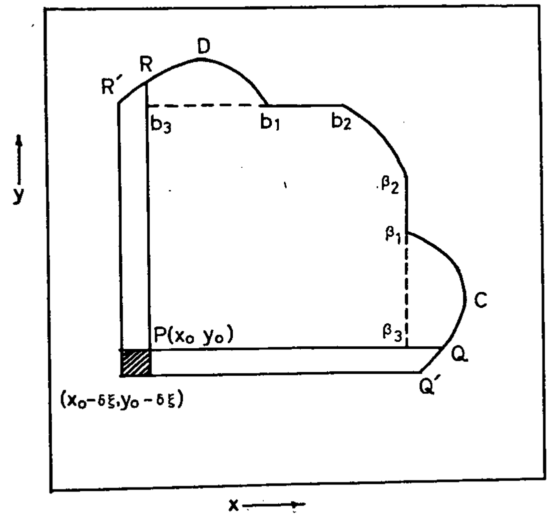
The problem will be approached by considering perturbations of $\theta(x,y)$ which have a small effect on I, not because they are small in magnitude, but because they are localized in a small region of the (x,y)-plane. The course of the argument is influenced by the location of the region in which θ is perturbed in relation to the horizontal and vertical segments of Γ and their projections into Σ, and three types of location are dealt with separately in the three succeeding sections.

§ 2. A Necessary Condition at a General Point of Σ

We shall consider a perturbation of θ which is not necessarily small in magnitude, but is localized in a small neighbourhood of a point P(x_0, y_0) of Σ. The course of the derivation depends to some extent on the location of P, as previously noted, and we first consider the case in which P does not lie on any horizontal or vertical segment of Γ, on the projection of any horizontal segment of the arc BCD into Σ, or on the projection of any vertical segment of the arc CDA into Σ. Such a point will be referred to as a 'general' point of Σ.

Specifically, let θ be changed to $\theta+\Delta\theta$ in the small square region $x_0-\delta\xi \to x_0$, $y_0-\delta\xi \to y_0$, as shown in fig. 2. This induces small but finite changes δu and δv in u and v within a region occupying the first quadrant of a pair of horizontal and vertical axes with the point $(x_0-\delta\xi, y_0-\delta\xi)$ as origin. u and v are unchanged elsewhere, so attention can be confined to the part of Σ lying in this quadrant, as indicated in fig. 2. The horizontal and vertical lines through (x_0, y_0) meet the bounding curve Γ at Q and R respectively, while the corresponding lines through $(x_0-\delta\xi, y_0-\delta\xi)$ meet Γ at Q′ and R′.

Fig. 2



The segment Q→R of Γ will be denoted by $\mathscr{C}$ and will, in general, include a number of finite horizontal and vertical segments. Figure 2 shows just one of each type for simplicity in drawing. Those horizontal segments of $\mathscr{C}$ which also belong to the arc BCD of Γ are produced back to meet either Γ or the line PR at a point which will be denoted by b_3 in the case of the horizontal segment b. As in Part I of this work, the end-points of the horizontal segment b will be denoted by b_1 and b_2, and this notation is illustrated in fig. 2. (If the produced segment meets Γ rather than PR, b_3 is, of course, identical with the point b_0 introduced in Part I.) Similarly the vertical segments lying in $\mathscr{C}$ and the arc CDA of Γ are produced back to meet either Γ or PQ in points such as β_3. Those parts of the arc $\mathscr{C}$ which belong neither to horizontal nor vertical segments will be denoted by $\tilde{\mathscr{C}}$, and the region enclosed by $\mathscr{C}$ and the straight lines PQ and PR will be denoted by $\mathscr{S}$.

When θ is perturbed as described the integrand of (3) is changed only on the arc $\mathscr{C}$ and the small arcs QQ′ and RR′, so:

$$\delta I = \delta I_1 + \delta I_2 + \delta I_3, \tag{4}$$

where

$$\delta I_1 = \int_{\mathscr{C}} (l\delta u + m\delta v)\, ds, \tag{5}$$

$$\delta I_2 = \int_{Q'}^{Q} (l\delta u + m\delta v)\, ds \tag{6}$$

and

$$\delta I_3 = \int_{R}^{R'} (l\delta u + m\delta v)\, ds. \tag{7}$$

δI will be evaluated in three stages; firstly, δI_1 will be related to changes in u and v along the lines PQ and PR; secondly, the contributions from PQ and PR will be related to variations in u and v at point P; and, thirdly, the variations in u and v at P will be related to $\Delta\theta$ and the contributions δI_2 and δI_3 will be added.

Stage 1

As in Part I of this work, variables λ_b and μ_β are introduced, associated with the horizontal and vertical segments respectively and satisfying:

$$\frac{d\lambda_b}{dx} = -\lambda_b \frac{\partial f}{\partial u} - l_b(x) \tag{8}$$

with

$$\lambda_b(b_2) = 0 \tag{9}$$

and

$$\begin{aligned} l_b(x) &= l(x) \quad \text{on} \quad b_1 \to b_2 \\ &= 0 \quad \text{on} \quad b_3 \to b_1 \end{aligned} \tag{10}$$

together with

$$\frac{d\mu_\beta}{dy} = -\mu_\beta \frac{\partial g}{\partial v} - m_\beta(y) \tag{11}$$

with

$$\mu_\beta(\beta_2) = 0 \tag{12}$$

and

$$\begin{aligned} m_\beta(y) &= m(y) \quad \text{on} \quad \beta_1 \to \beta_2 \\ &= 0 \quad \text{on} \quad \beta_3 \to \beta_1. \end{aligned} \tag{13}$$

We also consider the linearized form of eqns. (1) and (2) relating small changes in u and v, namely:

$$\frac{\partial}{\partial x}(\delta u) = \frac{\partial f}{\partial u}\delta u + \frac{\partial f}{\partial v}\delta v \tag{14}$$

and

$$\frac{\partial}{\partial y}(\delta v) = \frac{\partial g}{\partial u}\delta u + \frac{\partial g}{\partial v}\delta v. \tag{15}$$

Terms in $\delta\theta$ do not appear in these equations, since θ remains unchanged throughout the region considered.

From (8) and (14):

$$\frac{d}{dx}(\lambda_b\, \delta u) + l_b\, \delta u = \lambda_b \frac{\partial f}{\partial v}\delta v,$$

and integrating between the limits b_3 and b_2:

$$|\lambda_b \delta u|_{b_3}^{b_2} + \int_{b_3}^{b_2} l_b \delta u \, dx = \int_{b_3}^{b_2} \lambda_b \frac{\partial f}{\partial v} \delta v \, dx.$$

Since $\lambda_b(b_2) = 0$ and $l_b = 0$ on $b_3 \to b_1$, this gives:

$$\int_{b_1}^{b_2} l\delta u \, dx = \int_{b_3}^{b_2} \lambda_b \frac{\partial f}{\partial v} \delta v \, dx + (\lambda_b \delta u)_{b_3}. \tag{16}$$

Similarly, for a vertical segment β:

$$\int_{\beta_1}^{\beta_2} m\delta v \, dy = \int_{\beta_3}^{\beta_2} \mu_\beta \frac{\partial g}{\partial u} \delta u \, dy + (\mu_\beta \delta v)_{\beta_3}. \tag{17}$$

Now $\mathscr{S}$ is divided into sub-regions $\mathscr{S}_k$ by the projections of horizontal and vertical segments of $\mathscr{C}$. As in Part I, introduce variables ψ and χ in each sub-region, satisfying the differential equations:

$$\frac{\partial \psi}{\partial x} = -\psi \frac{\partial f}{\partial u} - \chi \frac{\partial g}{\partial u}, \tag{18}$$

$$\frac{\partial \chi}{\partial y} = -\psi \frac{\partial f}{\partial v} - \chi \frac{\partial g}{\partial v}. \tag{19}$$

Then if $\mathscr{C}_k$ is the (closed) boundary of the sub-region $\mathscr{S}_k$, it follows from Gauss's theorem, using eqns. (14), (15), (18) and (19) that

$$\oint_{\mathscr{C}_k} (\tau_2 \psi \delta u - \tau_1 \chi \delta v) \, ds = 0,$$

where $\boldsymbol{\tau} = (\tau_1, \tau_2)$ is the unit tangent vector to $\mathscr{C}_k$. (The argument is the same as that leading to eqn. (8) in Part I.) Adding these equations for all sub-regions then gives:

$$\left(\int_{\mathrm{P}}^{\mathrm{Q}} + \int_{\mathscr{C}} + \int_{\mathrm{R}}^{\mathrm{P}}\right)(\tau_2 \psi \delta u - \tau_1 \chi \delta v) \, ds + \sum_b \int_{b_3}^{b_1} (\chi_l - \chi_u) \delta v \, dx$$
$$+ \sum_\beta \int_{\beta_3}^{\beta_1} (\psi_1 - \psi_r) \delta u \, dy = 0, \tag{20}$$

where χ_l and χ_u are the values of χ on approaching the projection of b from below and above respectively, while ψ_1 and ψ_r are the values of ψ on approaching the projection of β from the left and right respectively.

Now in general $\tau_1 ds = dx$ and $\tau_2 ds = dy$, while on PQ $\tau_1 = 1$ and $\tau_2 = 0$, and on RP $\tau_1 = 0$ and $\tau_2 = -1$. Thus (20) may alternatively be written:

$$\int_{\mathscr{C}} (\tau_2 \psi \delta u - \tau_1 \chi \delta v) \, ds + \sum_b \int_{b_3}^{b_1} (\chi_l - \chi_u) \delta v \, dx + \sum_\beta \int_{\beta_3}^{\beta_1} (\psi_1 - \psi_r) \delta u \, dy$$
$$= \int_{\mathrm{P}}^{\mathrm{Q}} \chi \delta v \, dx + \int_{\mathrm{P}}^{\mathrm{R}} \psi \delta u \, dy. \tag{21}$$

Equations (16) and (17) may be re-arranged in the form:

$$\int_{b_1}^{b_2}\left(l\delta u-\lambda_b\frac{\partial f}{\partial v}\delta v\right)dx=\int_{b_3}^{b_1}\lambda_b\frac{\partial f}{\partial v}\delta v\,dx+(\lambda_b\,\delta u)_{b_3}, \tag{16'}$$

and

$$\int_{\beta_1}^{\beta_2}\left(m\delta v-\mu_\beta\frac{\partial g}{\partial u}\delta u\right)dy=\int_{\beta_3}^{\beta_1}\mu_\beta\frac{\partial g}{\partial u}\delta u\,dy+(\mu_\beta\,\delta v)_{\beta_3}. \tag{17'}$$

Adding these for all values of b and β to eqn. (21) and separating the first integral on the left-hand side of (21) into contributions from the horizontal and vertical segments and from $\tilde{\mathscr{C}}$, we obtain:

$$\begin{aligned}\int_{\tilde{\mathscr{C}}}(\tau_2\psi\delta u-\tau_1\chi\delta v)\,ds&+\sum_b\int_{b_1}^{b_2}\left[l\delta u-\left(\lambda_b\frac{\partial f}{\partial v}+\tau_1\chi\right)\delta v\right]dx\\&+\sum_\beta\int_{\beta_1}^{\beta_2}\left[m\delta v-\left(\mu_\beta\frac{\partial g}{\partial u}-\tau_2\psi\right)\delta u\right]dy\\&+\sum_b\int_{b_3}^{b_1}\left(\chi_l-\chi_u-\lambda_b\frac{\partial f}{\partial v}\right)\delta v\,dx\\&+\sum_\beta\int_{\beta_3}^{\beta_1}\left(\psi_{\mathrm{l}}-\psi_{\mathrm{r}}-\mu_\beta\frac{\partial g}{\partial u}\right)\delta u\,dy\\=\int_{\mathrm{P}}^{\mathrm{Q}}\chi\delta v\,dx+\int_{\mathrm{P}}^{\mathrm{R}}\psi\delta u\,dy&+\sum_b(\lambda_b\,\delta u)_{b_3}+\sum_\beta(\mu_\beta\,\delta v)_{\beta_3}.\end{aligned} \tag{22}$$

Now let ψ and χ be required to satisfy the boundary conditions derived in Part I of this work, namely:

$$\tau_2\psi=l \tag{23}$$

at points common to $\tilde{\Gamma}$ and $\Gamma-\Gamma_u$, where $\tilde{\Gamma}$ is that part of Γ which belongs neither to horizontal nor vertical segments:

$$-\tau_1\chi=m \tag{24}$$

at points common to $\tilde{\Gamma}$ and $\Gamma-\Gamma_v$:

$$\tau_2\psi-\mu_\beta\frac{\partial g}{\partial u}=l \tag{25}$$

on vertical segments belonging to the arc BCD, and

$$\tau_1\chi+\lambda_b\frac{\partial f}{\partial v}=-m \tag{26}$$

on horizontal segments belong to the arc CDA.

The discontinuities of ψ and χ on the projections of vertical and horizontal segments are also chosen so that

$$\psi_{\mathrm{l}}=\psi_{\mathrm{r}}+\mu_\beta\frac{\partial g}{\partial u} \tag{27}$$

and

$$\chi_l=\chi_u+\lambda_b\frac{\partial f}{\partial v}. \tag{28}$$

Conditions (23) to (28), together with eqns. (18) and (19), suffice to determine ψ and χ at all points of Σ and Γ when $\theta(x,y)$ is given.

Using these conditions eqn. (22) is considerably simplified to:

$$\int_{\mathscr{C}} (l\delta u + m\delta v)\,ds = \delta I_1 = \int_{\mathrm{P}}^{\mathrm{Q}} \chi\delta v\,dx + \int_{\mathrm{P}}^{\mathrm{R}} \psi\delta u\,dy + \sum_{b} (\lambda_b\,\delta u)_{b_3} + \sum_{\beta} (\mu_\beta\,\delta v)_{\beta_3}. \tag{29}$$

This relates δI_1 to changes in u and v on the lines PQ and PR and therefore completes Stage 1 of the argument.

Stage 2

It is now required to find δv on PQ and δu on PR for use in eqn. (29). In view of the localized nature of the perturbation in θ:

$$\delta v(x, y_0 - \delta\xi) = 0 \quad \text{for} \quad x > x_0, \tag{30}$$

so a Taylor expansion gives:

$$\delta v(x, y_0) = \left[\frac{\partial}{\partial y}(\delta v)\right]_{x,y_0-\delta\xi} \delta\xi + O(\delta\xi^2). \tag{31}$$

From eqns. (15) and (30)

$$\left[\frac{\partial}{\partial y}(\delta v)\right]_{x,y_0-\delta\xi} = \left[\frac{\partial g}{\partial u}\delta u\right]_{x,y_0-\delta\xi} = \left[\frac{\partial g}{\partial u}\delta u\right]_{x,y_0} + O(\delta\xi)$$

and using this in the right-hand side of eqn. (31) gives:

$$\delta v(x, y_0) = \delta\xi\left[\frac{\partial g}{\partial u}\delta u\right]_{x,y_0} + O(\delta\xi^2). \tag{32}$$

With this value of $\delta v(x, y_0)$ we then have:

$$\int_{\mathrm{P}}^{\mathrm{Q}} \chi\delta v\,dx = \delta\xi\int_{\mathrm{P}}^{\mathrm{Q}} \chi\frac{\partial g}{\partial u}\delta u\,dx + O(\delta\xi^2). \tag{33}$$

At all points of PQ, other than the points $\alpha_3, \beta_3, \gamma_3, \ldots, \nu_3$, where it is intersected by the projections of vertical segments of Γ, ψ satisfies eqn. (18). Thus, from eqns. (14) and (18):

$$\frac{\partial}{\partial x}(\psi\delta u) = \psi\left(\frac{\partial f}{\partial u}\delta u + \frac{\partial f}{\partial v}\delta v\right) - \delta u\left(\psi\frac{\partial f}{\partial u} + \chi\frac{\partial g}{\partial u}\right) = \psi\frac{\partial f}{\partial v}\delta v - \chi\frac{\partial g}{\partial u}\delta u. \tag{34}$$

But from (32):

$$\delta v = \delta u O(\delta\xi) + O(\delta\xi^2)$$

at all points of PQ, so the first term on the right-hand side of (34) is an order smaller than the second, and we may write:

$$\frac{\partial}{\partial x}(\psi\delta u) = -\chi\frac{\partial g}{\partial u}\delta u + O(\delta\xi), \tag{35}$$

which is valid at all points of PQ except $\alpha_3, \beta_3, \gamma_3, \ldots, \nu_3$.

The value of $\chi(\partial g/\partial u)\,\delta u$ given by eqn. (35) may not be substituted directly in the integrand of the right-hand side of eqn. (33), since ψ is discontinuous at the points $\alpha_3, \beta_3, \gamma_3, \ldots, \nu_3$ and the left-hand side of (35) does not take a finite value. However, we may write:

$$\int_{\mathrm{P}}^{\mathrm{Q}} \chi \frac{\partial g}{\partial u} \delta u\, dx = \left(\int_{\mathrm{P}}^{\alpha_3} + \int_{\alpha_3}^{\beta_3} + \int_{\beta_3}^{\gamma_3} + \ldots + \int_{\nu_3}^{\mathrm{Q}}\right) \chi \frac{\partial g}{\partial u} \delta u\, dx.$$

Equation (35) may then be used in each separate integral so that eqn. (33) becomes:

$$\begin{aligned}\int_{\mathrm{P}}^{\mathrm{Q}} \chi \delta v\, dx &= -\delta\xi\left(\int_{\mathrm{P}}^{\alpha_3} + \int_{\alpha_3}^{\beta_3} + \int_{\beta_3}^{\gamma_3} + \ldots + \int_{\nu_3}^{\mathrm{Q}}\right) \chi \frac{\partial g}{\partial u} \delta u\, dx + O(\delta\xi^2) \\ &= -\delta\xi[(\psi_{\mathrm{l}}\,\delta u)_{\alpha_3} - (\psi\delta u)_{\mathrm{P}} + (\psi_{\mathrm{l}}\,\delta u)_{\beta_3} - (\psi_{\mathrm{r}}\,\delta u)_{\alpha_3} + \ldots \\ &\qquad + (\psi\delta u)_{\mathrm{Q}} - (\psi_{\mathrm{r}}\,\delta u)_{\nu_3}] + O(\delta\xi^2)\end{aligned}$$

or

$$\int_{\mathrm{P}}^{\mathrm{Q}} \chi\delta v\, dx = \delta\xi\left\{(\psi\delta u)_{\mathrm{P}} - (\psi\delta u)_{\mathrm{Q}} - \sum_{\beta} [(\psi_{\mathrm{l}} - \psi_{\mathrm{r}})\,\delta u]_{\beta_3}\right\} + O(\delta\xi^2), \tag{36}$$

and by similar reasoning, using an equation analogous to (32), namely:

$$\delta u(x_0, y) = \delta\xi \left[\frac{\partial f}{\partial v}\delta v\right]_{x_0, y} + O(\delta\xi^2), \tag{37}$$

it can be shown that

$$\int_{\mathrm{P}}^{\mathrm{R}} \psi\delta u\, dy = \delta\xi\left\{(\chi\delta v)_{\mathrm{P}} - (\chi\delta v)_{\mathrm{R}} - \sum_{b} [(\chi_l - \chi_u)\,\delta v]_{b_3}\right\} + O(\delta\xi^2). \tag{38}$$

Equations (36) and (38) may now be used to determine the integrals on the right-hand side of eqn. (29), and the values of δu and δv in the sums on the right-hand side of this equation may be obtained from eqns. (32) and (37), with the result:

$$\begin{aligned}\delta I_1 = \delta\xi\Bigg\{&(\psi\delta u)_{\mathrm{P}} - (\psi\delta u)_{\mathrm{Q}} - \sum_{\beta}\left[\left(\psi_{\mathrm{l}} - \psi_{\mathrm{r}} - \mu_\beta \frac{\partial g}{\partial u}\right)\delta u\right]_{\beta_3} \\ &+ (\chi\delta v)_{\mathrm{P}} - (\chi\delta v)_{\mathrm{R}} - \sum_{b}\left[\left(\chi_l - \chi_u - \lambda_b \frac{\partial f}{\partial v}\right)\delta v\right]_{b_3}\Bigg\} + O(\delta\xi^2).\end{aligned}$$

Conditions (27) and (28) then show that each term in the sums over b and β vanishes separately, so this reduces to:

$$\delta I_1 = \delta\xi[(\psi\delta u)_{\mathrm{P}} - (\psi\delta u)_{\mathrm{Q}} + (\chi\delta v)_{\mathrm{P}} - (\chi\delta v)_{\mathrm{R}}] + O(\delta\xi^2), \tag{39}$$

and this completes Stage 2 of the argument. From now on we shall drop the terms of $O(\delta\xi^2)$ for simplicity in writing.

Stage 3

Finally, it is necessary to add δI_2 and δI_3 to δI_1, and to express δu_{P} and δv_{P} in terms of $\Delta\theta$.

From boundary conditions (23) and (24), $\psi = l/\tau_2$ at Q and $\chi = -m/\tau_1$ at R, so:

$$\delta\xi[(\psi\delta u)_{\mathrm{Q}} + (\chi\delta v)_{\mathrm{R}}] = [(l\delta u)\,\delta\xi/\tau_2]_{\mathrm{Q}} - [(m\delta v)\,\delta\xi/\tau_1]_{\mathrm{R}}. \tag{40}$$

Now if δs_Q is the length of the small displacement Q′Q and by δy_Q its vertical component, we have:

$$\delta y_Q = \tau_2 \delta s_Q.$$

But $\delta y_Q = \delta\xi$ since Γ is traversed in the direction of increasing y at Q, so this becomes:

$$\delta\xi/\tau_2 = \delta s_Q. \tag{41}$$

Similarly, if δs_R is the length of the small displacement RR′ and δx_R its horizontal component, we have:

$$\delta x_R = \tau_1 \delta s_R.$$

But $\delta x_R = -\delta\xi$, since Γ is traversed in the direction of decreasing x at R, so this becomes:

$$-\delta\xi/\tau_1 = \delta s_R. \tag{42}$$

Using (41) and (42) eqn. (40) becomes:

$$\delta\xi[(\psi\delta u)_Q + (\chi\delta v)_R] = (l\delta u\delta s)_Q + (m\delta v\delta s)_R \tag{43}$$

and using this, eqn. (39) becomes:

$$\delta I_1 + (l\delta u\delta s)_Q + (m\delta v\delta s)_R = [(\psi\delta u)_P + (\chi\delta v)_P]\,\delta\xi, \tag{44}$$

omitting terms of $O(\delta\xi^2)$ on the right-hand side.

Since Q′Q is small, we may write:

$$\delta I_2 = \int_{Q'}^{Q} (l\delta u + m\delta v)\,ds = (l\delta u + m\delta v)_Q\,\delta s + O(\delta\xi^2).$$

But according to eqn. (32):

$$\delta v = \delta u O(\delta\xi) + O(\delta\xi^2) \quad \text{on PQ},$$

so the above becomes:

$$\delta I_2 = (l\delta u\delta s)_Q + O(\delta\xi^2). \tag{45}$$

Similarly, using eqn. (37), it follows that

$$\delta I_3 = (m\delta v\delta s)_R + O(\delta\xi^2), \tag{46}$$

and using (45) and (46) in eqn. (44):

$$\delta I = \delta I_1 + \delta I_2 + \delta I_3 = \delta\xi(\psi\delta u + \chi\delta v)_P, \tag{47}$$

omitting terms of $O(\delta\xi^2)$.

From differential eqns. (1) and (2) for u and v, the changes δu_P and δv_P resulting from the perturbation $\Delta\theta$ can be found. It follows from eqn. (1) that

$$\delta u(x_0, y_0) = \delta u(x_0 - \delta\xi, y_0) + \delta\xi[f(u, v, \theta + \Delta\theta) - f(u, v, \theta)]_{x_0, y_0} + O(\delta\xi^2),$$

and since u is unchanged at $(x_0 - \delta\xi, y_0)$, the first term on the right-hand side of this vanishes and we may write:

$$\delta u(x_0, y_0) = \delta\xi(\Delta f)_{x_0, y_0} + O(\delta\xi^2), \tag{48}$$

where

$$(\Delta f)_{x_0,y_0} = [f(u,v,\theta+\Delta\theta)-f(u,v,\theta)]_{x_0,y_0}. \tag{49}$$

Similarly, using eqn. (2), it can be shown that

$$\delta v(x_0,y_0) = \delta\xi(\Delta g)_{x_0,y_0}+O(\delta\xi^2), \tag{50}$$

where

$$(\Delta g)_{x_0,y_0} = [g(u,v,\theta+\Delta\theta)-g(u,v,\theta)]_{x_0,y_0}. \tag{51}$$

Using (48) and (50) in eqn. (47) then gives:

$$\left.\begin{aligned} \delta I &= \delta\xi^2[\psi\Delta f+\chi\Delta g]_{x_0,y_0} \\ \text{or}\qquad \delta I &= \delta\xi^2[\Delta(\psi f+\chi g)]_{x_0,y_0}, \end{aligned}\right\} \tag{52}$$

omitting terms of higher order in $\delta\xi$.

If I is to be maximized by the function $\theta(x,y)$, it is necessary that $\delta I \leqslant 0$ for all $\Delta\theta$ and hence, from eqn. (52), it is necessary that

$$\Delta(\psi f+\chi g) \leqslant 0 \tag{53}$$

for all $\Delta\theta$ at any 'general' point (x_0,y_0). Note that $\Delta\theta$ is not necessarily small, so that $\theta' = \theta+\Delta\theta$ may be *any* other permissible control function. It therefore follows that $\theta(x,y)$ must be chosen at each point so that

$$H(\theta) = \psi f+\chi g \tag{54}$$

takes its largest possible value, regarded as a function of θ, for fixed values of ψ, χ, u and v. This is a result analogous to Pontryagin's maximum principle for extremal problems with ordinary differential equations as side-conditions, and provides a necessary condition for the maximization of I stronger than that obtained from the theory of the first variation in Part I of this work.

§ 3. A Necessary Condition at a Point Lying on the Projection of a Horizontal or Vertical Segment of Γ

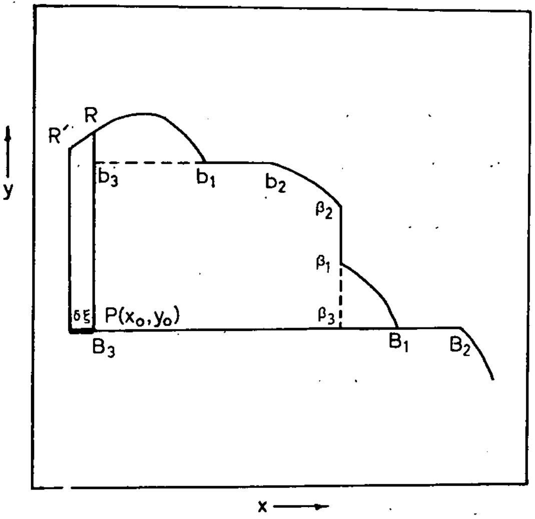
Figure 3 illustrates the case in which P lies on the projection of a horizontal segment B_2B_1 of Γ. In the notation previously used, P then coincides with B_3.

In place of the square used in the previous discussion, the perturbation of θ is now confined to a line segment $(x_0-\delta\xi,y_0)\rightarrow(x_0,y_0)$, as indicated in fig. 3. The result will be a perturbation of the integrand of I confined to the arc $B_2B_1\beta_1\beta_2b_2b_1RR'$ of Γ, as drawn in fig. 3. Only one vertical and one horizontal segment are indicated between B_1 and R but in general, of course, there may be a number of segments of each type. As before, the arc $B_2\rightarrow R$ will be denoted by $\mathscr{C}$ and the sub-set of points on this arc belonging neither to horizontal nor vertical segments will be denoted by $\tilde{\mathscr{C}}$.

Stage 1 of the argument for a general point P may be taken over unchanged in the present case and leads to an equation analogous to (22), except that Q is replaced by B_1 and it is important to remember that the value χ_u must be taken for χ on PB_1, since this is a line of discontinuity of χ. Thus eqn. (22) becomes:

$$\begin{aligned}\int_{\tilde{\mathscr{C}}}(\tau_2\psi\delta u-\tau_1\chi\delta v)\,ds+\sum_b\int_{b_1}^{b_2}\left[l\delta u-\left(\lambda_b\frac{\partial f}{\partial v}+\tau_1\chi\right)\delta v\right]dx\\ +\sum_\beta\int_{\beta_1}^{\beta_2}\left[m\delta v-\left(\mu_\beta\frac{\partial g}{\partial u}-\tau_2\psi\right)\delta u\right]dy\\ +\sum_b\int_{b_3}^{b_1}\left(\chi_l-\chi_u-\lambda_b\frac{\partial f}{\partial v}\right)\delta v\,dx\\ +\sum_\beta\int_{\beta_3}^{\beta_1}\left(\psi_\mathrm{l}-\psi_\mathrm{r}-\mu_\beta\frac{\partial g}{\partial u}\right)\delta u\,dy\\ =\int_\mathrm{P}^{\mathrm{B}_1}\chi_u\,\delta v\,dx+\int_\mathrm{P}^\mathrm{R}\psi\delta u\,dy+\sum_b(\lambda_b\,\delta u)_{b_3}+\sum_\beta(\mu_\beta\,\delta v)_{\beta_3}.\end{aligned}\tag{55}$$

Fig. 3



Now add to eqn. (55) the equation of type (16′) associated with the horizontal segment B, namely:

$$\int_{\mathrm{B}_1}^{\mathrm{B}_2}\left(l\delta u-\lambda_\mathrm{B}\frac{\partial f}{\partial v}\delta v\right)dx=\int_{\mathrm{B}_3}^{\mathrm{B}_1}\lambda_\mathrm{B}\frac{\partial f}{\partial v}\delta v\,dx+(\lambda_\mathrm{B}\,\delta u)_{\mathrm{B}_3}$$

and, at the same time, use the boundary conditions (23) to (28). The result is:

$$\begin{aligned}\int_{\tilde{\mathscr{C}}}(l\delta u+m\delta v)\,ds+\sum_b\int_{b_1}^{b_2}(l\delta u+m\delta v)\,dx\\ +\sum_\beta\int_{\beta_1}^{\beta_2}(l\delta u+m\delta v)\,dy+\int_{\mathrm{B}_1}^{\mathrm{B}_2}\left(l\delta u-\lambda_\mathrm{B}\frac{\partial f}{\partial v}\delta v\right)dx\\ =\int_\mathrm{P}^{\mathrm{B}_1}\left(\chi_u+\lambda_\mathrm{B}\frac{\partial f}{\partial v}\right)\delta v\,dx+\int_\mathrm{P}^\mathrm{R}\psi\delta u\,dy+\sum_b(\lambda_b\,\delta u)_{b_3}+\sum_\beta(\mu_\beta\,\delta v)_{\beta_3}+(\lambda_\mathrm{B}\,\delta u)_\mathrm{P}.\end{aligned}\tag{56}$$

Now since the perturbation in θ is confined to a line element collinear with $B_3B_1B_2$, $\delta v = 0$ at all points of $B_3B_1B_2$, and

$$\int_{B_1}^{B_2}\left(l\delta u - \lambda_B \frac{\partial f}{\partial v}\delta v\right)dx = \int_{B_1}^{B_2} l\delta u\, dx = \int_{B_1}^{B_2}(l\delta u + m\delta v)\,dx.$$

Thus the left-hand side of (56) is simply δI_1 the contributions to δI from the arc $B_2 \rightarrow R$ of Γ, and the first and last terms on the right-hand side vanish, giving:

$$\delta I_1 = \int_P^R \psi\delta u\, dy + \sum_b (\lambda_b\, \delta u)_{b_3} + (\lambda_B\, \delta u)_P. \tag{57}$$

This completes the analogue of stage 1 in the previous discussion.

Stage 2 of the previous discussion may also be used virtually unchanged to show that

$$\int_P^R \psi\delta u\, dy + \sum_b (\lambda_b\, \delta u)_{b_3} = \delta\xi[(\chi\delta v)_P - (\chi\delta v)_R] + O(\delta\xi^2),$$

so eqn. (57) becomes:

$$\delta I_1 + \delta\xi(\chi\delta v)_R = \delta\xi(\chi\delta v)_P + (\lambda_B\, \delta u)_P + O(\delta\xi^2). \tag{58}$$

The first term on the right-hand side of this vanishes, since $\delta v = 0$ at P. Furthermore, the reasoning which led to eqns. (43) and (46) remains valid, and in the present case shows that:

$$\delta\xi(\chi\delta v)_R = (m\delta v\delta s)_R = \delta I_3 + O(\delta\xi^2),$$

so (58) becomes:

$$\delta I = \delta I_1 + \delta I_3 = (\lambda_B\, \delta u)_P,$$

neglecting terms of higher order in $\delta\xi$.

But δu_P is given by eqns. (48) and (49) as before, so

$$\delta I = \delta\xi(\lambda_B\, \Delta f)_{x_0, y_0} = \delta\xi[\Delta(\lambda_B f)]_{x_0, y_0}, \tag{59}$$

neglecting terms of higher order in $\delta\xi$.

If I is to be maximized by the function $\theta(x, y)$, it is necessary that $\delta I \leqslant 0$ for all $\Delta\theta$ and hence, from eqn. (59), it is necessary that

$$\Delta(\lambda_B f) \leqslant 0$$

for all $\Delta\theta$. Since $\Delta\theta$ is not necessarily small it follows that $\theta(x, y)$ must be chosen at each point on B_3B_1 so that

$$H_B(\theta) = \lambda_B f \tag{60}$$

takes its largest possible value, regarded as a function of θ, for fixed values of λ_B, u and v. Once again this is a result analogous to Pontryagin's maximum principle, but the Hamiltonian H_B to be maximized along the projection of a horizontal segment differs from the Hamiltonian H defined by eqn. (54), which is to be maximized at points of Σ not lying on projections of horizontal or vertical segments of Γ.

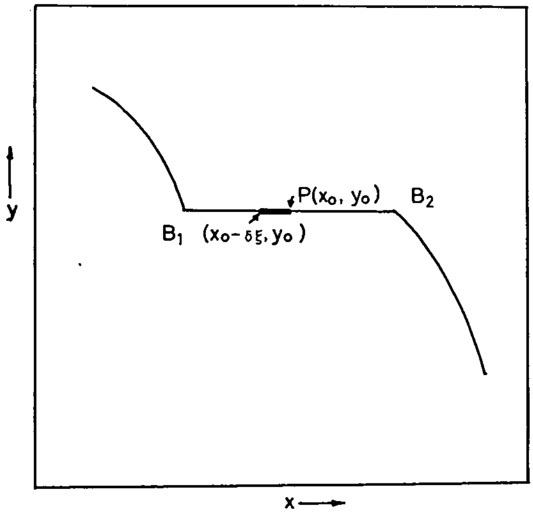
Applying the same reasoning to points lying on the projections of vertical segments, it is found that θ must be chosen at such points to maximize the Hamiltonian:

$$H_\beta(\theta) = \mu_\beta g. \tag{61}$$

§ 4. A Necessary Condition at a Point Lying on a Horizontal or Vertical Segment of Γ

Finally, it is necessary to consider the case in which P actually lies on one of the horizontal or vertical segments of Γ, as indicated in fig. 4 for the case of a horizontal segment. Then a perturbation of θ in the small interval $(x_0 - \delta\xi, y_0) \rightarrow (x_0, y_0)$ only affects the contribution to I from the segment PB_2 of Γ.

Fig. 4



On PB_2, δu satisfies:

$$\frac{d}{dx}(\delta u) = \frac{\partial f}{\partial u}\delta u$$

since $\delta v = 0$, and using this and eqn. (8), it follows that

$$\frac{d}{dx}(\lambda_B \delta u) = -l\delta u,$$

whence, integrating between P and B_2:

$$(\lambda_B \delta u)_P = \int_P^{B_2} l\delta u\, dx = \delta I.$$

δu_P is still given by eqn. (48), so it follows that

$$\delta I = \delta\xi(\lambda_B \Delta f)_{x_0 y_0} + O(\delta\xi^2)$$

or

$$\delta I = \delta\xi[\Delta(\lambda_B f)]_{x_0, y_0}, \tag{62}$$

neglecting terms of higher order in $\delta\xi$.

This is identical with eqn. (59), and it follows in the same way as before that θ must be chosen so that

$$H_{\mathrm{B}} = \lambda_{\mathrm{B}} f$$

is maximized. Similarly H_β, given by eqn. (61), must be maximized at all points of vertical segments of Γ.

§ 5. Conclusions

It has been shown that a necessary condition for θ to maximize I can be framed as the requirement that θ should be chosen at each point to maximize a certain function of θ, which may be constructed by integrating the given differential equations and other differential equations adjoint to them. To this extent the result resembles Pontryagin's maximum principle, valid for maximization problems with ordinary differential equations as side-conditions.

However, the Hamiltonian function to be maximized must be constructed in different ways, depending on whether or not the point considered lies on a horizontal or vertical segment of the bounding curve or the projection of such a segment into the interior of the region of interest. The adjoint variables must also have discontinuities of specified magnitude on crossing such lines.

Acknowledgments

This work was carried out while the author was Visiting Professor at Rice University in the session 1965–66 and was supported by the National Science Foundation grant for Systems Research GU–1153.

Reference

Jackson, R., 1966, *Int. J. Control,* 4, 127.

C7

The optimal use of mixed catalysts for two successive chemical reactions

R. Jackson
(University of Edinburgh)

Summary

When the conversion of a feedstock to a product takes place in two chemically distinct steps, each of which is promoted by a different catalyst, Gunn and Thomas recently showed that there are advantages to be gained by mixing the catalysts in a single reactor rather than carrying out the two reaction steps separately. In this paper the Maximum Principle is applied to the problem of determining the optimal variation in catalyst blend along the reactor, and for a simple first order kinetic scheme it is shown to lead to a complete solution in closed form.

Introduction.

Chemical reactions of economic importance often take place in several steps through a number of intermediate products, and each stage may be catalysed by a different cataly~~s~~tic substance. The simplest example of this is the pair of successive reactions

$$A \xrightarrow{1} B \xrightarrow{2} C \qquad \text{(i)}$$

which it is common practice to carry out in two physically separate reactors, the first to convert A to B and the second to convert B to the final product C. If both reactions are catalytic, the first reactor will then contain only the catalyst for reaction 1, while the second will contain the catalyst for reaction 2.

However, Gunn and Thomas[1] have pointed out that, in certain circumstances, there are advantages in mixing the two catalysts in a single reaction vessel. They quote the example of the reaction scheme

$$A \underset{2}{\overset{1}{\rightleftharpoons}} B \xrightarrow{3} C \qquad \text{(ii)}$$

which differs from the previous one only in that the first reaction is reversible. Then if the reactions are carried out in the conventional way, in two separately catalysed reactors, the yield of C is limited by the amount of B which can be produced in the first reactor, which is in turn limited by the equilibrium condition for the reversible reaction. If the two catalysts are mixed in a single vessel, however, B is removed by conversion to C and the equilibrium restriction is removed, permitting substantially higher yields to be achieved.

Gunn/

Gunn and Thomas limited their investigation to the use of a uniform catalyst mixture in a single isothermal tubular reactor, and were able to show that there is an optimum catalyst blend which gives the highest yield of C for a given length of reactor. They also remarked that it is clear on physical grounds that further improvements could be obtained by varying the catalyst blend along the reactor, but did not pursue this point further.

It is the purpose of the present paper to determine the optimum catalyst blend as a function of position in the reactor. A complete solution of this problem in explicit terms can be found using Pontryagin's Maximum Principle$^{(2)}$, and has the interesting property of using only a finite number of segments in the reactor, each containing uniformly blended catalyst. Thus the solution is of a pure "switching" type, but it is not a "bang-bang" solution making use only of the two pure catalysts.

Mathematical statement of the problem

Referring to reaction scheme (ii) we shall use x and y to denote the mole fractions of substances A and B in the mixture and will assume that all the reactions are of the first order and are carried out in an isothermal tubular reactor. The differential equations describing the variation of composition with distance along the reactor are then

$$\frac{dx}{dt} = f(k_2 y - k_1 x) \qquad (1)$$

$$\frac{dy}{dt} = f(k_1 x - k_2 y) - (1-f) k_3 y \qquad (2)$$

where/

where t denotes residence time from the instant of entry to the reactor. k_1 and k_2 are the velocity constants of reactions 1 and 2 respectively in a reactor where the catalyst consists entirely of the substance which catalyses the reactions $A \rightleftharpoons B$ while k_3 is the velocity constant of reaction 3 in a reactor where the catalyst consists entirely of the substance which catalyses the reaction $B \rightarrow C$. f, which we shall refer to as the catalyst blend, denotes the fraction of the catalyst formed by the substance which catalyses the reactions $A \rightleftharpoons B$ and this fraction can be varied as required along the reactor by suitable mixing of the two catalysts. Where the blend has the value f the effective velocity constants are fk_1, fk_2 and $(1-f)k_3$ as indicated in equations (1) and (2).

The feed will be assumed to consist of pure substance A, so the initial conditions are

$$x(0) = 1 \quad ; \quad y(0) = 0 \tag{3}$$

and we shall consider the problem of determining $f(t)$ subject to the physically necessary constraints

$$0 \leq f \leq 1 \tag{4}$$

so as to maximise the mole fraction of substance C present in the mixture at the reactor exit $t = T$ Thus the objective function to be maximised is

$$P = 1 - x(T) - y(T) \tag{5}$$

The/

The problem as stated is clearly of a form to which the Maximum Principle$^{(2)}$ is applicable. The adjoint equations corresponding to (1) and (2) are

$$\frac{d\lambda_1}{dt} = -f k_1 (\lambda_2 - \lambda_1) \tag{6}$$

and

$$\frac{d\lambda_2}{dt} = f k_2 (\lambda_2 - \lambda_1) + (1-f) k_3 \lambda_2 \tag{7}$$

and the appropriate boundary conditions corresponding to the objective function (5) are

$$\lambda_1(T) = \lambda_2(T) = -c \qquad (c > 0) \tag{8}$$

Then a necessary condition for optimality of $f(t)$ is that the Hamiltonian

$$\left.\begin{aligned} H &= f[(\lambda_2-\lambda_1)(k_1 x - k_2 y) + \lambda_2 k_3 y] - \lambda_2 k_3 y \\ &= fJ - \lambda_2 k_3 y \qquad (\text{say}) \end{aligned}\right\} \tag{9}$$

should take its greatest value (regarded as a function of f) for each t. Since H is linear in f , this implies that $f = 0$ or $f = 1$, depending on the sign of J at the point in question. f may take a value between these bounds only if J vanishes, and if this is the case over a finite interval of t , the corresponding part of the solution will be referred to as a <u>singular segment</u>.

At $t = T$, taking account of equation (8), the Hamiltonian is/

is seen to be

$$H = -cfk_3y + ck_3y$$

and since $c>0$, this is maximised by taking $f=0$. Thus the optimal catalyst blend starts back from $t=T$ with $f=0$, and will retain this value until J changes sign. At this point the optimal blend may switch to $f=1$ or, in certain circumstances, to an intermediate value of f corresponding to a singular segment. In fact, this second possibility is important in the present problem, so before proceeding further we will investigate in more detail the possible form of a singular segment.

Conditions for a singular segment.

If f is to take values between its bounds for the finite time interval $t_1<t<t_2$ it is clearly necessary that

$$J = (\lambda_2-\lambda_1)(k_1x-k_2y)+\lambda_2 k_3 y = 0 \quad (t_1<t<t_2) \qquad (10)$$

which, in turn, implies that $dJ/dt = 0$ for all $t_1<t<t_2$. Using the differential equations (1) and (2) and the adjoint equations (6) and (7), this condition can be reduced to the form

$$\frac{dJ}{dt} = k_3(\lambda_2 k_1 x - \lambda_1 k_2 y) = 0 \quad (t_1<t<t_2) \qquad (11)$$

Solving equation (11) for λ_1, and substituting into equation (10) we then obtain the following necessary condition for a singular segment/

segment

$$\frac{\lambda_2}{k_2 y}\left[k_2 k_3 y^2 - (k_1 x - k_2 y)^2\right] = 0 \quad (t_1 < t < t_2) \tag{12}$$

whence

$$\lambda_2 = 0 \tag{13a}$$

or

$$k_2 k_3 y^2 = (k_1 x - k_2 y)^2 \tag{13b}$$

But from equation (11) it follows that (13a) would imply $\lambda_1 = 0$, and hence $H = 0$ at all points of the singular segment. However, the form of H at $t = T$ has already been found and been seen to be maximised by $f = 0$. The corresponding maximum value of H is

$$H_{max}(T) = c k_3 y > 0$$

Thus a singular segment on which $H = 0$ cannot belong to the optimal solution, since it is known from the Maximum Principle that H_{max} must remain constant throughout the entire solution. The possibility (13a) can therefore be discounted and we are left with (13b), which reduces to

$$k_1 x = k_2 y \left(1 \pm \sqrt{\frac{k_3}{k_2}}\right) \tag{14}$$

Then from equation (11) the corresponding relation between the adjoint variables is

$$\frac{\lambda_1}{\lambda_2} = 1 \pm \sqrt{\frac{k_3}{k_2}} \tag{15}$$

Equation/

Equation (14) gives an algebraic equation for the singular segment in the x-y plane and shows that it may be one of two straight lines through the origin. It must also, of course, be a solution of differential equations (1) and (2) for a suitable choice of f, and to determine the necessary form of $f(t)$ we may compare the differential equation

$$\frac{dy}{dx} = \frac{k_1}{k_2\left(1 \pm \sqrt{\frac{k_3}{k_2}}\right)} \tag{16}$$

obtained from equation (14), with the corresponding differential equation obtained by dividing (1) and (2), namely

$$\frac{dy}{dx} = \frac{f(k_1x - k_2y) - (1-f)k_3y}{-f(k_1x - k_2y)} \tag{17}$$

Evaluating the ratio y/x in equation (17) from equation (14), it is found that (17) reduces to (16) if and only if

$$\frac{k_1}{k_2\left(1 \pm \sqrt{\frac{k_3}{k_2}}\right)} = \frac{\pm f\sqrt{k_3/k_2} - (1-f)(k_3/k_2)}{\mp f\sqrt{k_3/k_2}} \tag{18}$$

This may be simplified by writing

$$k_1/k_2 = \beta \quad ; \quad +\sqrt{k_3/k_2} = \alpha \tag{19}$$

and solving for f. Taking the upper signs in equation (18), we then obtain

$$f = \frac{\alpha(1+\alpha)}{\beta + (1+\alpha)^2} \tag{20}$$

and/

and correspondingly, from equations (14) and (15)

$$\lambda = \lambda_1/\lambda_2 = 1+\alpha \tag{21a}$$

and

$$z = y/x = \beta/(1+\alpha) \tag{21b}$$

where we have introduced the abbreviations λ and z for the ratios λ_1/λ_2 and y/x respectively. Similarly, taking the lower signs in equation (18) gives

$$f = \frac{-\alpha(1-\alpha)}{\beta+(1-\alpha)^2} \tag{22}$$

together with

$$\lambda = 1-\alpha \tag{23a}$$

and

$$z = \beta/(1-\alpha) \tag{23b}$$

Now if $\alpha < 1$ equation (22) gives a negative value for f, which is physically unacceptable, while if $\alpha > 1$ equation (23b) gives a negative value for z, which is also physically unacceptable. Thus equations (20) and (21) describe the properties of the only acceptable singular segment for this problem. From equation (20) it is easy to see that $f < 1$ for all $\alpha, \beta > 0$, so this value of f lies in the permitted interval whatever the reaction kinetics.

The singular segment just found has the interesting and unusual property that the control variable takes a constant value along it. Thus an optimal solution of the complete problem retains its "switching" character even when it includes a singular segment. It differs from a "bang-bang"/

"bang-bang" solution only by the introduction of a third switching level (given by equation (20)) between the upper and lower bounds of the control variable.

The singular segment (21b) is indicated on the x-y plane in Fig. 1

Form of the complete solution

We have already seen that the optimal solution starts back from $t=T$ with $f = 0$. Then from equation (1), $dx/dt = 0$, and the corresponding trajectory in the x-y plane is a line segment parallel to the y axis, as indicated in Fig. 1. The blend $f = 0$ is retained until J reaches the value zero as t decreases, then it is necessary to switch to $f = 1$, as at point E in Fig. 1. However, if the/

the vertical segment meets the singular segment at the point where $J = 0$ there is the alternative possibility of switching to the value of f given by equation (20) and following the singular segment for a time before switching to $f = 1$. This possibility is illustrated by the vertical segment CD in Fig. 1.

Having switched to $f = 1$, it can be shown that the condition $J = 0$ is not satisfied again, so a segment $f = 1$, corresponding to a line of slope - 1 in the $x - y$ plane, must be followed back to the initial conditions represented by the point A (1, 0). Thus the two possible ways of satisfying the Maximum Principle and the initial conditions are typified by the trajectories ABCD, making use of the singular segment, and AEF, which is a pure bang-bang trajectory. If the reactor is sufficiently short, corresponding to a small value of T, the switching must occur at a point on the segment $f = 1$ lying between A and B, and there is no solution using the singular segment. For small values of T, therefore, we expect to find only one blending policy which satisfies the Maximum Principle, namely an interval with $f = 1$ followed by an interval with $f = 0$. For larger values of T, on the other hand, there will be two alternative policies, one the bang-bang policy just described, and the second making use of an intermediate switch to a blend lying between the two limits. Which of these policies gives the larger value of the objective function can be found only by direct comparison of results, since both satisfy the Maximum Principle.

We shall now develop each of these solutions in greater detail, determining the switching points and the relation between the final value of the objective function P and the total residence time T. The time optimal solution will then be the one which gives the largest value of P at a given T, and this may be found by plotting the curve of P against T for/

for each case.

The bang-bang solution

Let t_s be the value of t at which the switch from $f = 1$ to $f = 0$ is made. Then $J(t_s) = 0$, and from equation (10) this implies that

$$\lambda = 1 + \frac{\alpha^2 z}{\beta - z} \qquad (24)$$

at $t = t_s$. From equations (6) and (7) it is easy to obtain a differential equation for λ, namely

$$\frac{d\lambda}{dt} = -f(1-\lambda)(k_1 + \lambda k_2) - (1-f)\lambda k_3 \qquad (25)$$

and on the segment $f = 0$ this reduces to

$$\frac{d\lambda}{dt} = -k_3 \lambda \qquad (26)$$

The segment $f = 0$ has the values $t = t_s$ and $t = T$ as its termini, and we know from equation (8) that $\lambda(T) = 1$. Thus, integrating equation (26)

$$\lambda(t_s) = e^{k_3(T - t_s)}$$

and using this in equation (24) permits $T - t_s$ to be expressed in terms of z_s, the value of z at the switching point

$$k_3(T - t_s) = \ln\left(1 + \frac{\alpha^2 z_s}{\beta - z_s}\right) \qquad (27)$$

A/

A second expression relating t_s and z_s can be obtained directly by integrating the differential equation for z forwards from $t = 0$. From equations (1) and (2)

$$\frac{dz}{dt} = f(1+z)(k_1 - k_2 z) - (1-f)k_3 z$$

and on the segment $f = 1$ this reduces to

$$\frac{dz}{dt} = (1+z)(k_1 - k_2 z)$$

But $z = 0$ at $t = 0$, so integrating between the limits $t = 0$ and $t = t_s$ gives

$$k_3 t_s = \frac{\alpha^2}{1+\beta} \ln\left[\frac{\beta(1+z_s)}{\beta - z_s}\right] \qquad (28)$$

Adding equations (27) and (28) then permits T to be expressed in terms of z_s

$$k_3 T = \frac{\alpha^2}{1+\beta} \ln\left[\frac{\beta(1+z_s)}{\beta - z_s}\right] + \ln\left(1 + \frac{\alpha^2 z_s}{\beta - z_s}\right) \qquad (29)$$

Finally we wish to calculate $x(T)$, $y(T)$ and hence P as functions of z_s. At $t = t_s$

$$y_s / x_s = z_s \qquad \text{and} \qquad y_s + x_s = 1$$

whence

$$x_s = \frac{1}{1+z_s} \quad \text{and} \quad y_s = \frac{z_s}{1+z_s} \qquad (30)$$

On the terminal segment $f = 0$, x does not change, so

$$x(T) = x_s = 1/(1+z_s) \qquad (31)$$

while/

while $y(T)$ may be related to y_s by integrating equation (2) after setting $f = 0$, with the reslt

$$y(T) = y_s e^{-k_3(T-t_s)}$$

Using equations (27) and (30), this reduces to

$$y(T) = \frac{z_s}{1+z_s}\left[\frac{1}{1+\frac{\alpha^2 z_s}{\beta - z_s}}\right] \tag{32}$$

Thus

$$P = 1 - x(T) - y(T) = 1 - \frac{1}{1+z_s} - \frac{z_s}{1+z_s}\left[\frac{1}{1+\frac{\alpha^2 z_s}{\beta - z_s}}\right] \tag{33}$$

Equations (29) and (33) provide the relation between P and T in parametric form, with z_s as a parameter. Similarly equations (28) and (29) provide a parametric relation between t_s and T. Note that $T \to 0$ as $z_s \to 0$ and $T \to \infty$ as $z_s \to \beta$ so the complete range of bang-bang solutions is obtained as z_s traverses the finite interval $0 \to \beta$.

The solution making use of the singular segment

The path ABCD in Fig. 1 represents a solution which makes use of the singular segment. Clearly this is possible only if $T > t_1$, where t_1 is the value of t at which the singular segment is reached along a trajectory $f = 1$ starting from the initial conditions. But if t_s and z_s are replaced by t and z in equation (28), we have a general relation holding at any point of the segment $f = 1$ issuing from the initial conditions. If, in particular, we set $z = \beta/(1+\alpha)$ in this, we obtain the value $t = t_1$ immediately

$$k_3 t_1 = \frac{\alpha^2}{1+\beta} \ln\left(\frac{1+\alpha+\beta}{\alpha}\right) \tag{34}$$

Suppose we leave the singular segment (point C in Fig. 1) at $t = t_2$, when $y = y_2$. On the singular segment equation (2) reduces to

$$\frac{1}{k_3}\frac{dy}{dt} = \frac{-\beta y}{\beta + (1+\alpha)^2}$$

and this may be integrated to obtain the following relation:

$$\ln\left(\frac{y_1}{y_2}\right) = \frac{\beta}{\beta + (1+\alpha)^2} k_3 (t_2 - t_1) \tag{35}$$

where $y_1 = y\ (t_1)$, the value of y at the intersection of the singular segment and the line $x + y = 1$. Clearly

$$y_1 = \frac{\beta}{1+\alpha+\beta}$$

so equation (35) reduces to

$$k_3 (t_2 - t_1) = \frac{\beta + (1+\alpha)^2}{\beta} \ln\left[\frac{\beta}{y_2 (1+\alpha+\beta)}\right] \tag{36}$$

On the singular segment we know that $\lambda = 1+\alpha$, while at $t = T$, $\lambda = 1$. Thus, integrating equation (26) between these limits along the final segment with $f = 0$ (CD in Fig. 1):

$$k_3 (T - t_2) = \ln(1+\alpha) \tag{37}$$

Then adding equations (34), (36) and (37) gives the value of T:

$$k_3 T = \ln(1+\alpha) + \frac{\alpha^2}{1+\beta}\ln\left(\frac{1+\alpha+\beta}{\alpha}\right) + \frac{\beta+(1+\alpha)^2}{\beta}\ln\left[\frac{\beta}{y_2(1+\alpha+\beta)}\right] \qquad (38)$$

We then wish to express P in terms of y_2 so as to obtain a parametric relation between P and T, with y_2 as a parameter. From the equation of the singular segment

$$x(T) = x(t_2) = y(t_2)\left(\frac{1+\alpha}{\beta}\right) = y_2\,\frac{1+\alpha}{\beta} \qquad (39)$$

while by integrating equation (2) along the final segment f = 0

$$y(T) = y_2\, e^{-k_3(T-t_2)} = y_2/(1+\alpha) \qquad (40)$$

where we have made use of equation (37). Finally, from equations (39) and (40)

$$P = 1 - x(T) - y(T) = 1 - y_2\left(\frac{1}{1+\alpha} + \frac{1+\alpha}{\beta}\right) \qquad (41)$$

Equations (38) and (41) then provide the desired parametric relation between P and T, while equations (34) and (37) give the two switching times which determine the optimum blending policy.

The solution is a valid alternative to the pure bang-bang solution only when $t_2 > t_1$, of course; in other words, when

$$y_2 < \frac{\beta}{1+\alpha+\beta} \qquad (42)$$

Numerical results

To/

To illustrate the nature of the solution consider the values of the kinetic constants taken by Gunn and Thomas[(1)], namely

$$k_1 = k_3 = 1 \quad ; \quad k_2 = 10$$

A curve giving P as a function of $k_3 T$ can be computed for the bang-bang solution from the parametric equations (29) and (35), and a corresponding curve for the solution making use of the singular segment can be computed from equations (38) and (41). These curves are presented in Fig. 2, from which it is seen that the solution making use of the singular segment gives the larger value of P over the whole range $k_3 T > 0{\cdot}411$, and therefore represents the optimum blend policy over this range of T. When $k_3 T < 0{\cdot}411$, on the other hand, inequality (42) is violated, so the bang-bang solution is the only solution satisfying the maximum principle and therefore represents the optimum blend policy. $k_3 T = 0{\cdot}411$ corresponds to point A in Fig. 2.

Since P represents the mole fraction of the desired product C in the mixture leaving the reactor, it is clear from Fig. 2 that the optimum blend policy makes use of the singular segment for all reactors which are capable of giving a reasonable yield of C. Only for very short reactors giving an unacceptably small yield would the bang-bang solution be optimal. Using equations (34) and (37) it is found that the optimal blend policy takes f = 1 for an interval of length,

$$k_3 t_1 = 0{\cdot}1363$$

at the entry to the reactor, and takes f = 0 for an interval of length

$$k_3 (T - t_2) = 0{\cdot}2747$$

before the reactor exit. Throughout the remainder of the reactor length, whatever/

whatever it may be (provided $k_3T > 0{\cdot}411$ of course), the blend takes the intermediate value

$$f = 0{\cdot}2272$$

computed from equation (20). This is illustrated in Fig. 3, which shows the optimum blend policy for $k_3T = 1{\cdot}0$.

Physically the above choice of reaction velocity constants corresponds to a system in which the conversion is strongly limited by the reversibility of the reaction $A \rightleftharpoons B$. Thus a large proportion of the total reactor volume available is taken up by mixed catalyst which enables the reactions to circumvent this limitation. If a smaller value is taken for k_2 , thus reducing the importance of the reverse reaction, the interval occupied by mixed catalyst is decreased and the initial and terminal regions occupied by pure catalysts become more important. Indeed it is not difficult to see that $t_1 \to \infty$ when $k_2 \to 0$, so in the limit where the first reaction is irreversible, the bang-bang solution is optimal in all cases and the catalyst should never be mixed.

Fig. 2 shows that the optimal solution using a mixed catalyst may give spectacular increasesin yield compared with the conventional arrangement of two reactors in series, in which the reactions $A \rightleftharpoons B$ and $B \to C$ sre separately catalysed.

References

1. Gunn, D.J. and Thomas, W.J. Chem. Eng. Sci. 20, 89 (1965).

2. Pontryagin, L.S. et al. "The Mathematical Theory of Optimal Processes", Pergamon Press (1964).

Captions for diagrams

Fig. 1: Solution trajectories in the x-y plane

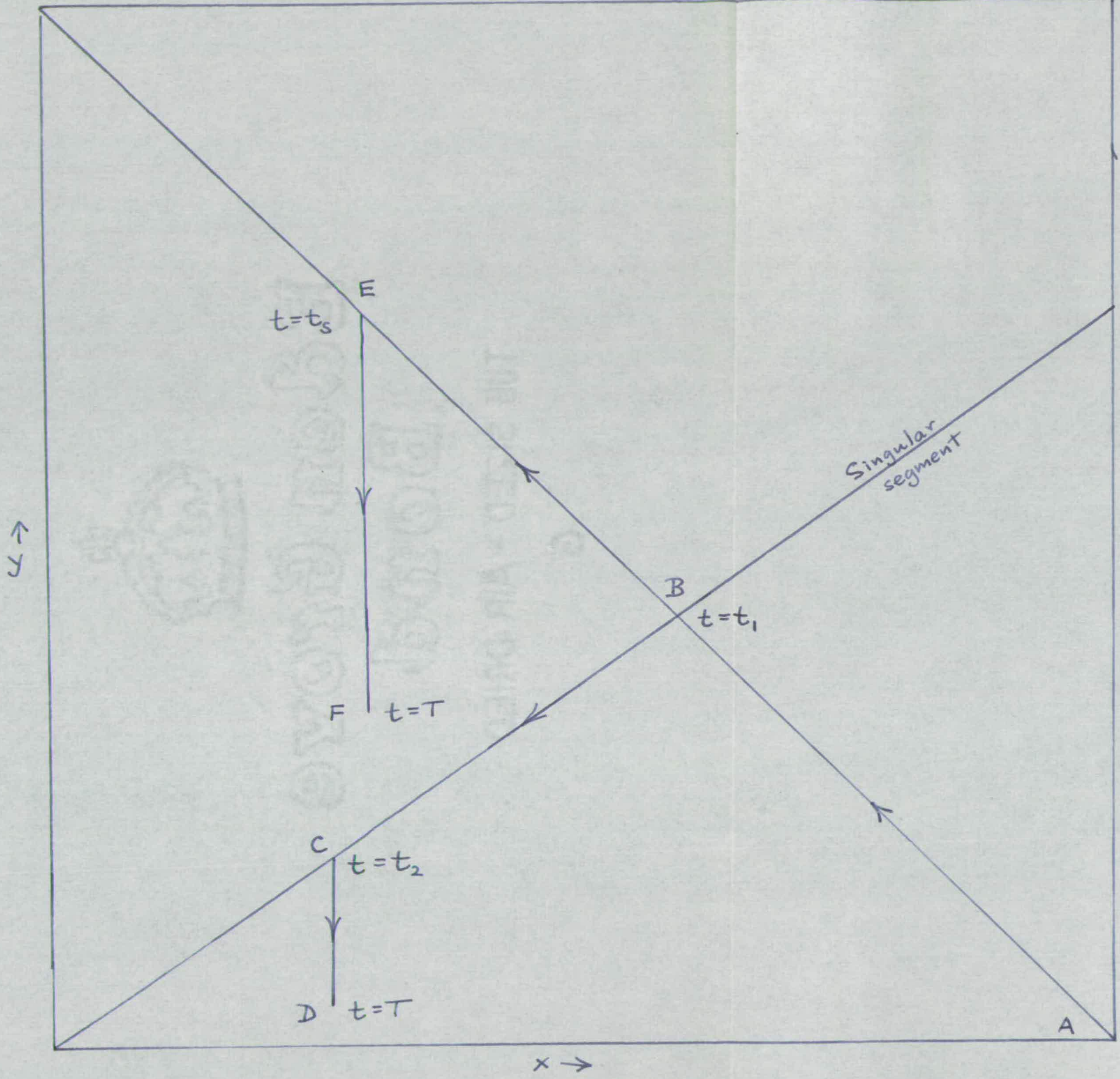
Fig. 2: Optimum yield as a function of reactor length for the two solutions satisfying the Maximum Principle.

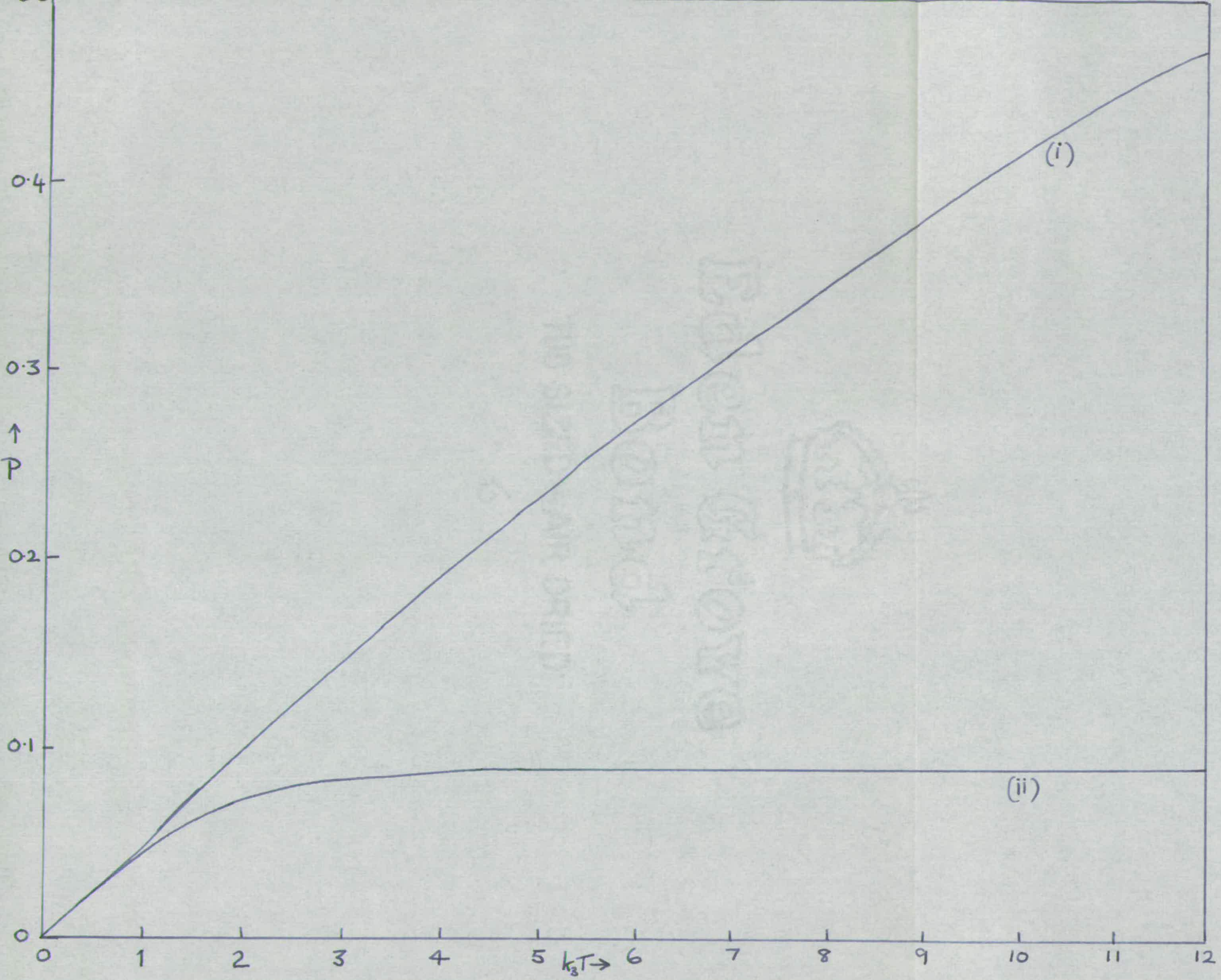
$k_1 = k_3 = 1, \quad k_2 = 10$

(i) Solution using singular segment, (ii) Bang-bang solution.

Fig. 3: Optimal blend policy when $k_3T = 1.0$

$k_1 = k_3 = 1, \quad k_2 = 10$


E
t=t_s
F
t=T
B
t=t_1
Singular segment
C
t=t_2
D
t=T
A
y ↑
x →


0·4
0·3
↑
P
0·2
0·1
0
0
1
2
3
4
5
$k_BT \rightarrow$
6
7
8
9
10
11
12
(i)
(ii)

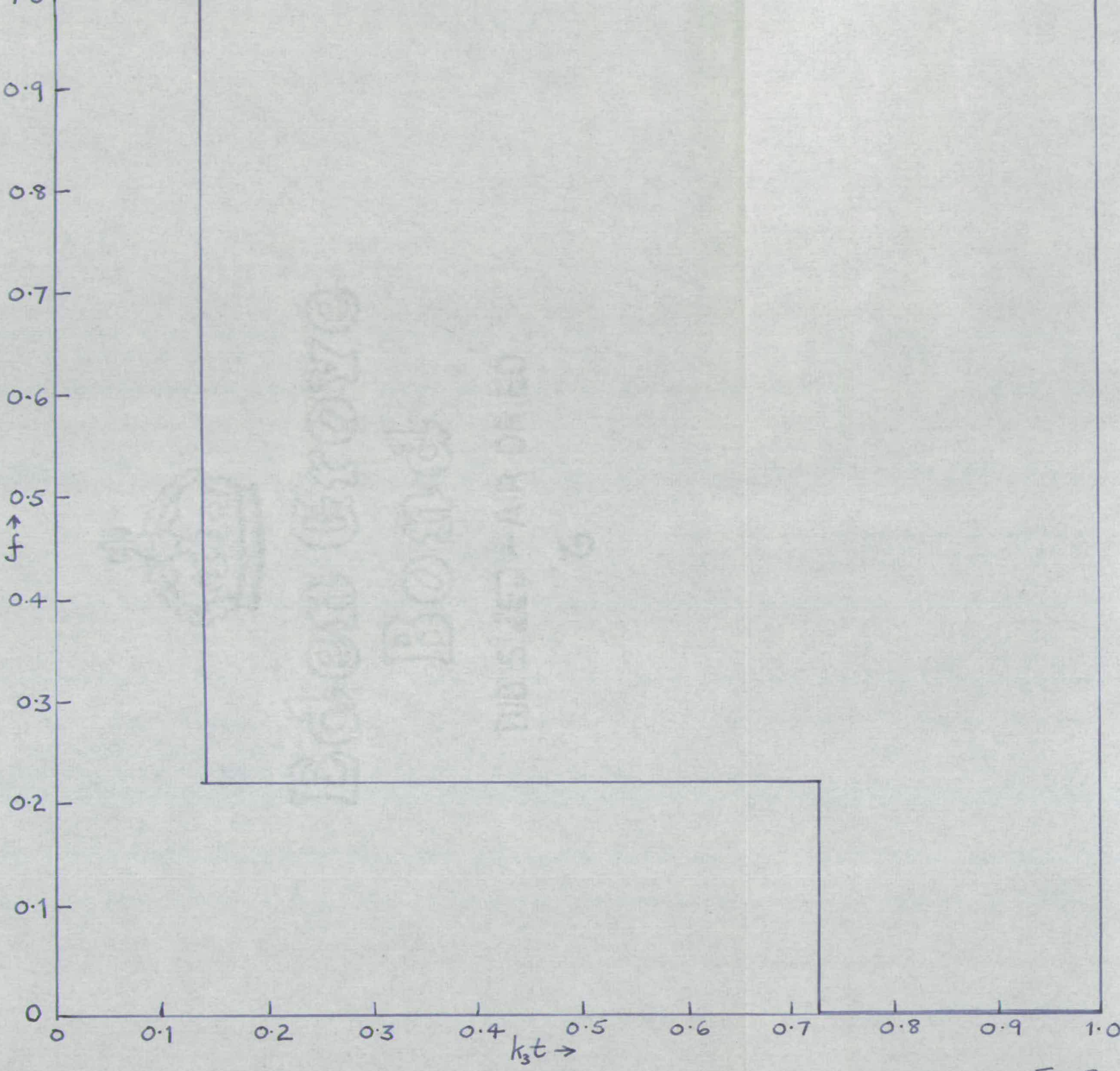


Fig 3

C8

REACTOR OPTIMIZATION PROBLEMS FOR REVERSIBLE EXOTHERMIC REACTIONS

D. C. Dyson, F. J. M. Horn, R. Jackson, and C. B. Schlesinger
Department of Chemical Engineering
Rice University

ABSTRACT

Optimization of tubular reactors for exothermic reversible reactions is considered. For the cases where there are no side reactions and there is no decay of catalyst various types of temperature control are investigated. The remaining cases considered are a) stable catalyst, irreversible decay of product and b) decaying catalyst, stable product. In each of these cases perfect indirect temperature control is treated. Particular attention is given to devising numerical methods which take advantage of the structure of the problem in question and to convenient representation of data.

I. INTRODUCTION

Some of the earliest applications of physico-chemical principles to the design and operation of efficient industrial chemical reactors were made for processes involving a single reversible reaction such as ammonia synthesis ($N_2 + 3H_2 = 2NH_3$), the water gas shift reaction ($CO + H_2O = CO_2 + H_2$) and the oxidation of sulphur dioxide ($2SO_2 + O_2 = 2SO_3$).

The objective of this paper is to discuss some typical reactor optimization problems arising with such reactions. Importance will be given to the mathematical formulation of the different problems and to the convenient representation of results. Only the general ideas behind the methods of solution will be discussed. For mathematical details we shall refer to other papers. The reactions under consideration are exothermic which means that the equilibrium conversion decreases with increasing temperature; on the other hand, the rate of the forward reaction increases with increasing temperature, and in fact, it was early recognized that in the interests of catalyst economy the temperature should be high in the first part of the reactor (where the reverse reaction is slow because of lack of products) and low in the last part. In order to achieve this, two methods of cooling are employed in practice:

1) indirect cooling by means of heat exchangers, and

2) direct cooling by adding cold gas to the mixture.

Often in practice the reactor consists of a set of stages in each of which the reaction takes place adiabaticly while between the stages the cooling is achieved either by method 1) or 2). This type of reactor will be called an adiabatic cascade. In general, the maximum conversion obtainable for given total amount of catalyst, total mass flow, and inlet composition increases as the number of stages increases. It is of practical interest to know the limit of the performance of a multistage reactor as the number of stages approaches infinity. This is equivalent to considering a reactor in which the temperature can be controlled by either method 1) or 2) at any point along the reaction tube. This type of control will be called "perfect". Perfect control can be approximated either by using many adiabatic stages (the cross-section of which can be arbitrarily large) or by removing heat from or adding cold gas to the reaction zone. In the latter case the problem of transport of heat and mass perpendicular to the main direction of flow arises. Examples in this class are the U-tube and the push-pull reactor which will be treated later.

In some cases (e.g. adiabatic packed bed reactors) catalytic reactors for the reactions under investigation can be well approximated by the ideal tubular reactor

INSERT p. 3

In other cases, for example indirectly cooled reactors, this plug flow model is an important limiting case, the properties of which are of practical interest because they normally set an upper limit for the performance of realizable reactors.

model for which it is assumed that the fluid velocity, composition, temperature and pressure are constant across any plane of section and that there is no mixing (by any mechanism, e.g. diffusion or convection) in the direction of main flow. INSERT The pressure will be assumed to be constant along the reactor as the influence of the pressure drop on the kinetics is negligible in the industrial cases mentioned here.

Ammonia synthesis, the water gas shift reaction, and the oxidation of sulphur dioxide can be described by a single stoichiometric reaction. Also, in these cases the catalyst used is fairly stable so that catalyst deterioration should not be taken into account in the mathematical model. The reactor material balance is then described by a single ordinary differential equation. In Section II of this paper we will deal with this situation. In Section III a short account will be given of problems arising when the reaction kinetics is more complex, for one of the following reasons:

1) the number of stoichiometrically independent reactions is greater than one.

2) : catalyst deterioration has to be taken into account.

In the first case the reactor material balance is described by more than one ordinary differential equation. In the latter case partial differential equations are needed since there are two independent variables; that is distance and time. Only perfect indirect temperature control will be considered in the last part of the paper.

II. OPTIMUM PROBLEMS WITH A SINGLE STOICHIOMETRICALLY INDEPENDENT REACTION AND WITH STABLE CATALYST

A. Perfect Indirect Control

The concept of the ideal tubular reactor with perfect indirect temperature control was introduced early and the problem of finding the relationship between temperature and position in this reactor such that the volume required for a given duty is minimum was solved by Leitenberger[1] and others:[2-13] the optimum policy may be obtained by maximizing the reaction rate with respect to temperature at each point in the reactor.

If t_f^* is the minimum reactor volume required to change the fractional conversion y of some reference reactant (not present in stoichiometric excess) which flows through the reactor in a steady stream at a given total mass flowrate, from 0 at the inlet to y_f at the outlet, then it can be shown that for a large class of reaction rate expressions if the t_f^*, y_f relation is plotted with scales proportional to log t_f^* and log $(y_f/(1-y_f))$ respectively, the curve approaches a straight line as $y_f \rightarrow 1$. These scale transformations have been found useful for the representation of the minimum volume, y_f relations for a wide class of reactors over a considerable range of y_f (see Figures 4 and 5).

The curve C^* divides the y_f, t_f plane (Figure 1) into two regions, such that only points on C^* and in the shaded region below C^* are attainable by tubular reactors with any temperature control. The boundary points of the region (the curve C^*) represent reactors which are optimal with respect to various objective functions of practical significance (see also Section III.A).

The boundary curve is easily obtained numerically from the rate expression. Sometimes it can be expressed in terms of known functions.[11] In order to compare various reactor types numerically calculations have been carried out for an example. In this example the reaction rate expression was chosen as*

$$v(y,T) = H(1-y)e^{-A/T} - H'ye^{-A'/T} \qquad (1)$$

the sets of parameters used are given in Table 1 and the results of the calculations are plotted in Figures 3 and 5.

B. Perfect Direct Control

In this section we consider a reactor which has a main feed F_o (see Figure 11) and a supplementary feed F_m. We shall

*For the definitions of the symbols see Appendix treat only reactors where the compositions of F_o and F_m are the same.

In the appendix the conversion y, the variable ω (which is proportional to the total mass flow), the volume variables m and t, and the reaction rate v(y,T) are defined.

No special form is assumed for the function v(y,T) but the thermochemistry is assumed to be such that the adiabatic temperature rise coefficient is constant.* (See Table 1.)

The minimum volume problem which we will now discuss for this reactor is as follows.

For given reaction rate function and thermochemistry, feed conversion y_m, supplementary feed temperature T_m, exit conversion y_f, and for $\omega = 1$ at the exit from the reactor determine the minimum volume $m_{f\ min}$ where the temperature T_o of the main feed is freely adjustable** and the feed distrubution $\omega(m)$ is also freely adjustable (except that ω is, of course, to be a non-decreasing function of m, i.e. material may leave the reactor only at the right hand end in Figure 10).

Consider the contours of the reaction rate v in the plane of y and T. For non-autocatalytic single reversible exothermic reactions the rate contours for expressions proposed in the literature have the shape shown in Figure 2.

*In Reference 14 the corresponding discussion covers the case of an arbitrary thermochemistry.

**It is more convenient to treat an equivalent problem where the variable C (see appendix) is taken as being freely adjustable. C and w(m) determine T_o

The subsidiary feed condition is represented by the point O. The lines of constant enthalpy (adiabatic trajectories for a tubular reactor) are straight lines with the direction DE. Consider a point in the reactor corresponding to D. If no cold gas is added the state of the mixture will change such that the point representing it moves in the direction DE as m increases.

If cold gas is added at a very high rate so that the change of the state of the mixture by reaction is negligible in comparison to the change of the state due to mixing, the point representing the state of the mixture will move in the direction DF as m increases. All directions pointing into the shaded region bounded by the two directions just mentioned can be obtained by choosing the rate of addition of cold gas appropriately. A point on the line OB however, can move only in the direction of this line (upwards or downwards). Points to the left of this line can not be reached from a starting point to the right of this line.

Indirect cooling can be considered as a limiting case of direct cooling. If T_m is given the value $- \infty$ (a purely mathematical device, of course) then direct cooling becomes equivalent to indirect cooling because by adding an infinitesimally small amount of subsidiary feed one can produce any given drop in the temperature without changing the composition. In this case the direction corresponding to DO becomes the direction parallel to the T axis. The locus of points where the rate is maximum with respect to displacements in this direction (i.e. with respect to temperature variations) corresponds to optimal perfect indirect control. The curve L is the locus of points at which the rate is maximal with respect to displacements in the direction DO. To follow this locus is to maximize the reaction rate at each point in the reactor with respect to the available control. As pointed out such a policy is known to be optimal for perfect indirect control. One might be tempted to surmise that the optimal control would trace out this curve L in Figure 2; however, this is not so in general. That this policy is not optimal for $y_f = y_{f_1}$ in Figure 2 is clear, and it has, in fact, been established that it is not, in general, optimal for $y_f = y_{f_2}$ in Figure 2 either.

The optimal trajectory may be found by methods which have already been established.[14,30]

Except in certain special cases (which arise when T_m is unreasonably hot with respect to y_f and will not be discussed here) the optimal control may be described as follows.

Consider a given flowrate w_o and the corresponding value of C such that the initial temperature and composition are represented by the point A in Figure 2. Now let w be increased in such a way that the curve L in Figure 2 is traced out until $w = 1$ (point D' in Fig. 2): then w is maintained equal to 1 until y attains its prescribed value y_f (point B'' in Fig. 2)

Then for some value of C this is the optimal policy. (Note that both w_o and the breakpoint B' are completely determined by C.) There is no simple way to determine the optimum value of C for a given y_f. However, by choosing several values of C one can determine the corresponding y_f, m_f relations and the optimum relation will be their envelope. (See Figure 3 for the results corresponding to $y_m = 0.01$, $T_m = 300°$ and the first set of parameters in Table I. The corresponding ~~result for perfect indirect control is also plotted.)~~ One can shorten the numerical work by making use of the fact that only one integration along the curve L (up to the furthest breakpoint) is required. The contribution to m [illegible] to integration along L up to other breakpoints is then easily determined.[14,30]

To compute the curves shown in Figure 3 with a relative precision better than 0.0001 in m_f, not more than 4 seconds are required on an ICT Atlas computer.

These optimal controls have a limit as C → 0 which may be thought of as the entire portion AB of curve L in Figure 2 mapped onto the entrance of the reactor (m=0) at which place $w = 0$, and w(m) so adjusted to keep y and T in the reactor constant at values corresponding to B in Figure 2 until w = 1, at which stage w is made ≡ 1. Such a policy is, of course, identical to the tank-tube policy,[14] and it is, in fact optimal in some cases. Point B in Figure 4 corresponds to such a control.

If $y_m = 0$, then for many reaction rate expressions (such as Equation (1)), the point on the L locus corresponding to A in Figure 2 has a coordinate $T = \infty$ and the above theory breaks down. In such cases the optimum relation may be estimated precisely by using a slight modification of the above procedure.[14]

C. Indirect Control of an Adiabatic Cascade

It has already been noted that by making $T_m = -\infty$ direct control becomes equivalent to indirect. The equations mentioned in the following subsection can be easily specialized for this case and the equations thus obtained are well known*.[25] In Figure 5 results are given for N = 1,2,3, ∞ stages forthe first set of parameters in Table 1.

D. Direct Control of an Adiabatic Cascade

In this case, for given J, y_m and T_m and kinetics we wish to adjust the N-1 ratios of catalyst masses and the N-1 bypass ratios and the main feed temperature in such a way as to end up with w = 1, and a given conversion y_{e_N} and a minimum total mass of catalyst. (See Figure 9.)

There have been many ways proposed for treating this problem.[17-21] The most efficient by far makes use of the equations found by taking the derivatives of the total mass of

*For a survey see page 23 of Reference 14.

catalyst:

$$m_f = \sum_{i=1}^{i=N} \omega_i \int_{y_{a_i}}^{y_{e_i}} \frac{dy}{v(y,T)} \qquad (2)$$

with respect to a suitably chosen set of 2N-1 free variables taking into account the restraint relations (heat and mass balances for the subsidiary feed addition):

$$\omega_j y_{e_j} + (\omega_{j+1} - \omega_j) y_m = \omega_{j+1} y_{a_{j+1}} \; ; \; j = 1,2, \ldots N-1 \qquad (3)$$

$$\omega_j T_{e_j} + (\omega_{j+1} - \omega_j) T_m = \omega_{j+1} T_{a_{j+1}} \; ; \; j = 1,2 \ldots N-1 \qquad (4)$$

If this is done one obtains a set of equations[14-17] first derived by K. Konoki which can, in general*, be solved by treating a series of easy one-dimensional problems if T_{a_1} is guessed. For m_f each T_{a_1} one obtains a rea$\overset{c}{\wedge}$tor which is optimal for some y_{e_N}. By varying T_{a_1} the optimal y_{e_N}, m_f relation is obtained. A procedure described in Reference 14 based on Konoki's equations was found to be over 100 times as fast as a dynamic programming method.[19]

In Figure 4 results calculated for N = 1,2,3, and ∞ are given for the first set of reaction rate function parameters in Table 1, and for $y_m = 0$, $T_m = 600$°K. Now T_{a_1} varies along each

*The discussion of these equations is complicated[14] and will not be given here.

of the curves for N = 1,2, and 3, being high for low y_{e_N} and decreasing as y_{e_N} increases. At the ends of the curves N = 3 and N = ∞ close to B;t [illegible] $T_{a_1} = T_m$. The point A represents the case where T_m is relatively too hot to be any use in the case of a two stage reactor. It also represents the degenerate point for 3, 4, 5....∞ staged reactors (the missing portions of the curves for the 3 and ∞ staged reactors were not computed but can be expected to follow the N = 2 curve closely and meet at the point A).

E. U-Tube Reactors: Empty-Full and Push-Pull Reactors

Consider the reactor shown in Figure 6 in which heat (but not mass) may be transferred across the dividing wall. Two cases are important.

1) No reaction takes place in the left hand tube; e.g. where the right hand tube is packed with catalyst but the left hand tube is not. This we will call the empty-full reactor which has found practical application in ammonia synthesis and SO_2 oxidation.

2) Both tubes are identical and reaction takes place in each tube. Two such reactors are shown in Figure 8. By a symmetry argument it has been shown that these two are equivalent* to the countercurrent scheme in Figure 7. Reactors

*plug flow with no axial dispersion of mass or heat assumed, and heat transfer orthogonal to flow with all resistance lumped into the boundary. A constant overall heat transfer coefficient was used as in Reference 22.

of these types are called push-pull reactors because they are also equivalent, (see footnote, page 12), to the limiting operation of a pebble heat exchange reactor (Figure 11) with flow in the direction of the arrows for one half cycle and in the opposite direction for the other provided that the cycling is neither so slow that the pebble temperatures vary appreciably with time nor so fast that back-mixing becomes serious.

The problem of minimizing the volume (for $w \equiv 1$) for zero inlet conversion, parametricly in the exit conversion (with inlet temperature and heat transfer coefficient as the two free variables) was solved by using a perturbation technique for the derivatives and a modified iterative gradient method.[23,24]

In the case of the empty-full reactor only the volume of the right hand tube (Figure 6) was considered as the reactor volume but for the push-pull reactor both sides were taken.*

The results are shown in Figure 5 for the first set of rate parameters in Table 1. For the push-pull reactor (P-P) the broken line indicates that the minimum was not determined with great precision for the corresponding range of values of y_f. The empty-full reactor (E-F) results fell almost exactly on the line for the optimum two stage adiabatic reactor with indirect

*We have in mind a solid catalyzed fluid reaction where only the volume which is packed with catalyst is important.

intercooling, and so are not shown.*

The push-pull reactor has internal heat exchange (pebbles) which may in some cases be provided more cheaply than interstage cooling. The mathematically equivalent full-full reactor requires more volume than the empty-full reactor but never twice as much so that one might expect that in cases where volume is expensive, per se (as in high pressure processes) and not on account of the catalyst required to pack it, one could improve on a given empty-full reactor (such as a Haber Bosch ammonia converter) by using fewer tubes, each of a larger diameter, but packed with catalyst.

F. Single Stirred Tank

Also shown in Figure 5 is the corresponding result for a single (adiabatic) C.S.T.R. captioned (M). It is of interest to compare it with the one stage adiabatic tubular reactor (captioned N = 1) and to notice that at high conversions its performance is superior. Such is not the case for the second set of parameters[14] for which the tube is always better especially at high conversions.

*For other reaction rate function parameters which we have studied the E-F reactor was found to require from 10% less to 10% more volume than the corresponding two stage reactor.

An explanation for the superiority of the tank at high conversions in Figure 5 is to be found in the exceptionally high activation energies used in the kinetics. Under these conditions the heat feed-back in the tank more than compensates for the mass action loss due to the mixing.

Note: the inlet temperature to the adiabatic tank (M) was freely adjustable as was the inlet temperature to the tube (N = 1) and that these were not forced to be the same as has been done in other studies, e.g. Reference 26.

Note also that the tank plot (M) is an exact straight line on our transformed coordinates.

III. OPTIMUM PROBLEMS INVOLVING MORE THAN ONE STOICHIOMETRICALLY INDEPENDENT REACTION OR CATALYST DETERIORATION

A. A Reaction in which the Desired Product Decays

Consider a reaction system in which the desired product is formed by an exothermic reversible reaction and decays by an irreversible reaction. The simplest case of such a system is represented by

$$A \rightleftharpoons B \rightarrow C \quad .$$

where A is the raw material, B is the desired product and C is a waste material. The fact that there are now two stoichiometrically independent reactions instead of one causes some difficulties in the proper formulation of the optimum problem as well as in its solution. In the following a geometrical interpretation will be utilized to overcome these difficulties.

Suppose y_1 and y_2 are the ratios of the molar flowrates of B and C respectively to the molar feed flowrate of A. For given pressure and initial composition the reaction rates then become functions of y_1, y_2, and T. The change of composition along a tubular reaction therefore can be described by the two differential equations

$$\frac{dy_1}{dt} = r_1(y_1, y_2, T) \tag{5}$$

$$\frac{dy_2}{dt} = r_2(y_1, y_2, T) \tag{6}$$

where t is an appropriately chosen measure for the distance in the reactor. For instance, t may be the volume of the part of the reactor ~~part~~ from the inlet to the ~~point~~ section under consideration divided by the flowrate of A at the inlet.

We shall consider in this section perfect temperature control only, that is, any function T(t) which is piecewise continuous and subject to

$$T_l \leqslant T(t) \leqslant T_u$$

will be considered as a possible temperature policy. The introduction of temperature limits is in general necessary (quite apart from other physical reasons) in order to prevent the temperature from becoming negative or infinite in the optimum case.

If there is no B and C present at the inlet the boundary conditions which have to be considered together with Equations (5) and (6) are:

$$\left.\begin{array}{l} y_1 = 0 \\ y_2 = 0 \end{array}\right\} \text{for } t = 0 \tag{7}$$

For any given temperature policy T(t) the solution of Equations (5) and (6) will trace out a path in the space spanned by y_1, y_2, and t. Consider the set of all admissible policies $\{T(t)\}$. To this set there corresponds a set of reaction paths which form a region called the attainable region,[15] in the above mentioned space (see Figure 12). Each point belonging to this region is attainable by at least one admissible policy, that is, there exists an admissible T(t) such that conversions y_1, y_2 (corresponding to the first coordinates of the point) can be obtained in the time t (corresponding to the third coordinate of the point). The attainable region thus defined will, of course, lie within the stoichiometrically attainable region which is defined by:

$$y_1 + y_2 \leqslant 1 \quad ; \quad y_1, y_2 \geqslant 0 \tag{8}$$

We shall discuss at first how the attainable region can be used in order to solve special optimization problems and then we shall discuss methods to calculate the attainable region.

The attainable region depends only on the initial conditions for the given system, once the various parameter values have been fixed. The region may be used in conjunction with any objective function depending on y_1, y_2, and t. We will illustrate this by means of an example.

Consider the recycle system shown in Figure 13, with reaction described by Equations (5) and (6) carried out in a tubular reactor. A fixed amount of A enters this reactor and product B is desired. Pure A enters the system and separation may be considered complete, so that only A is present in the recycle. A mass balance over the system will yield the flows shown, assuming unit flow of A into the reactor.

In this example let us assume that the profit P per mol of A entering the reactor has to be maximized and that this profit is given by the following simple relation:

$$P = y_1 c_b - (y_1 + y_2)c_a - (1 - y_1 - y_2)c_r - t c_c \tag{9}$$

where c_a = cost of raw material/mol. feed

c_r = cost of recycling/mol. recycle

c_c = cost of reactor/unit volume

c_b = value of product/mol. B formed

t = volume/mol. of A entering

Only under very special circumstances will such an objective function be of any practical significance. However, the essential points of the following discussion apply to any objective function, however complicated, as long as the arguments of this function are y_1, y_2, and t only. From Equation (9) it follows that:

$$(C_a - C_r - C_b)y_1 + (C_a - C_r)y_2 + C_c t + C_r + P = 0 \qquad (10)$$

This equation represents a plane in the (y_1, y_2, t) space. Note that $C_a > C_r$ for recycle to be economical and $C_b > C_a$ for system to be profitable. The coefficients of y_2 and t are thus positive and that of y_1 is negative, while $(C_r + P)$ is positive. Planes of constant P are thus parallel planes and a typical such plane is represented in Figure 12. The attainable region must now be considered together with these planes to solve the optimum problem. Since P increases in the direction indicated by the arrow, the profit will be maximum on that plane which just touches the surface of the attainable region. The point of tangency will then provide the solution.

It follows that the border of the attainable region will be of special interest to the solution of optimum problems. The border can be found by means of Pontryagin's Maximum Principle.

If Equations (5) and (6) are integrated together with

$$\left.\begin{aligned} \frac{d\lambda_1}{dt} &= -\frac{\partial v_1}{\partial y_1}\lambda_1 - \frac{\partial v_2}{\partial y_1}\lambda_2 \\ \frac{d\lambda_2}{dt} &= -\frac{\partial v_1}{\partial y_2}\lambda_1 - \frac{\partial v_2}{\partial y_2}\lambda_2 \end{aligned}\right\} \tag{11}$$

and the temperature is chosen such that the expression

$$\lambda_1 v_1(y_1, y_2, T) + \lambda_2 v_2(y_1, y_2, T) \tag{12}$$

assumes a maximum with respect to T at any time t then the solution will trace out a path at the border of the region. By integrating the equations for various initial values of $\lambda_1 : \lambda_2$ and by minimizing instead of maximizing the expression given above a set of paths at the border can be calculated and the border can thus be determined.

Once the border of the attainable region is known optimum reactors can easily be calculated for any objective function. The calculation of the border itself does not require any quantitative knowledge of the economics of the process. What has to be known is the set of economica~~lly~~ly important variables which in our example are y_1, y_2, and t and the kinetics of the reaction.

Calculations have been carried out for first order kinetics with Arrhenius rate constants. In this case the functions v_1 and v_2 in Equations (5) and (6) are given by:

$$v_1 = H_1 e^{-E_1/RT}(1-y_1-y_2) - y_1\left(H_1' e^{-E_1'/RT} + H_2 e^{-E_2/RT}\right)$$
$$v_2 = H_2 e^{-E_2/RT} y_1 \tag{13}$$

It is convenient to introduce a dimensionless time t' and a dimensionless control variable, z, which will replace the temperature T as follows:

$$t' = tH_1 \left(\frac{H_1}{H_1'}\right)^{\frac{E_1}{E_1'-E_1}} \tag{14}$$

$$z = \left(\frac{H_1'}{H_1}\right)^{\frac{E_1}{E_1'-E_1}} e^{-E_1/RT} \tag{15}$$

With these transformations Equations (5)and (6) become:

$$\frac{dy_1}{dt'} = -y_1\left(z + z^{\rho} + a z^{\delta}\right) - y_2 z + z$$
$$\frac{dy_2}{dt'} = y_1 a z^{\delta} \tag{16}$$

ρ , δ , and a are given by:

$$\rho = E_1'/E_1 , \quad \delta = E_2/E_1 , \quad a = \frac{H_2}{H_1}\left(\frac{H_1'}{H_1}\right)^{\frac{1-\delta}{\rho-1}} \tag{17}$$

Adjoint equations are obtained if t is replaced in Equation (11) by t' and v_1 and v_2 are replaced by v_1' and v_2', i.e. the functions of y_1, y_2, and z on the right hand sides of Equation (16). Similarly, along a border path the expression:

$$\lambda_1 v_1' + \lambda_2 v_2' \qquad (18)$$

must be maximum with respect to z at any t'. If temperature limits are to be taken into account and z_1 and z_u are the values of z corresponding to the lower temperature limit T_1 and upper limit T_u the expression (18) must assume its maximum within the interval:

$$z_\ell \leq z \leq z_u \qquad (19)$$

Numerical integrations have been carried out for the values:

$$\rho = \frac{4}{3}, \sigma = \frac{2}{3}, a = \frac{1}{4}, z_e = 0, z_u = 10^6 \qquad (20)$$

The results are shown in isometric representation in Figure 17. The coordinate $\hat{t}$ is defined as:

$$\hat{t} = \frac{t'}{1+t'} \qquad (21)$$

By this transformation the infinite time interval

$$0 \leq t' \leq \infty$$

is represented by the finite interval:

$$0 \leq \hat{t} \leq 1$$

The lines a, b, ..., h in Fig. 17 are border trajectories obtained by integrating (16) together with its adjoint equations and under consideration of the optimum condition. Each such line corresponds to a ratio $\lambda_1 : \lambda_2$ chosen at t' = 0. All trajectories start at the point 0 and all trajectories shown in the diagram end at the point E. The lines intersecting the trajectories represent intersections of the border with planes of constant $\hat{\tau}$. If the cost of reaction volume is insignificant ($c_c = 0$ in Equation (9)) then only the border of the projection of the attainable region onto the y_1-y_2 plane is of importance. This border of the projection is identical with the intersection at $\hat{\tau} = 1$. Furthermore, this line is identical to the projection of the trajectory a onto the y_1-y_2 plane. In the example in question the y_1-y_2 projection of the attainable region does not fill out the stoichiometrically attainable region. The maximal obtainable conversion to A_2 is about 43%. If however, $\sigma > 1$ and $\rho > 1$ (the latter is necessary for an exothermic reaction) the projected region would completely fill the stoichiometrically attainable region (in the limit of very large Z_u) and the maximal obtainable conversion to A_2 would be 100%. If only the conversion to A_2 and the reactor size matters while the conversion to A_3 is not important (no utilization of A_3 or unconverted A_1 in the reactor effluent is possible) the projection of the region onto the $\hat{\tau}$-y_1 plane has to be considered. It can be seen that the border of this projection is not generated by a trajectory in contrast to the previously discussed case.

B. Exothermic Reaction with a Decaying Catalyst

Very often an exothermic reaction carried out in a tubular reactor is catalyzed by a solid catalyst, present in the form of a packing, and the activity of the catalyst decays with increasing time. In general, the rate of decay of the catalyst will depend on temperature, and possibly also on the composition of the reaction mixture, and since these are not the same at all points of the reactor an uneven decay occurs. In particular, the instantaneous rate of decay will depend on the reactor temperature at each point, so the pattern of catalyst decay at any time will depend on the complete previous history of the temperature profile in the reactor. This leads to an interesting type of optimization problem, in which the current temperature profile influences the whole future course of the reaction by leaving its imprint on the pattern of catalyst decay.

If the changes in catalyst activity are slow compared with the speed of response of the reactor to changes in imposed conditions, the departure from an instantaneous steady state is always small, and the course of the reaction is determined by an equation of the form:

$$\frac{\partial y}{\partial t} = f(x, y, T) \tag{13}$$

where t is the distance along the reactor ($0 \leqslant t \leqslant t_e$), y is the conversion, T is the temperature, and x is a variable measuring the catalyst activity.

Typically, for a reversible reaction, with first order kinetic~~s~~ behavior in both directions, f would take the form:

$$f(x, y, T) = x\left[(1-y)k(T) - y\,k'(T)\right] \tag{14}$$

The rate of decay of the catalyst activity at any point will certainly depend on the temperature, and may also depend on the composition of the reaction mixture and the activity itself, so that:

$$\frac{\partial x}{\partial \tau} = g(x, y, T) \tag{15}$$

where τ is the time.

When the temperature $T(t, \tau)$ is specified in the domain of interest

$$0 \leqslant t \leqslant t_e$$
$$0 \leqslant \tau \leqslant \tau_e$$

where τ_e is the total time between catalyst changes, Equations (13) and (15) can be solved subject to boundary conditions

$$y = y_0 \quad \text{at } t = 0 \quad (\text{all } 0 \leqslant \tau \leqslant \tau_e)$$

and

$$x = x_0 \quad \text{when } \tau = 0 \quad (\text{all } 0 \leqslant t \leqslant t_e) \tag{16}$$

where x_o is the uniform initial activity of the catalyst. The problem is then to choose T(t,τ) so as to maximize the total yield of product in the time interval (0, τ), in other words to maximize an objective function:

$$P = \int_0^{\tau_e} y(t_e, \tau)\, d\tau \tag{17}$$

A solution will be sought using a direct method of the calculus of variations and for this purpose it is necessary to consider a small change in temperature policy from T(t,τ) to T(t,τ) + δ(t,τ) and to relate the consequent change δ P to δ T. This can easily be done by introducing two new variables λ and μ defined by the differential equations:

$$\begin{aligned} \frac{\partial \lambda}{\partial t} &= -\frac{\partial f}{\partial y}\lambda - \frac{\partial g}{\partial y}\mu \\ \frac{\partial \mu}{\partial \tau} &= -\frac{\partial f}{\partial x}\lambda - \frac{\partial g}{\partial x}\mu \end{aligned} \tag{18}$$

together with the boundary conditions:

$$\left.\begin{aligned} &\lambda = 1 \text{ at } t = t_e \ (\text{all } 0 \leqslant \tau \leqslant \tau_e) \\ &\mu = 0 \text{ when } \tau = \tau_e \ (\text{all } 0 \leqslant t \leqslant t_e) \end{aligned}\right\} \tag{19}$$

it is then easy to show[29] that:

$$\delta P = \int_{t=0}^{t_e}\int_{\tau=0}^{\tau_e} \left[\mu \frac{\partial g}{\partial T} + \lambda \frac{\partial f}{\partial T}\right] \delta T\, dt\, d\tau \tag{20}$$

The gradient of P in the function space $T(t,\tau)$ is defined as:

$$P_T = \mu\frac{\partial q}{\partial T} + \lambda\frac{\partial f}{\partial T} \qquad (21)$$

Equation (20) provides the basis for a very simple computational procedure in which the rth approximation $T_r(t,\tau)$ to the optimum policy is replaced by $T(t,\tau)$ where:

$$T(t,\tau) = T_r(t,\tau) + l\,P_{T_r} \qquad (22)$$

with P_{T_r} computed from the temperature policy $T_r(t,\tau)$.

Using the temperature policy $T(t,\tau)$, a value of P can then be computed, and l is varied to maximize this value of P. If l_m is the corresponding value of l, the (r + 1)th approximation to the optimum policy is taken to be:

$$T_{r+1} = T_r + l_m P_{T_r} \qquad (23)$$

and the process is repeated.

This procedure corresponds to the gradient method widely used in the problem of maximizing a function with a finite number of variables.[28]

This procedure was tested for an example in which the reaction rate was taken to be of the form (14), with k and k' depending on temperature according to:

$$k = k_o \exp(-e_1/T) \; ; \; k' = k_o' \exp(-e_2/T)$$

while the rate of catalyst decay was given by:

$$\frac{\partial x}{\partial \tau} = - \frac{T}{T_c} x$$

where T_c is a constant. The values taken for the various constants were:

$y_o = 0.06$ (extent of reaction at inlet)

$T_c = 250°K$

$x_o = 1$

$k_o = 6900$

$k_o' = 3 \times 10^7$

$e_1 = 6000°K$

$e_2 = 10{,}000°K$

$t_e = \tau_e = 1$

(Note that the above values of t_e and τ_e result from the use of dimensionless scaled values of distance and time, and this scaling also makes k_o and k_o' dimensionless, as indicated.) Figures 14, 15, and 16 show T_2, T_4, and T_6, the temperature policies after two, four, and six ascents respectively, plotted as functions of t for $\tau = 0$ and for $\tau = \tau_e$. The temperature profiles for intermediate values of τ, not shown on the diagrams, are, as one would expect, intermediate in nature relative to these two.

The corresponding values of the objective function P are also indicated on these diagrams. The search was started from an initial temperature policy (T_o) = 600°K (all t and τ), giving P = 0.2720 and after six ascents P is increasing only slowly.

APPENDIX

Stoichiometric Parameters, Conversion, Reaction Rate Functi

Suppose that the reaction taking place in the reactor is represented by:

$$\sum_{i=1}^{i=n} \nu_i M_i = 0 \tag{A1}$$

where M_i is the chemical formula of the ith species and ν_i is the stoichiometric coefficient if M_i is a product, minus the stoichiometric coefficient if M_i is a reactant, and zero if M_i is inert, and n is the total number of species present. M_i is a reactant not present in stoichiometric excess which we will call the reference substance.

Consider a sample of the feed F_o and suppose it to be reacted completely in reverse direction, (i.e. until at least one product disappears). Suppose that in this (reacted) sample there are present α_i moles of M_i and that the sample is of such size that $\alpha_1 = 1$. The set of parameters (α_i) will be termed the basic composition of the feed F_o.

Let y be the fractional conversion of M_1 at some point in the reactor, and y_m in the main feed F_o and in the subsidiary feed F_m.

It is convenient to introduce the quantity ω as the flow rate the reference substance would have if the material in the stream were instantly converted to completion in the reverse direction.

It is clear that the composition of a stream is completely determined by the variable y and the parameters α_i, so that with a given pressure and a given set (α_i) in mind the reaction rate v, i.e., the number of moles of reference substance converted per unit time unit volume by a homogeneous reaction, may be written as a function of y and the temperature T:

$$v = v(y, T) \tag{A2}$$

In the case of a solid catalyzed fluid reaction taking place in a tube packed with fixed granular catalyst we can still use the form (A2) if we make y and T fluid phase variables (assumed constant in the fluid part of a cross section) and consider the mass flowrate per unit sectional area (which together with y and T and the catalyst characteristics will determine the rate) to be either constant or else without influence.

For the solid catalyzed fluid reaction v is based on the mass of catalyst rather than the reactor volume.

Let m be the volume of the reactor (or mass of catalyst in the case of the solid catalyzed reactor) between the main feed inlet and some plane of reference A in Figure 17. The feed addition policy may be specified by assigning a relation between w and m: w(m).

Mass Balance, Enthalpy Function, Heat Balance

It is convenient to introduce a new variable, x:

$$x = w(y - y_m) \tag{A3}$$

in terms of which we may write a differential mass balance for the reactor:

$$\frac{dx}{dm} = r\left(\frac{x}{w} + y_m, T\right) \tag{A4}$$

For a tubular reactor w is constant, and if we put t = m/w, (A5), in Equation (A4) we obtain:

$$\frac{dy}{dt} = r(y, T) \tag{A6}$$

In general, T will not be explicitly specified but must be determined from an enthalpy balance, so the relation between enthalpy composition and temperature will be required.

We will assume (as is commonly done) that the enthalpy of the reaction mixture is linear in y and T. It follows that for an adiabatic feed bypass reactor we may write the enthalpy balance:

$$\left.\begin{aligned} T &= Jy + T_m - Jy_m + \frac{C}{\omega} \\ C &= \overset{w}{\omega}_0 (T_0 - T_m) \end{aligned}\right\} \qquad \text{(A7)}$$

The numerical calculations were all made for J = 158.5°K* and for the reaction rate expression:

$$v(y,T) = H(1-y)e^{-A/T} - H'ye^{-A'/T} \qquad (1)$$

with the two separate sets of parameters shown in Table 1.

These sets of parameters have the following properties.

1) The differences A'-A are constant, corresponding to the heat of reaction for SO_2 oxidation.

2) The ratios A'/A are respectively 1.5 and 2. For these simple ratios the t_f^*, y_f relation for perfect indirect control can be obtained in closed form.[11]

3) The ratios H'/H are constant.

It follows that the equilibrium relation between y and T is the same for each set.

The second set of parameters corresponds to the experimental

*A typical figure for a lean Pyrites roast gas.

data of Schytil and Schwalb[16] for the oxidation of a lean SO_2 gas at 440°C.

Note: units for H and H' are not important because the quantities of interest in this study are the ratios of reactor volumes required in different cases and these ratios are, of course, invariant against multiplication of H and H' by the same positive constant.

Acknowledgments

This work was supported by the systems grant from the National Science Foundation (No. GU 1153)

REFERENCES

1. Leitenberger, W., Chem. Fabr. 12, 281 (1939).

2. Denbigh, K. G., Trans. Faraday Soc. 40, 352 (1944).

3. Annable, D., Chem. Eng. Sci. 1, 145 (1952).

4. Temkin, M. and V. Pyzhev, Acta Physchim. U.R.S.S. 12, 327-356 (1940).

5. Bilous, O. and N. R. Amundson, Chem. Eng. Sci. 4, 81-92, (1956).

6. Calderbank, P. H., Chem. Eng. Prog. 49, 585-590 (1953).

7. Boreskov, G. H. and M. G. Slinko, Chem. Eng. Sci. 14, 249 (1961).

8. Katz, S., Ann. N. Y. Acad. Sci. 84, 443 (1960).

9. Aris, R., Chem. Eng. Sci. 13, 197 (1961).

10. Horn, F. and L. Kuchler, Chem. Ing. Tech. 31, 1 (1959).

11. Horn, F., Chem. Eng. Sci. 14, 77 (1961).

12. Horn, F., Z. Elektrochem. 65, 209 (1961).

13. Horn, F., Doctoral Thesis, Technical University, Vienna, (1958).

14. Dyson, D. C., Ph.D. Thesis, University of London, 1966.

15. Horn, F., Chem. Eng. Sci. 20, 293-303 (1965), supplement.

16. Schytil, F. and H. Schwalb, Chem. Eng. Sci. 14, 367-373 (1961).

17. Konoki, K., Kagaku Kogaku 24, 569-71 (1960).

18. Aris, R., The Optimal Design of Chemical Reactors, Academic Press, 1961.

19. Kung-You Lee and R. Aris, Ind. Eng. Chem. Proc. Des. Devel. 2, (4), 300-306 (1963).

20. Horn, F., Chem. Eng. Sci. 14, 20-21 (1961).

21. Bakemeier, H., H. Detzer, and R. Krabetz, Chim. Ing. Tech. 37, 429-433 (1965).

22. van Heerden, C., Ind. Eng. Chem. 45, 1242 (1953).

23. Powell, M. D., and R. Fletcher, The Computer Journal 6, 183 (1963).

24. Davidon, W. C., Argonne Natl. Lab. Report ANL 5990 (Rev.).

25. Kramers, H. and K. Westerterp, Chemical Reactor Design and Operation, Netherlands U. Press, 1963.

26. Aris, R., Can. J. Chem. Eng. 40, 87-92 (1962).

27. Schlesinger, C. B., Doctoral Thesis, University of London, 1966.

28. Optimization Techniques with Application to Aerospace Systems, ed. G. Leitmann, Academic Press, Chapter VI.

29. Jackson, R., Joint I. Chem. E., A.I.Ch.E. Symposium, London, June 1965, Proceedings of Section 4.

30. D. C. Ryan and F. J. M. Horn. To appear in Journal of Optimization Theory and Applications.

NOMENCLATURE

(Defining Equations in Brackets)

a	See (17).
f	Reaction rate, (14).
g	Catalyst activation function, (15)
l	See (22) and following text.
m.	See text between (A2) and (A3).
t	Measure of distance from entrace of reactor. See (A5), (5), and (13) and adjacent texts.
t'	See (14).
$\hat{t}$	See (21).
v	Reaction rate. See (A2), (1), (5), and (6).
W → ω	Reference flow. See text between (A1) and (A2).
x	Catalyst activity, (15), (16) and text below (23).
y	Fractional conversion of reference reactant. See Appendix.
y_1, y_2	See text above (5).
z	See (15).
C	See (A7).
C_a, C_b, etc.	See Text below (9).
E	Activation energy.
H, H'	Parameters in (1). See Table 1 and last paragraph of Appendix.
H_1, H_1', etc.	Parameters in (13).
J	Adiabatic temperature rise coefficient, (A7).
N.	Number of stages.
P	Profit, (9); (17).
P_T	Gradient of P in function space, (21).
R	Gas constant.
T	Temperature, degrees Kelvin

λ	Adjoint variable, (18).
λ_1,λ_2	Adjoint variables, (11).
μ	Adjoint variable, (18).
ρ	See (17).
ld $\rightarrow$ δ	See (17)
na τ	Time, see text below (15).

Subscripts(apply to Section II only)

a_j	Referring to entrance to stage j.
e_j	Referring to exit of stage j.
f	Referring to exit of reactor.
m	Referring to subsidiary feed: See Fig. 10.
0	Referring to the main entrance to the reactor.

Titles for Figures

Figure 1: Attainable region for a single reaction with stable catalyst.

Figure 2: Reaction rate contours and optimal trajectories for perfect control in the y, T plane.

Figure 3: Perfect control trajectories in the m, y plane.

Figure 4: Comparison of N-stage directly controlled adiabatic stage reactors.

Figure 5: Comparison of indirectly controlled reactors.

Figure 6: U-Tube reactor.

Figure 7: Counter Current heat exchange reactor.

Figure 8: Modified counter current heat exchange reactor.

Figure 9: N-stage directly controlled adiabatic stage reactor.

Figure 10: Distributed feed reactor.

Figure 11: Push-pull reactor in push mode.

Figure 12: Attainable region (hill) and plane of constant objective function (triangle). Arrow indicates direction of gradient of objective function.

Figure 13: Reactor-separator system.

Figure 14: Temperature profile for time $\tau = 0$ and $\tau = \tau_e$ after 2nd ascent.

Figure 15: Temperature profile for time $\tau = 0$ and $\tau = \tau_e$ after 4th ascent.

Figure 16: Temperature profile for time $\tau = 0$ and $\tau = \tau_e$ after 6th ascent.

Figure 17: Attainable region in y_1, y_2, $\hat{t}$ space. Lines a - h are optimal trajectories.

TABLE 1

Reaction rate and thermochemical parameters for equation (1)

	1st set	2nd set
	8.7678 x 10^{14}	1.7536 x 10^{7}
	1.2442 x 10^{30}	2.4835 x 10^{12}
	2.2748 x 10^{4}	1.1374 x 10^{4}
	3.4122 x 10^{4}	2.2748 x 10^{4}
	158.5	158.5

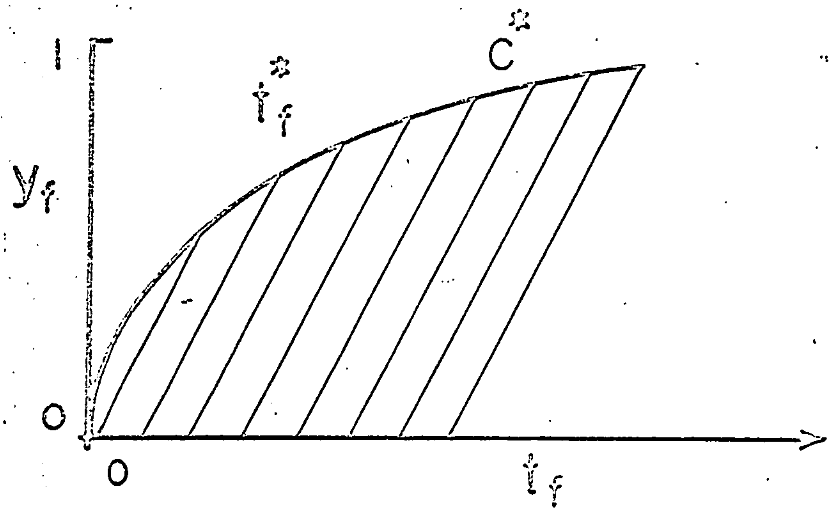


FIGURE 1

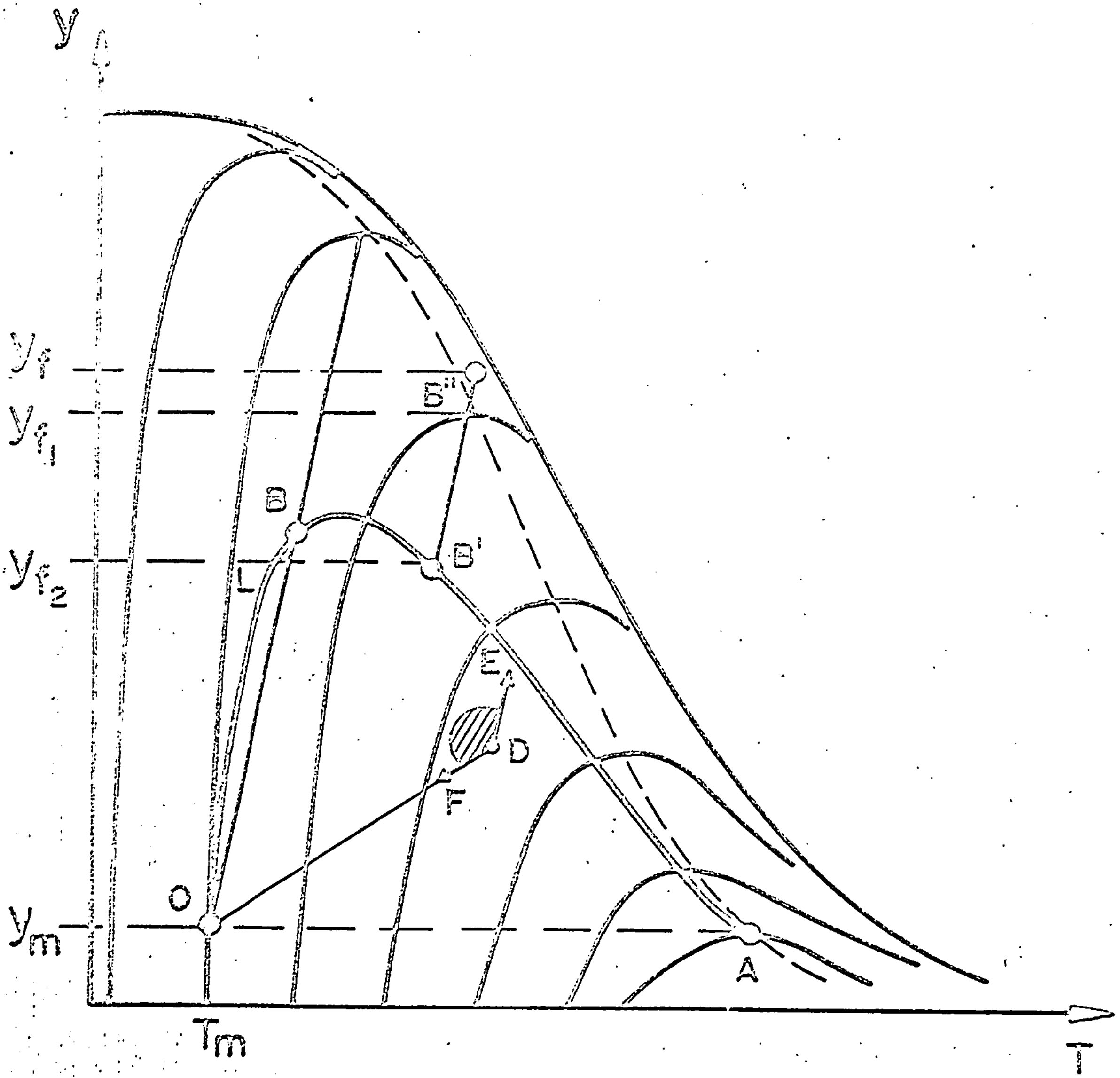


FIGURE 2

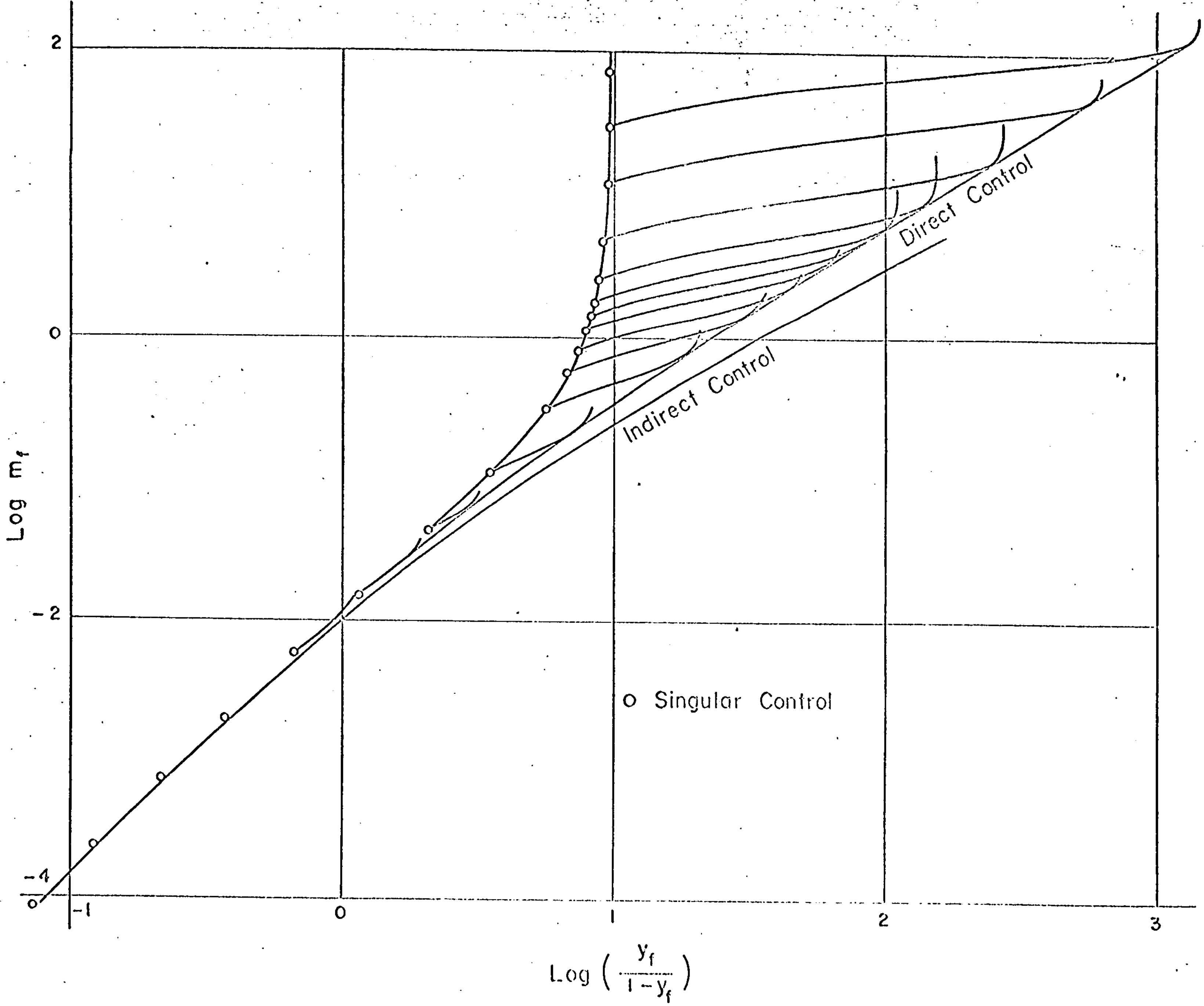

2
0
-2
-4
-1
0
1
2
3
Log m_f
Log $\left(\frac{y_f}{1-y_f}\right)$
Direct Control
Indirect Control
o Singular Control

FIGURE 3

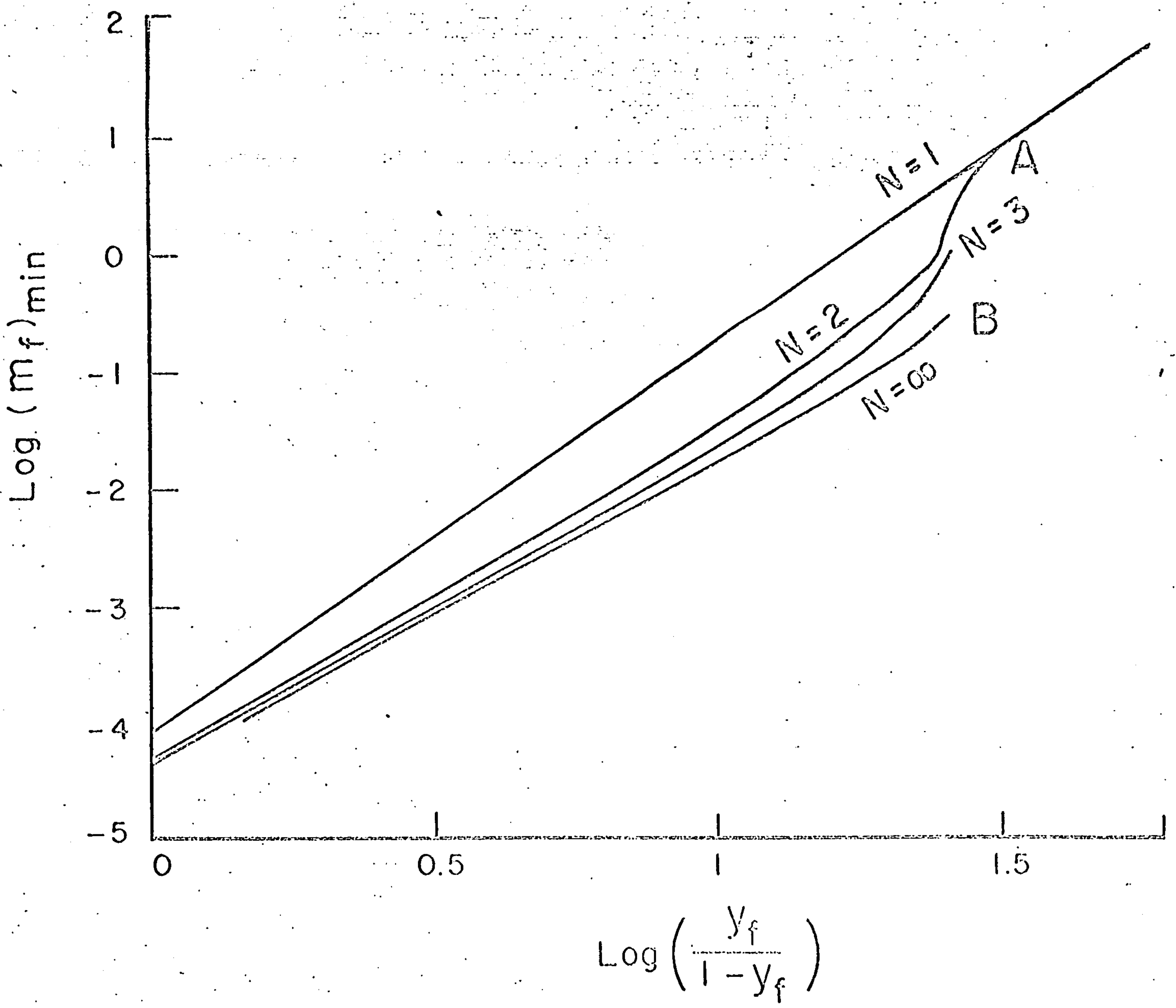
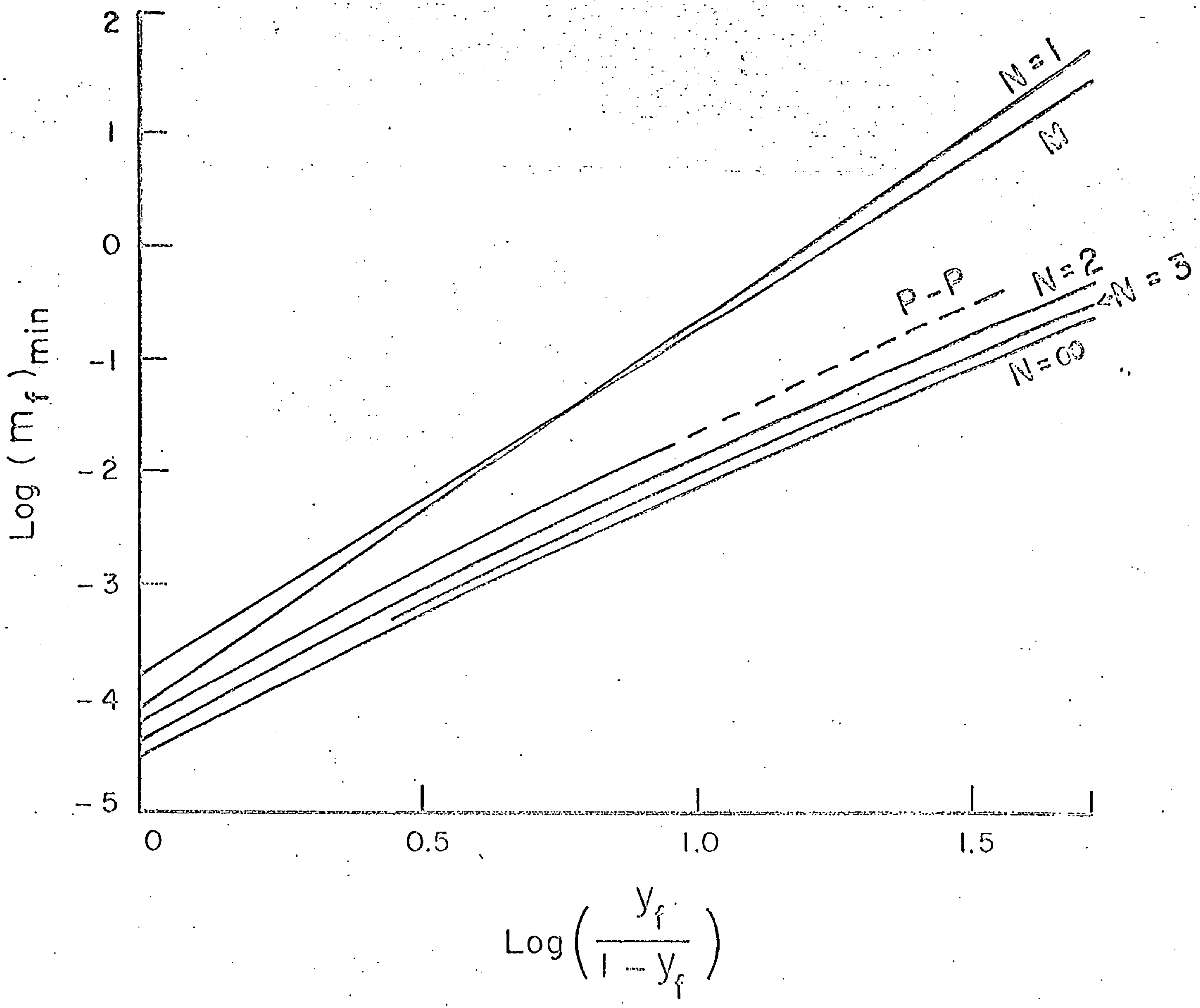

2
1
0
-1
-2
-3
-4
-5
Log $(m_f)_{min}$
0
0.5
1
1.5
Log $\left(\frac{y_f}{1-y_f}\right)$
N=1
N=2
N=3
N=∞
A
B

FIGURE 4



Log $(m_f)_{min}$
2
1
0
-1
-2
-3
-4
-5
0
0.5
1.0
1.5
Log $\left(\frac{y_f}{1-y_f}\right)$
N=1
M
P-P
N=2
N=3
N=∞

FIGURE 5

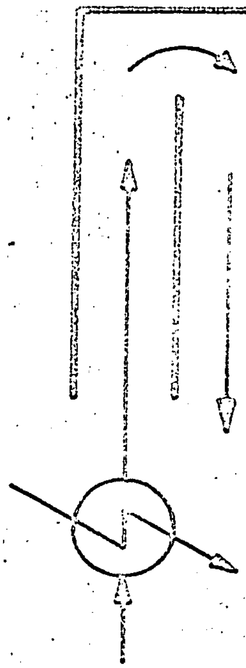


FIGURE 6

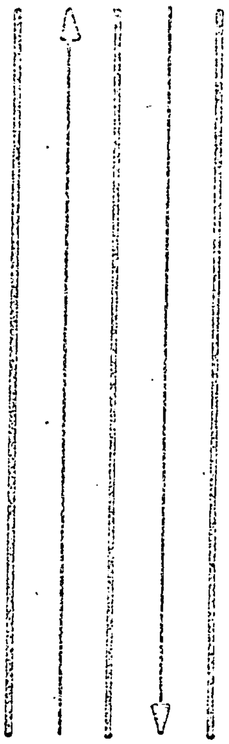


FIGURE 7

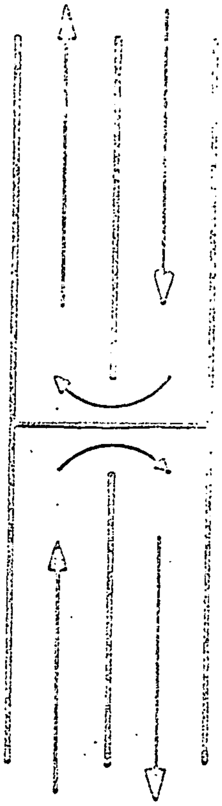


FIGURE 8

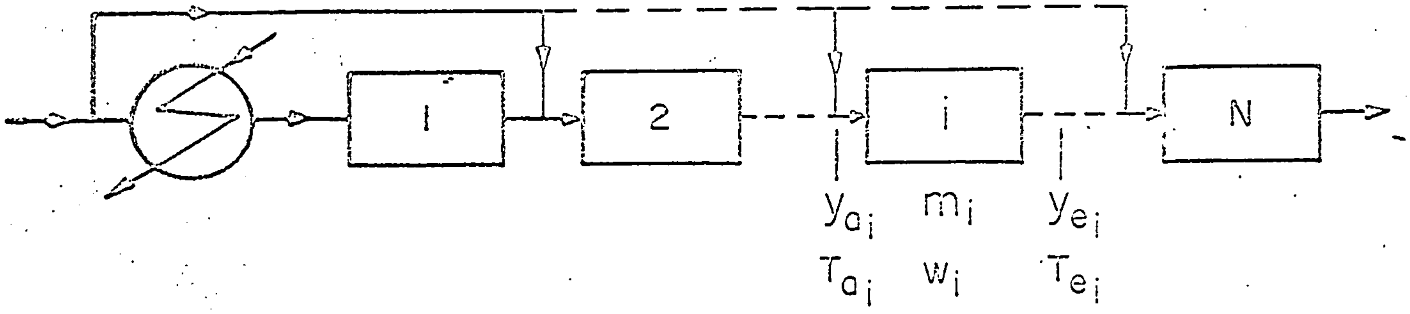


FIGURE 9

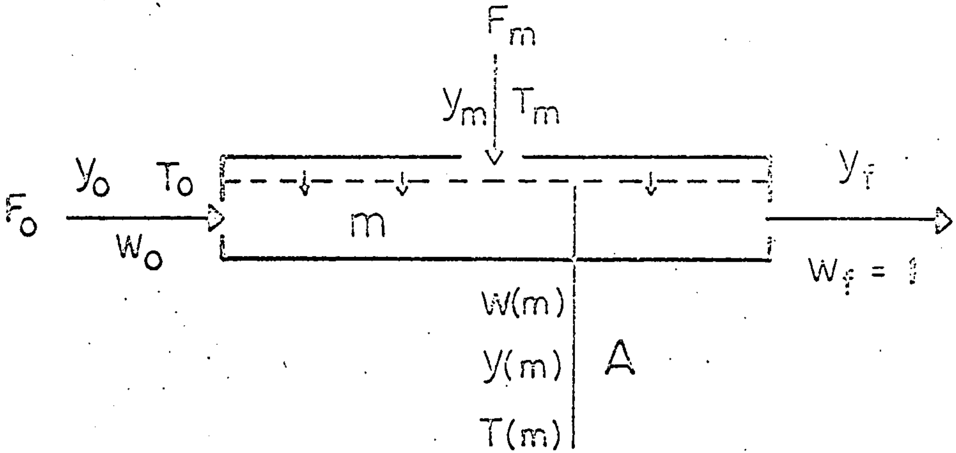


FIGURE 10

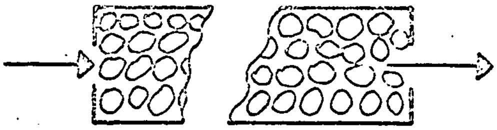


FIGURE 11

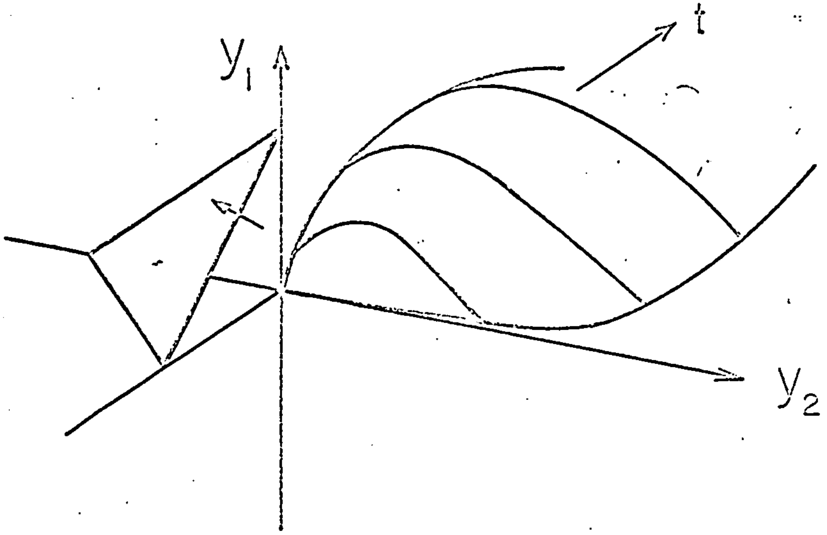


FIGURE 12

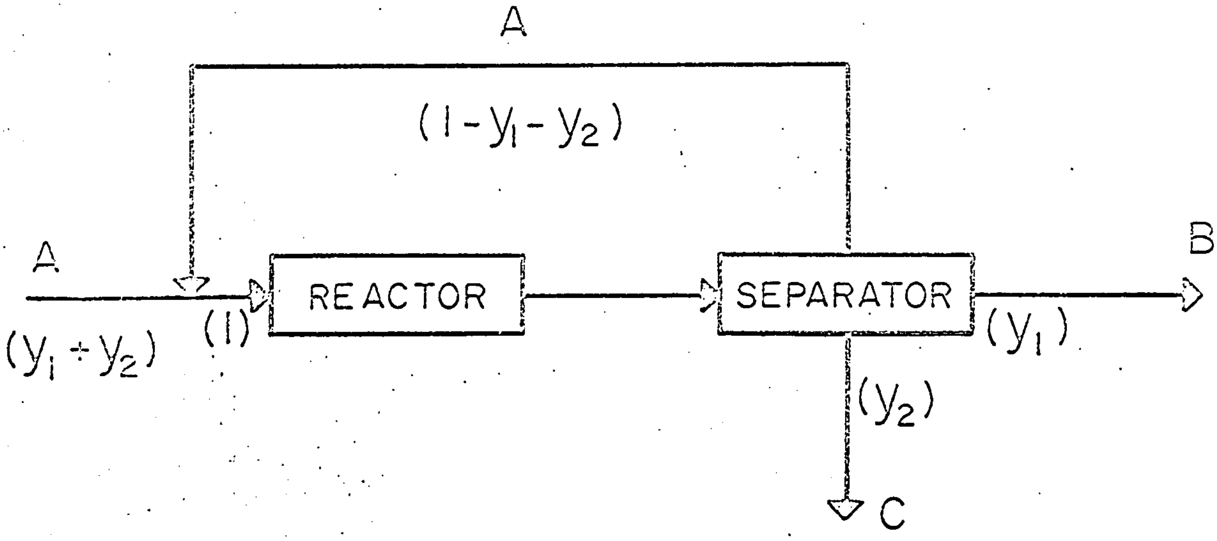


FIGURE 13

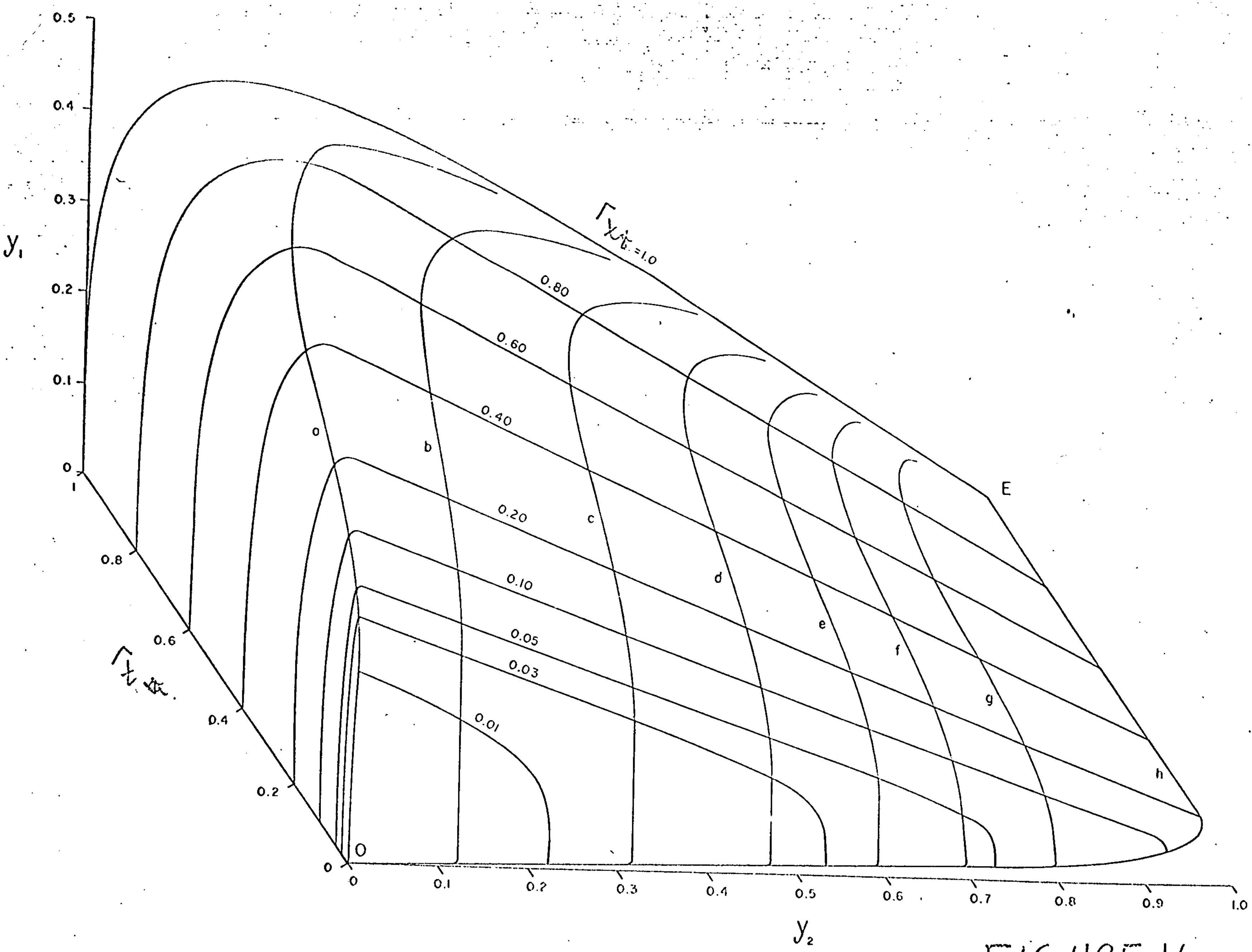

0.5
0.4
0.3
0.2
0.1
0
y_1
$\Gamma_{yt} = 1.0$
0.80
0.60
0.40
0.20
0.10
0.05
0.03
0.01
a
b
c
d
e
f
g
h
E
O
1
0.8
0.6
0.4
0.2
0
$\Gamma_{t\,st}$
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
y_2

FIGURE 14

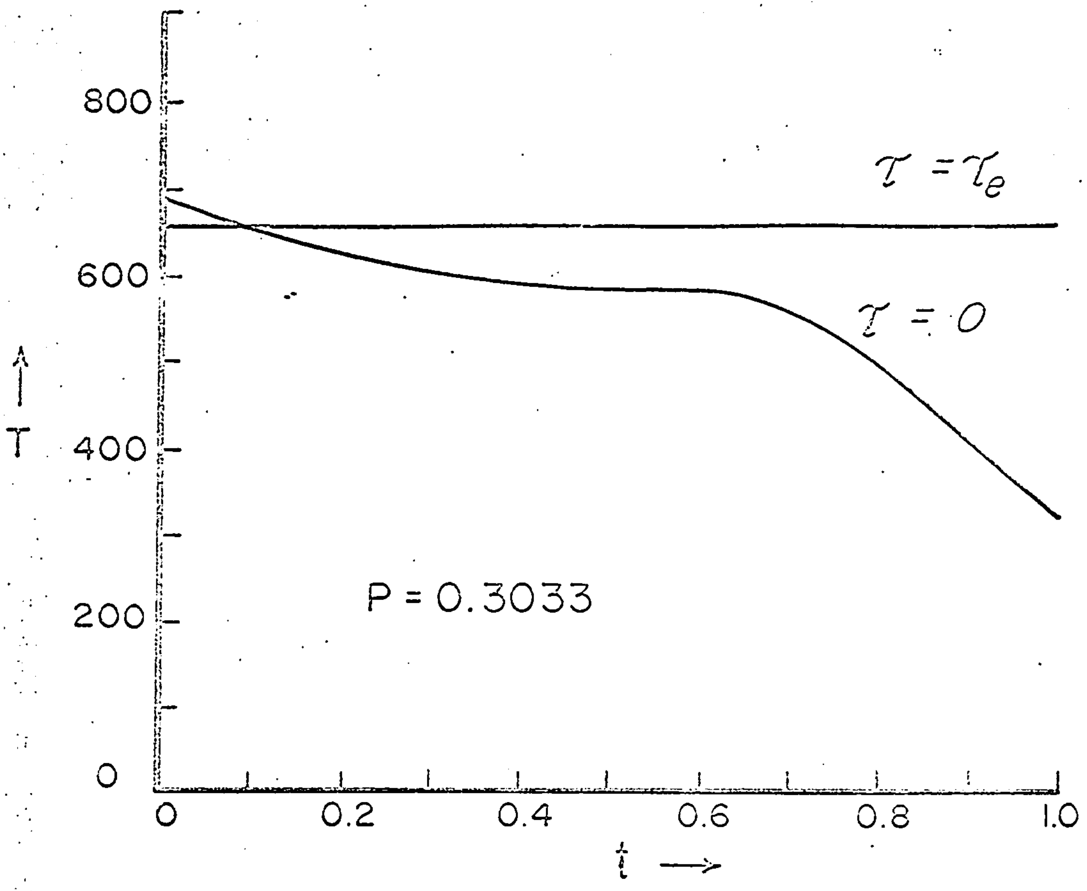


FIGURE 15

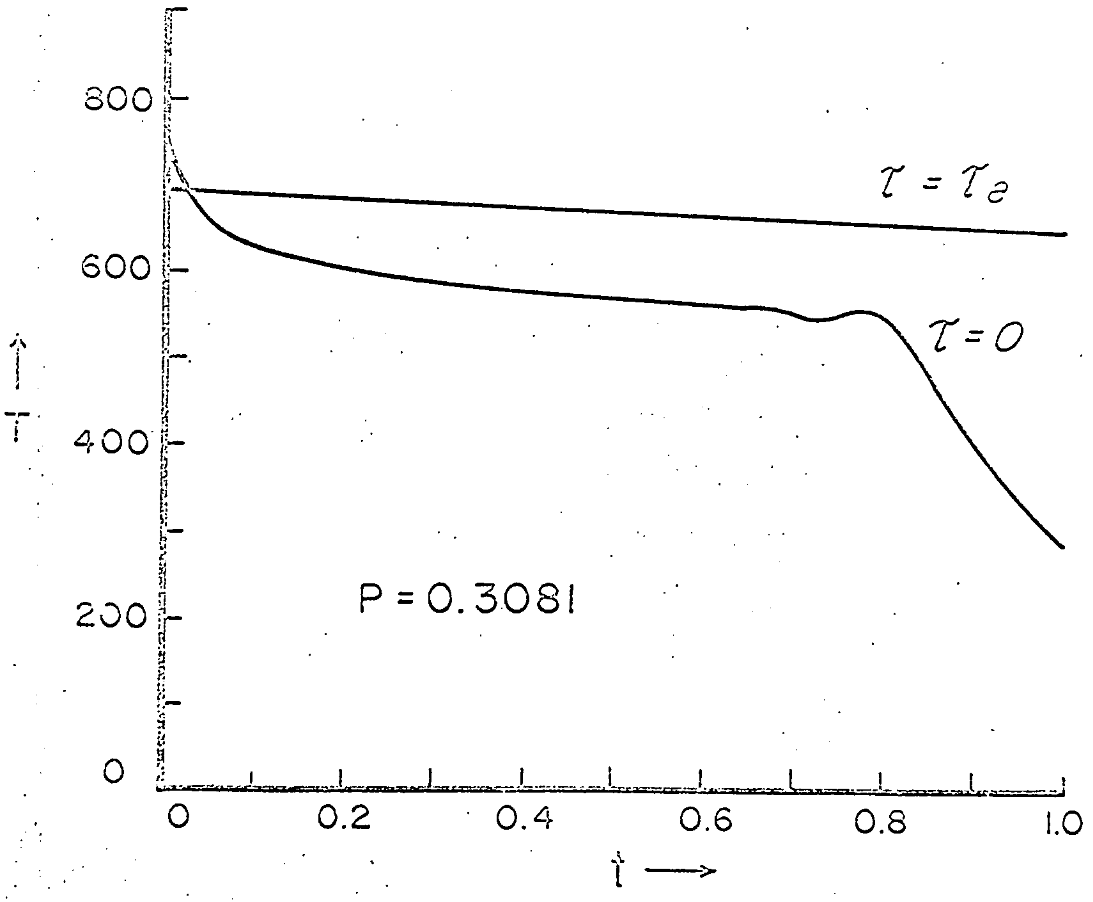


FIGURE 16

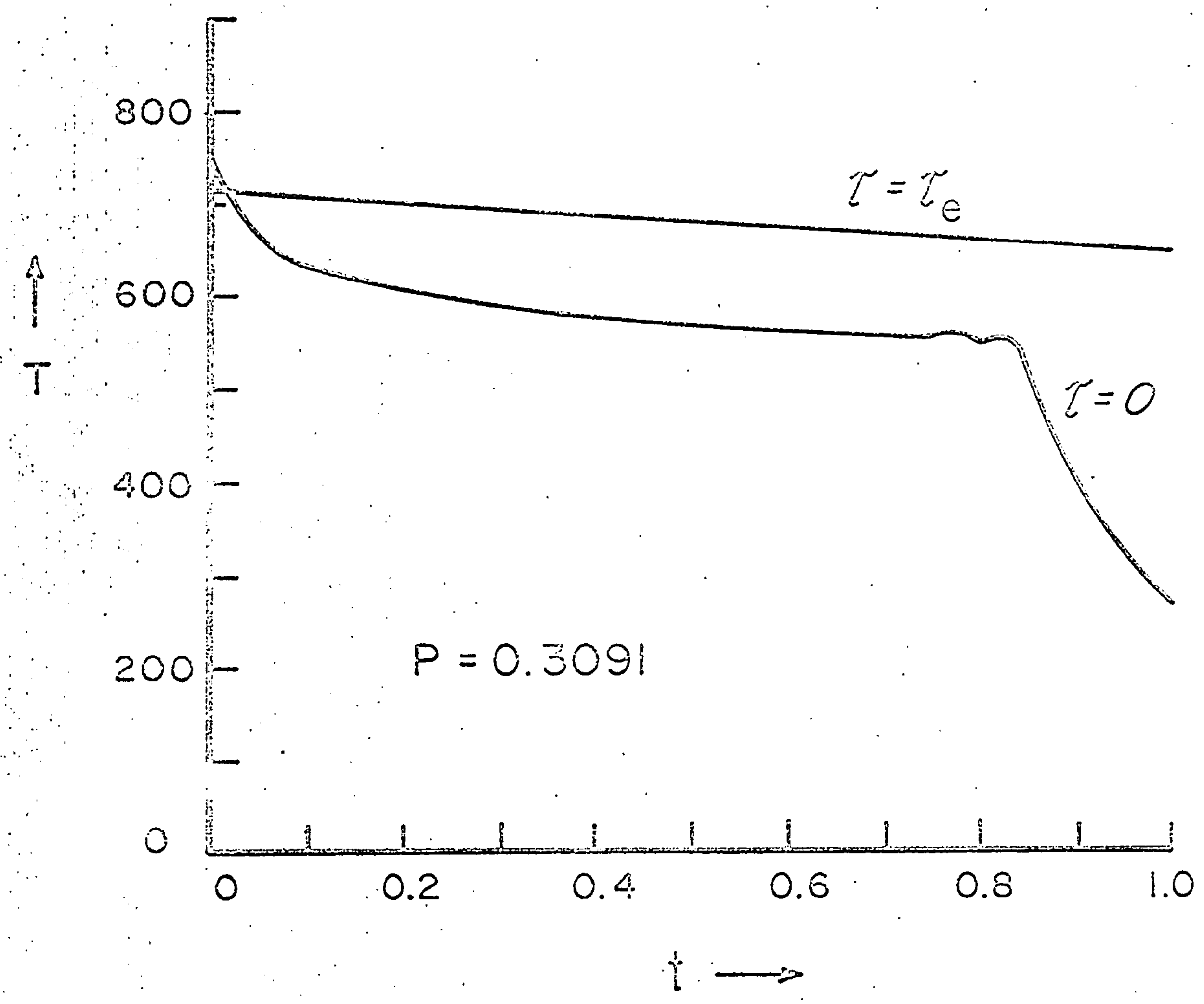


FIGURE 17

Group D

A celebrated theorem of chemical thermodynamics determines the minimum energy required to separate a given mixture of substances into a number of other mixtures. The case of greatest practical interest is that in which the final mixtures approximate to the pure components themselves, in which case the problem may be described as the determination of the minimum energy of separation. The minimum is attained only when the separation is reversible, of course, and can therefore only be approximated by any real separation device. Publication D1 shows how a binary distillation column can, in principle, be operated reversibly in such a way as to attain the thermodynamic minimum separation energy, and some practical schemes for improving the energy economy are based on this analysis. These schemes form the basis of a patent granted in the names of the authors of the publication.

When the mixture distilled contains more than two components it is not possible, even in principle, to attain the thermodynamic minimum energy of separation in a distillation column. Nevertheless there is still a lower bound on the energy of separation by distillation, even though this is higher than the thermodynamic bound. The problem of determining this bound for the distillation of multicomponent mixtures still remains unsolved, to the writer's knowledge.

ENERGY REQUIREMENTS IN THE SEPARATION OF MIXTURES BY DISTILLATION

By J. R. FLOWER, Ph.D.,* and R. JACKSON, M.A.†

SYNOPSIS

The reversible operation of a continuous distillation column is briefly described and it is shown how closely this can be approximated in principle, both using heat pumps with an independent heat-transfer medium and using direct compression of the overhead vapour. The corresponding systems are impractical idealisations but have obvious practicable analogues which are discussed in more detail; quantitative illustrations are given. In particular, it is shown that systems may be designed, making direct use of compressed overhead vapour, which are economically attractive in certain cases.

Introduction

It has long been recognised that the process of continuous distillation, as normally carried out in an adiabatic column, is highly irreversible and consequently the energy requirement greatly exceeds the theoretical minimum demanded by the thermodynamic properties of the feed and product streams. When the energy is provided in the form of mechanical work, as in low-temperature gas separations, there is a thermodynamically determined minimum amount of work required to accomplish the separation. As early as 1923, van Nuys[1-5] gave a very thorough discussion of this point and a few years later Dodge and Housum[6] analysed in detail a number of systems for separating the main components of air. Subsequently Hausen[7] proposed a model for an ideal distillation system which would, in principle, achieve the limiting performance thermodynamically permissible for a binary separation. In later years a large number of papers and patents relating to low-temperature gas separation have appeared, and it will suffice to mention papers by Haselden[8,9] and the book by Ruhemann[10] which give many references.

In conventional distillation systems operating above ambient temperature, where energy is supplied in the form of heat rather than mechanical work, the approach has, on the whole, been less systematic than in the field of gas separation, though many methods of reducing heat consumption have been described and patented.[11]

It is the purpose of the present paper to describe an ideal distillation system somewhat different in structure from that of Hausen,[7] to discuss the energy losses which necessarily accompany various approximate physical realisations of this system, and hence to arrive at practically realisable systems which secure a significant proportion of the energy economies associated with the ideal system at reasonable capital cost.

Reversible Distillation

In its most general form a distillation column separates a feed stream into a number of product streams, in the course of which heat is exchanged with a number of reservoirs at different temperatures and mechanical work is performed on the system. If Q_i is the heat absorbed from a resevoir at temperature T_i ($i = 1, 2..N$) and W is the mechanical work performed on the system during the distillation of a quantity F of the feed, it is easy to show from the second law of thermodynamics that:

$$\sum_{i=1}^{N} Q_i/T_i \leqslant \sum_{p=1}^{P} G_p S_p - FS_F = \Delta S \quad . \quad . \quad (1)$$

where G_p is the quantity of the pth product corresponding to a quantity F of feed, S_p is the unit entropy of this product and S_F is that of the feed.

A straightforward enthalpy balance also gives:

$$\sum_{i=1}^{N} Q_i + W = \sum_{p=1}^{P} G_p H_p - FH_F = \Delta H \quad . \quad (2)$$

where H_p and H_F are the unit enthalpies of products and feed respectively.

With a heat pump or vapour compression system, heat is exchanged with only a single reservoir (at temperature T, say) and work is done on the system by a compressor. Equations (1) and (2) then degenerate into the familiar expression giving the minimum work of separation, namely:

$$W \geqslant \Delta H - T\Delta S \quad . \quad . \quad . \quad (3)$$

In conventional distillation, on the other hand, no mechanical work is performed on the system and equations (1) and (2) yield rather less familiar lower limits for the heat absorbed at the reboiler, Q_B, and the heat rejected at the condenser, Q_D, namely:

$$Q_B \geqslant (\Delta H - T_D\Delta S)/(1 - T_D/T_B) \quad . \quad (4)$$

and:

$$Q_D \geqslant (\Delta H - T_B\Delta S)\Big/\left(\frac{T_B}{T_D} - 1\right) . \quad . \quad (5)$$

where T_B and T_D are the temperatures in the boiler and condenser respectively.

The limits imposed on the performance by equations (1) to (5) are attainable only by strictly reversible systems, but a conventional adiabatic distillation is far from reversible since on each plate heat and mass transfer occurs between a liquid and a vapour with which it is not in equilibrium.

In the case of a binary distillation it was shown by van Nuys,[1-5] and by Dodge and Housum,[6] that it is possible to arrange that the liquid and vapour are everywhere in equilibrium by varying the reflux flow along the length of the column in such a way that it operates everywhere in a "pinched" condition. This can be accomplished by supplying or removing heat in the necessary amounts and leads, of course, to a column of infinite capital cost, with an infinite number of theoretical stages. It has frequently been assumed[9] that the same is true for separations involving more than two components but, as has been pointed out by Timmers and by Beek,[12,13] variation of the heat exchange between the column and its surroundings does not provide sufficient disposable parameters to ensure equilibrium with respect to all components in the liquid and the vapour at all points of the column, so multicomponent distillations cannot be rendered reversible in this way.

* Department of Chemical Engineering, Houldsworth School of Applied Science, University of Leeds.

† Chemical Engineering Laboratories, University of Edinburgh and Heriot-Watt College, Chambers Street, Edinburgh 5.

For a binary mixture the necessary distribution of heat exchange along the column can be calculated very simply if an enthalpy-composition diagram is available. This is illustrated by Fig. 1, in which f, d, and b represent the feed, distillate, and bottom product respectively, all three being liquids at their bubble points in this case. The distance dd′ gives Q_{DO}/D, where Q_{DO} is the heat to be removed in the condenser to condense the flow D of distillate. If l and m represent liquid and vapour in equilibrium at some point above the feed and n is the point in which lm produced meets the vertical through d, the distance nd′ gives $\Delta Q'/D$, where $\Delta Q'$ is the additional amount of heat, over and above Q_{DO}, which must be removed between the top of the column and the point where the liquid has the composition represented by point l. By choosing various positions for l between d and f, it is therefore possible to plot the heat to be removed above any point as a function of the liquid composition (or equivalently the temperature) at that point. Similarly, if t represents the liquid at a point below the feed, the distance yv gives the heat to be supplied to the column between the feed and this point, per unit of bottom product, and again this quantity can be plotted as a function of liquid composition (or equivalently temperature) in the part of the column below the feed.

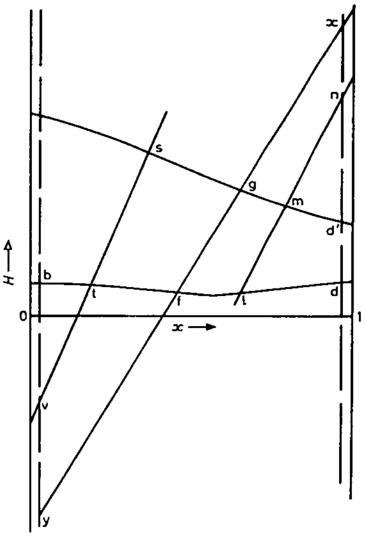


Fig. 1.—*H–x diagram, illustrating the construction of the Q-curve*

The two curves obtained as just described can be combined into a single curve showing the net amount of heat to be supplied to the column (including the condenser) above any point as a function of the temperature at the point in question. This will be denoted by $Q(T)$ and the resulting curve, which will be called the Q-curve, will prove very useful in dealing with thermal effects accompanying the distillation. Fig. 2 shows a Q-curve, constructed in this way, for the separation of a methanol–water mixture at its boiling point, containing 90% by weight of methanol, into a distillate containing 99·95% methanol and a bottom product containing 1% methanol. It is based on the enthalpy-composition diagram and equilibrium data of Plewes, Jardine, and Butler[14] for a pressure of one atmosphere and is typical of the simplest form which these curves can take.

Enthalpy-composition diagrams are, of course, seldom available in practice, and one must then derive the form of the Q-curve from the same approximate assumptions as lead to the McCabe–Thiele construction with straight operating lines. With these assumptions, and a suitable choice of the units used in expressing the composition, it is possible to arrange that the latent heat of vaporisation per unit of liquid takes a constant value L, independent of composition, and the thermal quantities which determine the form of the Q-curve can be obtained from the following equations:

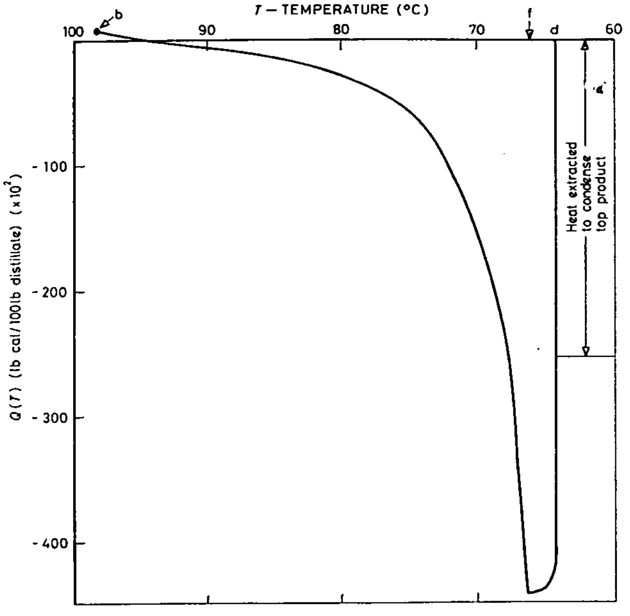


Fig. 2.—*Q-curve for the methanol–water system described*

$$s_1 = \frac{\Delta Q'/LD}{1 + \Delta Q'/LD} \quad . \quad . \quad . \quad (6)$$

and:

$$s_2 = \frac{1 + \Delta Q/LB}{\Delta Q/LB} \quad . \quad . \quad . \quad (7)$$

where s_1 is the slope of a line joining the point (x, y) (above the feed) on the equilibrium curve in the x-y plane to the point (x_D, x_D) representing the composition of the distillate, and s_2 is the slope of a corresponding line joining a point (x, y) on the equilibrium curve below the feed to the point (x_B, x_B) representing the bottom product. $\Delta Q'$ is the heat to be removed from the column between the top and the point above the feed where the liquid composition is x, and ΔQ the total amount of heat to be supplied to the system (including the boiler) below the point between the feed and the bottom of the column where the liquid composition is x. B and D are the flows of bottom product and distillate respectively. The construction is illustrated in Fig. 3.

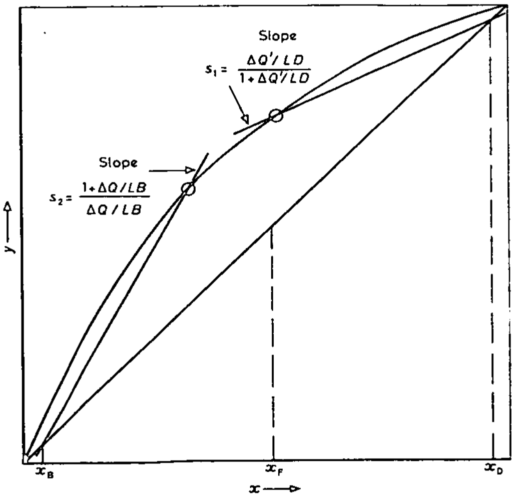

Slope
$s_1 = \frac{\Delta Q'/LD}{1+\Delta Q'/LD}$
Slope
$s_2 = \frac{1+\Delta Q/LB}{\Delta Q/LB}$
y →
x_B
x_F
x_D
x →

Fig. 3.—*Construction of the Q-curve from a McCabe–Thiele diagram*

Examination of Fig. 2 shows that reversible operation does not, in itself, reduce the total amount of heat to be supplied to the column, which is equal to the difference in ordinates of the Q-curve, $Q(T_B) - Q(T_F)$, whether it is all supplied at the boiler, as in an adiabatic column, or is distributed to make the system reversible. Reversible operation therefore permits part of the heat supplied to be of lower grade than that required in the boiler, but when the only available source of heat is a medium hotter than the boiler temperature, there is no advantage in distributing the heat supply. However, since the majority of the heat supplied in the lower part of the column is subsequently rejected to a cooling medium in the upper part or the condenser, there may be a considerable economic advantage in operating some form of heat-pumping system. Such a system was described by Hausen,[7] who used a compressor to deliver a hot heat-transfer medium to the boiler. The medium was then reduced in pressure through a sequence of expansion engines, passing through exchangers on the plates of the column between successive expansions, and finally returning to the suction of the compressor after serving to condense the distillate at the top of the column. By suitable adjustment of the expansion ratios it was possible to distribute the heat supply to plates below the feed and the heat removal from plates above the feed in such a way as to make the column reversible. At the same time, of course, the number of plates, and correspondingly the number of expansion engines, became infinite. Hausen's arrangement is of course impracticable for the reason common to all reversible devices, the necessity for an infinite amount of equipment (in this case distillation plates and expansion engines), but it suffers from the further disadvantage that even finite approximations to it are hardly practicable, as expansion engines working through the small pressure drops involved are not a practicable means of recovering energy to offset the work performed in the main compressor. In fact, Hausen's arrangement is an unnecessarily complicated way of using mechanical work to realise a reversible system, and we shall describe a simpler arrangement using no expansion engines which has the further advantage that its finite approximations are practically realisable systems.

Conventional heat-pumping systems for distillation columns operate between the condenser as source and the reboiler as sink, either directly by compression of the overhead vapour or indirectly using a secondary heat-transfer medium, and systems of both types have been extensively described[11] and have formed the basis of many patents. However, inspection of the Q-curve in Fig. 2 immediately suggests that this type of system can be greatly improved, since the major part of the thermal requirements of the system can be met by operating heat pumps through quite small temperature differences between sources above the feed point and sinks below it, corresponding to the steep sides of the deep, narrow valley in the Q-curve. Only a very small proportion of the heat to be supplied to the lower part of the column need be pumped through temperature differences approaching $(T_B - T_D)$, and correspondingly the work consumption of the heat pumps using the distributed sources and sinks required for reversible operation is much smaller than that of a single pump operating between the condenser and reboiler. Furthermore, the reversible mode of operation crowds the heat supply to the column and the heat removal from the column as closely as possible below and above the feed point respectively; any attempt to supply or remove more heat than the quantity corresponding to reversible operation in a given interval adjacent to the feed would lead to a "pinch", and would prevent the specified product compositions being obtained. A heat-pumping system using a minimum amount of work is therefore obtained by linking infinitesimal sources and sinks above and below the feed by an infinite array of heat pumps, each driven by a compressor. The individual sources and sinks can be linked in pairs in an infinite number of different ways; for example sources and sinks at successively increasing temperatures may be linked together or, as indicated in Fig. 4, the coolest source may be linked to the hottest sink and all the other sources and sinks connected in reverse sequence. However, the particular pattern of connection is of no importance since it can easily be shown that the total work of heat pumping is the same for all patterns.

Any system of this type therefore operates, in principle, with the minimum work for separation as given by the inequality (3) and, unlike Hausen's arrangement, requires no expansion engines.

Let dQ be the amount of heat to be supplied to a small element of the column about the temperature T and assume that it is made available by a heat pump which extracts an amount of heat dQ' from an element of the column about the temperature T'. Then, assuming the heat pump is reversible, we have:

$$dQ' = \frac{T'}{T}\, dQ \quad . \quad . \quad . \quad . \qquad (8)$$

and:

$$dW = \frac{T - T'}{T}\, dQ \quad . \quad . \quad . \qquad (9)$$

where dW is the work done in the heat pump. Equations (8) and (9) are differential equations for T' and W as functions of T, since Q' is given in terms of T' and Q in terms of T by the ordinates of the two branches of the Q-curve, one on each side of the feed point. W is the total amount of work performed in the heat pumps supplying heat to the temperature interval $T_F \rightarrow T$ in the column. The equations may be integrated outwards in the directions of increasing T and decreasing T' from the feed point, starting with $W = 0$ and $T' = T_F$ when $T = T_F$. This actually corresponds to the pattern of connection of sources and sinks indicated in Fig. 4.

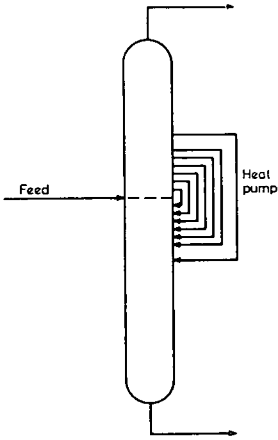


Fig. 4.—*Optimum heat-pumping system*

The integration may terminate either when Q reaches the value $Q(T_B)$, leaving a finite residue of heat to be removed at the condenser, or when Q' reaches the value zero, leaving a finite residue of heat to be supplied to the boiler. Which of these occurs in any given case is determined by the sign of the entropy change for the overall separation process.

In certain circumstances it may be economic to terminate the sequence of heat pumps at some temperature T lower than T_B even if Q' has not reached the value zero. The cost per unit of mechanical work is greater than that of heat, and if the cost of dW units of work exceeds that of dQ units of heat, where dW and dQ are related by equation (9), it is clearly uneconomic to go further with the heat-pumping sequence. Since the ratio dW/dQ increases with T, there will then be some temperature at which the heat-pumping sequence should be terminated to give minimum total energy costs, namely the temperature at which the costs of dW and dQ are equal. In most cases the temperature difference between condenser and reboiler is not large enough to satisfy this condition and heat pumping should therefore be employed to the fullest possible extent.

Another interesting approach to the problem of energy economy is to replace heat pumps using mechanical work by pumps which consume heat from a reservoir at the boiler temperature T_B; for example, absorption refrigerators. In this arrangement the system consumes heat from a single reservoir at temperature T_B however complicated a system of heat pumps is used, and a rational comparison of energy economies can be made without raising the question of the relative costs of heat and mechanical work. If the thermal heat pumps are reversible, and if dq is the heat absorbed from the reservoir at T_B by the heat pump operating between a source in the neighbourhood of temperature T' and a sink in the neighbourhood of temperature T, the equations corresponding to (8) and (9) are:

$$\mathrm{d}Q' = \frac{T'}{T}\left(\frac{T_B - T}{T_B - T'}\right)\mathrm{d}Q \quad . \quad . \quad (10)$$

and:

$$\mathrm{d}q = \frac{T_B}{T}\left(\frac{T - T'}{T_B - T'}\right)\mathrm{d}Q \quad . \quad . \quad (11)$$

These may be integrated outwards from the feed in the same way as equations (8) and (9), and the integration terminates either when Q reaches the value $Q(T_B)$ or when Q' reaches the value zero. The corresponding value of q is then the total amount of heat absorbed from the reservoir at temperature T_B.

It can be seen from the above descriptions that the form of the Q-curve gives a very good qualitative idea of the economies likely to be obtainable by the use of heat pumps. The total amount of heat to be supplied to the boiler of a conventional adiabatic column is given by $Q(T_B) - Q(T_F)$, and therefore corresponds to the vertical span of the Q-curve. On the other hand the energy (heat or work) required by the heat pumps is determined by the width of the valley about the feed point. Thus, if the Q-curve is more or less flat except for a deep, narrow valley at the feed point, the total heat required for adiabatic distillation is large, but the minimum energy requirement is small if heat pumps are correctly deployed. Such a system therefore offers considerable scope for energy economies. If the valley about the feed point is broad and flat, on the other hand, there is relatively little to be gained by heat pumping.

Quantitative Results for Particular Systems

In this section some numerical results will be quoted which have been obtained by applying the theory of the foregoing section to specific examples, and the effects of various departures from the ideal reversible system will be compared.

Two examples will be considered. The first is the methanol–water system whose Q-curve is given in Fig. 2, and the second a mixture of acetic acid and water, containing 28% by weight acid and at its boiling-point, which is separated at atmospheric pressure into a bottom product containing 99·0% acid and a top product containing 98·5% water. The Q-curve for the second example can be constructed from the enthalpy-composition chart and equilibrium data of Lemlich, Gottslich, and Hoke,[15] and in both examples energy consumptions were calculated by integration of equations (8) and (9) or equations (10) and (11) in the manner already described, making use of the Q-curves. Costs quoted are based on 1d/kWh for electric power, and 10 shil and £1 per ton, which are thought to span a reasonable range of costs for industrial heating steam. The results are given in Table I in which Q_{B0} is the boiler heat required for conventional adiabatic distillation at minimum reflux, Q_T is the total heat required from a source at temperature T_B with the optimum heat-pumping system using pumps which consume heat rather than work, W is the work consumption of the optimum heat-pumping system using heat pumps which consume electric power, C_1 is the energy cost for adiabatic distillation at minimum reflux, and C_2 is the cost of electric power in the optimum heat-pumping system. The two values given for C_1 refer to steam costs of 10 shil and £1 per ton respectively. The basis of the calculations is 100 lb of distillate in the methanol–water system and 100 lb of bottom product in the acetic acid–water system. The reduction in heat consumption or cost is very striking in both cases.

TABLE I.—*Energy Costs in Distillation*

	Methanol–water (Basis 100 lb distillate)	Acetic acid–water (Basis 100 lb bottom product)
Q_{B0} (lb cal)	45 730	452 000
Q_T (lb cal)	9375	62 807
W (lb cal)	850	2500
C_1 (pence)	4·57/9·14	45·2/90·4
C_2 (pence)	0·45	1·32

The results of Table I refer to idealised systems which represent the limits of sequences of real systems of increasing cost and complexity. If we disregard for the moment any mechanical inefficiencies of the heat pumps, there are three major reasons why practical systems will fall short of this performance. Firstly, heat must be transferred to and from the column across heat-transfer surfaces of finite area with which there will be associated finite temperature drops; secondly, the infinite set of heat pumps providing a continuous distribution of heat transfer along the column must be replaced by a finite number of pumps supplying and removing heat at a finite number of points, and thirdly, the column may contain only a finite number of plates of finite area while the ideal systems operate the column in a "pinched" condition at all points and therefore require an infinite number of theoretical plates. Since the pressure drop over a plate is zero in the ideal system, the plates are assumed to have zero hold up and infinite area.

It is interesting to see the relative effect on the cost of distillation of relaxing each of the first two idealisations.

To calculate the cost of power for a heat-pumping system which is ideal in every respect save the presence of a finite temperature difference ΔT associated with each heat-transfer surface, equations (8) and (9) are simply modified by replacing T by $T + \Delta T$ and T' by $T' - \Delta T'$, and integrated as before. For the methanol–water system, with $\Delta T = \Delta T' = 10°$ C, this increases the power cost from the value 0·45d/100 lb distillate, given in Table I, to 1·80d/100 lb distillate.

To investigate the effect of departure from the ideal heat distribution, we consider a very simple system comprising a single heat pump operating between a source in the upper part of the column and a sink in the lower part, and supplying enough heat to create a local pinch at the sink. The economics of such a system clearly depend on the position chosen for the sink, but by carrying out calculations with various sink temperatures it is possible to find the system with lowest total energy costs. If steam is charged at £1/ton and electrical energy at 1d/kWh, this is found to be a system delivering heat to a sink at 82° C and gives a total energy cost of 1·72d/100 lb distillate.

It is seen that even this gross simplification of the optimum heat distribution has a smaller economic effect than the presence of 10° C temperature drop across each of the heat-transfer surfaces, and it must be concluded that the reduction of these temperature drops is of prime importance. The simplest way of accomplishing this is to eliminate one of the two temperature drops completely by directly compressing vapour drawn from the top of the column and supplying heat by condensing it at a point or points in the lower part rather than by using an external heat-transfer medium separated from the contents of the column by heat-transfer surfaces at both source and sink. What can be achieved in principle with systems of this type will be discussed in the next section.

Optimum Direct Compression Systems

The systems described above are idealised systems and their performance is the limit of what can be achieved by realisable heat-pumping systems of increasing complexity. Although they are not themselves of any practical significance, their study is valuable in that it sets a limit to the attainable performance of real systems in the same way as the Carnot engine sets a limit to the attainable efficiency of real heat engines. We now turn our attention to systems which supply heat in the lower part of the column by compression of the overhead vapour of the column itself, and enquire, once again, what is the best performance attainable in principle by such a system.

To supply heat at any point in the lower part of the column the overhead vapour must be compressed to such a pressure that its saturation temperature exceeds the temperature within the column at the point in question. Heat can therefore be supplied at a number of points by compressing parts of the overhead vapour stream to appropriate pressures, then condensing the compressed vapour in thermal contact with the contents of the column. The condensate formed at higher pressures could in principle be let down through expansion engines to heaters at lower pressure thus recovering some mechanical work and giving up more heat to the column at each stage, but for reasons discussed earlier we exclude the possibility of expansion engines. We must then be content with flashing the condensate formed at higher pressures down through valves to heaters at lower pressures, thus recovering some extra heat from it but forgoing the benefit of recovering work at the same time. A system of this type with three heating stages is illustrated in Fig. 5 in which E_1, E_2, and E_3 are the three heat exchangers, C_1, C_2, and C_3 are the three compressors handling overhead vapour, and V_1, V_2, and V_3 are the three valves through which high-pressure condensate from one stage is flashed to the exchanger at the next highest pressure. The condensate from the final exchanger is flashed to the reflux drum, B, which provides the top product and the reflux in the usual way. Any vapour not used in the compression system may be condensed in the externally-cooled condenser, A, which also deals with vapour produced by flashing condensate through V_1. The balance of the column heat requirements not satisfied by the exchangers E_1, E_2, and E_3 must be supplied in an externally-heated reboiler, R.

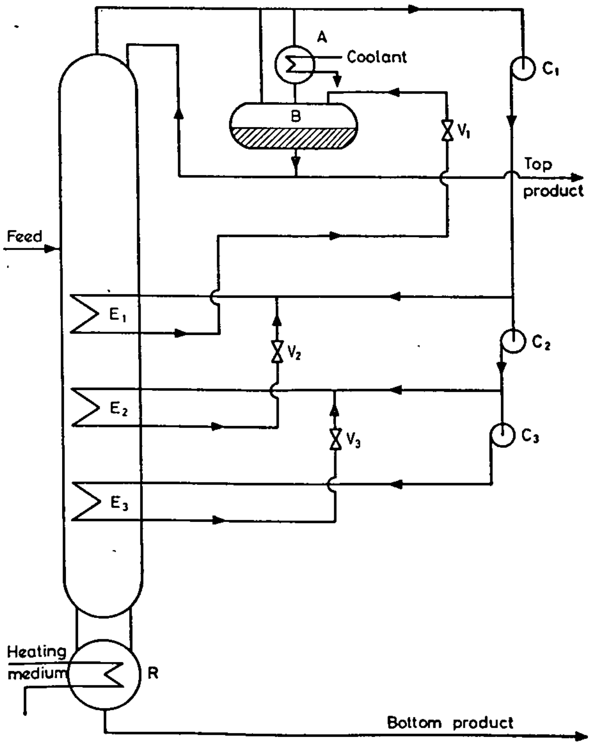


Fig. 5.—*Three-stage direct-compression system*

The work required to compress the overhead vapour is smallest when the vapour is condensed at as low a temperature as possible, so the aim should be to transfer heat from the condensing vapour to the column contents as far up the column as possible. The limit to which this distribution may be pushed is set by the Q-curve, and the minimum work of compression is expended when the heat supply to the lower section of the column from condensing overhead vapour is distributed as specified by the lower branch of the Q-curve. This, of course, entails the provision of an infinite number of plates and associated heat exchangers in the lower part of the column and makes use of an infinite sequence of infinitesimal compression stages as indicated schematically in Fig. 6. The system shown in Fig. 6 may be regarded as the limit of a sequence of practicable systems of the type shown in Fig. 5, with increasing complexity, and correspondingly its performance provides a limit which may be approached but not bettered by a practical system.

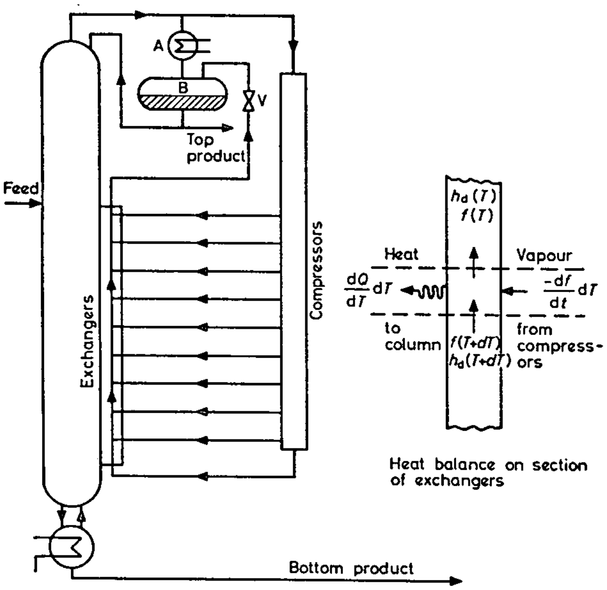


Fig. 6.—*The optimum direct-compression system*

In order to calculate the total work of compression in a system of this type it is necessary to know the distribution of flow of compressed overhead vapour as a function of the temperature at the point in the column to which heat is supplied. If f is the total flow of condensate down the temperature gradient in the heaters at a level in the column where the temperature is T, then the flow of compressed vapour condensed in the temperature interval T to $T + dT$ is $(-df/dT)dT$ and an enthalpy balance on an elementary section of the column and heaters in Fig. 6 between temperatures T and $(T + dT)$ gives the following differential equation for f:

$$\frac{d}{dT}[(H_d - h_d)f] = -\frac{dQ}{dT} + f\frac{dH_d}{dT} \quad . \qquad (12)$$

where H_d is the unit enthalpy of compressed (superheated) vapour which delivers its heat by condensing at temperature T, h_d is the unit enthalpy of the corresponding condensed liquid, and Q is the ordinate of the Q-curve at temperature T. The method of integrating equation (12) depends on whether or not the amount of overhead vapour available is sufficient, after compression, to supply all the heat requirements of the column without any additional external heat supply. If it is, equation (12) may be integrated upwards from the base of the column, starting from the initial condition $f = 0$ at $T = T_B$. If it is not, on the other hand, and the total flow of overhead vapour available for compression is $\bar{f}$, the integration must proceed downwards from the feed point, starting with the initial condition $f = \bar{f}$ at $T = T_F$. The value of $\bar{f}$ is easily seen to be given by:

$$\bar{f} = (R+1)D\left[\frac{H_{do} - h_{do}}{H_{do} - h_d(T_F)}\right] \quad . \quad . \quad (13)$$

where R is the reflux ratio at the top of the column, D is the flow of top product, H_{do} and h_{do} are the enthalpies of overhead vapour and its equilibrium liquid at the top of the column, and $h_d(T_F)$ is the enthalpy of liquid top product at the feed temperature T_F. The two cases are probably best distinguished in practice by starting the integration from $T = T_B$ and proceeding up the column. If f reaches the value $\bar{f}$ before T reaches the feed temperature T_F, this must be abandoned and the integration must be re-started from the feed. In either case the integration must be carried out numerically, using values of Q from the Q-curve and values of H_d and h_d from suitable thermodynamic approximations; for example:

$$H_d(T) = H_{do} + w(T) \quad . \quad . \quad . \quad (14)$$

$$h_d(T) = h_{do} + c(T - T_D) \quad . \quad . \quad (15)$$

where c is the specific heat of the liquid top product and $w(T)$ is the work required to compress unit quantity of the vapour from the top of the column to a pressure such that its saturation temperature is T. It is assumed that the top product is sufficiently pure to be treated as a single substance in these calculations.

Having obtained f by integration of equation (12), we obtain the total work of compression of overhead vapour from:

$$W = \int w \mathrm{d}f$$

$$= \int_{T_o}^{T_F} w(T)\frac{\mathrm{d}f}{\mathrm{d}T}\mathrm{d}T \quad . \quad . \quad . \quad (16)$$

where T_o is the highest temperature to which heat is supplied by compressed vapour. The value of T_o will be identical with that of T_B if sufficient vapour is available.

These calculations were carried through for the methanol–water separation already discussed. In this case there is sufficient vapour to supply all the heat requirements of the column when compressed, and the energy cost is simply the cost of electrical energy to drive the compressors. With a unit energy cost of 1d/kWh, this was found to be 0·46d/100 lb distillate. This should be compared with the value 0·45d given in Table I for the ideal reversible system, and it is seen that the irreversibilities associated with the present system have an almost negligible effect on the energy costs. The calculations were also repeated evaluating H_d and h_d in equation (12), not at temperature T, but at $T + \Delta T$, with $\Delta T = 10°$ C. Physically this corresponds to a 10° C temperature drop allowed across the heat-transfer surfaces. The energy cost is then found to be 1·23d/100 lb distillate, compared with a value 1·80d found above for the optimum system with an external heat-transfer medium when 10° C temperature drop is allowed at each heat-transfer surface. Thus the departure from ideal heat distribution in the system of Fig. 6 compared with the system of Fig. 4 is more than repaid by the elimination of one of the two heat-transfer surfaces. Although this conclusion has been reached for one particular system, there is good reason to suppose it is true in the majority of cases where heat pumping is likely to be an attractive proposition, for these are systems in which a substantial part of the energy needs to be pumped through only small temperature differences, and the heat-transfer temperature drop will then be correspondingly important. It is probably true, therefore, that a system employing direct compression of the overhead vapour is preferable to one using a separate heat-transfer medium, and should normally be used unless there are mechanical or chemical factors which make it impracticable or expensive to handle the overhead vapour in compressors.

Practical Direct Compression Systems

Practical approximations to the ideal direct compression system of Fig. 6 are based on a system with a finite number of compression stages like the one shown in Fig. 5. In practice, the reflux ratio at the top of the column must be sufficiently high to avoid a pinch above the feed and give a reasonable number of plates in the upper part of the column. The quantities of heat supplied in the exchangers E_1, E_2, and E_3 must also be less than the limiting quantities deducible from the Q-curve so that pinches are not generated at these positions. The temperature at any point in the column depends on pressure as well as composition, so in order to keep the pressure of the condensing compressed vapour as low as possible, and hence to reduce the work of compression as far as possible, it is important to choose low pressure drop plates or packing and not to load the column too heavily.

To keep the temperature drop across the heat-transfer surfaces in exchangers E_1, E_2, and E_3 as small as possible a large heat-transfer-surface area must be provided, and this would be difficult to accommodate within the column as indicated in Fig. 5, since the tubes would need to be submerged in the liquid hold-up on a plate. It is possible to arrange internal heat exchangers of a completely different design from an ordinary distillation plate, but probably the most practical method of dealing with this problem is to draw off liquid at the point to be heated and carry out the heating in an external tube and shell exchanger, thus obtaining a system of the type illustrated in Fig. 7. Strictly, the combination of a plate within the column and a separate vaporiser, as shown in Fig. 7, is not equivalent to a perfectly-mixed plate with submerged heating tubes on which the theoretical calculations have been based. It is possible, though rather complicated, to modify the thermodynamic calculations to apply rigorously to a system of the type shown in Fig. 7, but in practice this is hardly justifiable. The difference between the two calculations is not very large, and actually the arrangement of an ideal plate and separate external vaporiser gives slightly greater enrichment than one theoretical plate with built-in heating. Thus calculations based on the system of Fig. 5 will, in principle, give slightly pessimistic results when applied to the system of Fig. 7.

The simplest practical compression system is one in which part of the overhead vapour is compressed and condensed in a single heat exchanger, supplying heat to just one point in the column between the feed and the reboiler. The balance of the heat requirements of the column are then supplied by an external heating medium in a reboiler at the base of the column. In such a system the position of the intermediate exchanger has an important effect on the overall economics of the process. In conventional vapour recompression schemes, the reboiler itself is the exchanger in which compressed overhead vapour is condensed, but there will be few cases in which this is the best arrangement. As the position of the intermediate exchanger moves up the column from the reboiler, the amount of expensive external heat required at the boiler increases, but the amount of compressed vapour to be condensed in the intermediate exchanger decreases, and so does the work of compression per unit of this vapour. When the exchanger is near the bottom of the column the decrease in work of compression on displacing it upward usually more than pays for the extra heat required at the boiler: so the total energy cost decreases. Near the feed, on the other hand, the total energy cost usually increases when the intermediate exchanger is displaced upward and it follows that there is frequently some position of the intermediate exchanger, between the feed point and the base of the column, which gives minimum total energy cost.

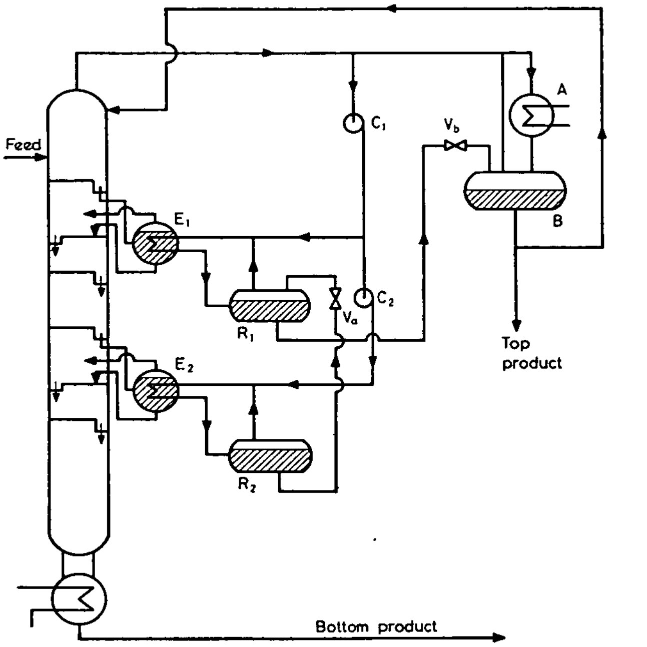


Fig. 7.—*Practical three-stage direct-compression system with external heat exchangers*

When considering a compression system with any number of stages it is clearly important to find the best positions in the column for the intermediate heat exchangers in which compressed vapour is to be condensed, and in the Appendix an approximate method of locating these positions is described.

Of course, when assessing the relative economics of different arrangements, one must take the capital cost of the system into account as well as the energy costs. The capital cost of the compressors and exchangers decreases as the positions of the exchangers move up the column towards the feed, with the result that the most economic system, taking account of both energy and capital costs, has its exchangers rather nearer to the feed than the positions which minimise the energy costs alone. This will be illustrated by the figures given below for specific systems.

Calculations to determine the optimum single-stage compression systems have been carried out for several separations[16] and we shall quote detailed results for the methanol–water and acetic acid–water separations described earlier in the paper. In the methanol–water example the reflux ratio is taken to be 1 : 1, the isentropic compressor efficiency is assumed to be 70% and a temperature drop of 10° C is allowed across the heat-transfer surface. The total energy costs were calculated for conventional adiabatic distillation, for a conventional vapour-recompression system with condensation of the compressed vapour in the reboiler, and for single-stage compression systems of the type described here with heat supply to liquids of various compositions by condensation of compressed vapour.

The number of theoretical plates varied only between 17 and 20 for all the systems considered, so it was not worth while adjusting the reflux ratio to maintain the number of plates strictly constant. The results are given in Table IIA, in which x is the weight fraction of methanol in the liquid on the heated plate, and the two cost figures quoted correspond to unit costs of 10 shil/ton and £1/ton for heating steam. In both cases it is seen that the total energy cost passes through a minimum which is substantially smaller than the costs for either adiabatic distillation or conventional vapour recompression. The optimum liquid composition x depends, of course, on the unit cost of the heating steam.

TABLE IIA.—*Costs of Separation: Methanol–Water System*

System	Total energy cost per 100 lb distillate (pence)	
Vapour recompression to reboiler	5·25	5·25
Compression to heat liquid with $x = 0·2$	3·8	4·05
$x = 0·4$	3·1	3·5
$x = 0·5$	2·9	3·5
$x = 0·6$	2·75	3·6
$x = 0·7$	2·8	4·1
$x = 0·8$	3·1	5·2
Conventional adiabatic column	5·15	10·3

In the separation of acetic acid and water the reflux ratio is taken to be 5 : 1, the compressor efficiency is assumed to be 70% and a temperature drop of 10° C is allowed across the heat-transfer surface, as before. A set of calculations corresponding to those for the methanol–water system were carried out and it was found that the number of theoretical plates in the column varied between 23 and 30 in different cases. The calculations were based on a McCabe–Thiele diagram with a fictitious molar weight of 100 for acetic acid to give a constant pseudo-molar latent heat of vaporisation.

TABLE IIB.—*Costs of Separation: Acetic Acid–Water System*

System	Total energy cost per 100 lb bottom product (pence)	
Vapour recompression to boiler	47·5	47·5
Compression to heat liquid with $x = 0·2$	39·0	42·0
$x = 0·4$	31·5	34·0
$x = 0·5$	29·5	32·5
$x = 0·6$	27·7	32·0
$x = 0·7$	27·5	35·0
$x = 0·8$	32·0	48·0
Conventional adiabatic column	82·0	163·0

The results are presented in Table IIB, where x denotes the pseudo-mole fraction of water in the liquid on the heated plate and, as before, the cost figures quoted refer to unit costs of 10 shil/ton and £1/ton for heating steam. Once again there is a value of x for which the total energy cost is a minimum.

The costs in Tables IIA and IIB should be compared with those given in Table I for ideal reversible heat-pumping systems. Although the single-stage systems of Tables II come nowhere near the minimum energy cost attainable in principle, they still show a very substantial saving over conventional distillation.

The above results give no indication of the effect of capital costs on the economics of the process, so further calculations were carried out to give estimates of capital costs and equipment specifications for plants of specified annual output. In the methanol–water example the basis was taken to be a plant producing 22 000 tons per annum of the 99·95% by weight methanol top product, and in the acetic acid–water example a plant producing 5000 tons per annum of the 90% by weight concentrated acid. The installed capital costs of compressors were estimated from a chart given by Vilbrandt and Dryden[17] (converting costs to sterling at the rate £1 ≡ $3). The capital cost of extra heat-transfer equipment (over and above that required for conventional adiabatic distillation), extra plates, valves, and piping was estimated using approximate methods such as those described by Sawistowski and Smith.[18]

The results are set out in Tables IIIA for the methanol–water system and IIIB for the acetic acid–water system, and show the power of the motor required to drive the compressors, the intake volume of the compressor and the pressure rise, in addition to capital and energy costs.

Table IIIA.—*Plant to Produce* 22 000 *ton/a of Methanol*

System	Energy cost (steam at 10 shil/ton)	Energy cost (steam at £1/ton)	Motor power	Intake volume of compressor	Pressure rise	Installed cost of compressor and motor	Total installed cost of heat-pumping equipment
	(£/a)	(£/a)	(hp)	(ft^3/min)	(lb/in^2)	(£)	(£)
Vapour recompression to boiler	10 770	10 770	425	2567	56·1	18 000	35 400
Compression to heat liquid containing 50 % wt methanol	5960	7210	186	2386	17·7	7000	22 200
Compression to heat liquid containing 60% wt methanol	5690	7430	156	2269	14·9	6000	20 400
Compression to heat liquid containing 70% wt methanol	5750	8420	122	2042	12·5	4500	17 700
Conventional adiabatic column	10 500	21 000	—	—	—	—	—

Table IIIB.—*Plant to Produce* 5000 *ton/a Acetic Acid*

System	Energy cost (steam at 10 shil/ton)	Energy cost (steam at £1/ton)	Motor power	Intake volume of compressor	Pressure rise	Installed cost of compressor and motor	Total installed cost of heat-pumping equipment
	(£/a)	(£/a)	(hp)	(ft^3/min)	(lb/in^2)	(£)	(£)
Vapour recompression to boiler	22 300	22 300	835	8900	23	36 000	83 000
Compression to heat liquid containing 21% wt water	13 500	15 300	430	8400	10	17 000	67 500
Compression to heat liquid containing 30% wt water	12 800	16 300	352	8200	8·8	15 000	64 000
Compression to heat liquid containing 42% wt water	15 000	22 400	287	7300	7·2	12 000	54 300
Conventional adiabatic column	38 200	76 400	—	—	—	—	—

For acetic acid, due allowance is made for construction in stainless steel because of the corrosive nature of the substance handled.

When heating steam costs £1/ton it is seen that the reduction in energy costs, compared with those of a conventional adiabatic column, pays off the extra capital investment in between one and two years. This is a very satisfactory return on capital, and even with the lower steam cost the return on capital is still quite good.

Clearly in more detailed design studies it would be necessary to consider various reflux ratios and the effect of changing the temperature drop allowed across the heat-transfer surface. It would also be important to consider the possible advantages of using two or more stages of compression and condensation, as indicated in Fig. 7. In the acetic acid–water separation, a two-stage system can be obtained by compressing overhead vapour to 10 lb/in^2 gauge and condensing most of it at this pressure. The remainder, quite small in quantity, can then be further compressed in a second compressor to 23 lb/in^2 gauge, after which it can be condensed in the reboiler to supply the remainder of the column heat load and eliminate the need for heating steam. The pressure ratio of the second compressor remains reasonably small and the intake volume is only 1680 ft^3/100 lb acid product. The total energy costs of the separation are then reduced to 27d/100 lb acid, independent of the cost of heating steam, and this should be compared with the best costs of 27·5d/100 lb acid and 32·0d/100 lb acid given in Table IIB for a single-stage system, with steam costs of 10 shil and £1/ton respectively. Clearly there is no justification for the second stage when the steam cost is 10 shil/ton, but the extra capital cost may be worth while when the steam cost is £1/ton.

During start-up, the reboiler would be heated by an independent heat supply until a sufficient overhead vapour rate had been established to enable the compressors to operate satisfactorily. Since the heat load at the reboiler and condenser will frequently be determined by these start-up conditions rather than by the steady-state conditions, the savings in energy costs obtained by using compressed overhead vapour in the reboiler must be balanced against the small saving in capital cost of the reboiler and condenser surface and a possible loss in flexibility of the column.

Finally, it must not be inferred from the examples worked here that the application of the direct compression system is confined to binary mixtures of only moderate non-ideality. The methanol–water and acetic acid–water systems were chosen only because of the availability of accurate thermodynamic data. Indeed it is doubtful whether distillation is the best method of effecting the latter separation.

Discussion

In a McCabe–Thiele diagram, the heat to be supplied to the reboiler of a distillation column is determined by the slope of the operating line at the base of the column, decreasing as this slope increases. The principle of the present method is to increase this slope by condensing compressed overhead vapour and hence supplying heat to the column at some point or points between the feed plate and the reboiler.

Freshwater[19] has proposed an alternative method of increasing the slope of the lower operating line which is applicable when the form of the equilibrium relations is such that there is a pinch above the feed plate. In an adiabatic column, the reflux ratio must be sufficiently large to avoid this pinch, and this has the effect of unduly reducing the slope of the lower operating line. It is therefore suggested that high reflux should be generated locally in the region of the pinch by using a heat pump over quite a small temperature interval. The reflux ratio at the top of the column can then be reduced without meeting difficulty at the pinch, and correspondingly the slope of the lower operating line is increased and the heat consumption at the boiler reduced. However, the economy in heat consumption obtainable in this type of system is still limited by the creation of a pinch at the feed point, whereas the systems described here permit considerably greater economies. This is borne out by the relatively modest reduction in energy consumption (about 20%) quoted by Freshwater for a worked example.

Actually the effect on the operating lines of heat transfer to and from the column is erroneously represented in the diagrams of Freshwater's paper, but it is not known whether the numerical calculations quoted are subject to a corresponding error.

It is hoped that this paper has shown that vapour recompression, often regarded as impracticable in most distillation columns because of the large temperature difference between condenser and reboiler, may be an attractive proposition if the compressed vapour is condensed at some point, or points, in the column more appropriate than the reboiler.

Acknowledgments

The authors wish to acknowledge the help they received from discussions with Dr. N. Macleod and with Mr. A. J. Hunter, who first roused their interest in this subject. They were also greatly assisted by discussions with Mr. J. J. Franses, Mr. S. Pridham, and Mr. D. M. Goldin.

APPENDIX

Approximate Method of Locating Best Positions for Intermediate Heat Exchangers

Consider the system of Fig. 5 with a reboiler, R, and N exchangers, $E_1, \ldots, E_n, \ldots, E_N$, at positions where the temperatures of the contents of the column are $T_1, \ldots, T_n, \ldots, T_N$, numbered in order of increasing temperature. If the heat liberated by the flashing condensate entering exchanger E_n is small compared with the total heat supply, $q^*(T_n)$, at exchanger E_n, the associated work of compression, $W^*(T_n)$, is:

$$W^*(T_n) = q^*(T_n) \cdot \frac{w(T_n)}{[H_d(T_n) - h_d(T_n)]} \quad . \quad (17)$$

The total cost of supplying the energy requirements of the column is:

$$\text{T.E.C.} = C_3\left\{Q_B + \sum_1^N q^*(T_n) \cdot \frac{w(T_n).C}{[H_d(T_n) - h_d(T_n)]}\right\} \quad (18)$$

where C_3 is the unit cost of heat supplied to the reboiler and C is the ratio of the unit cost of power supplied to the compressors to that of heat supplied to the reboiler. If the reboiler is heated by condensing compressed overhead vapour:

$$C = \frac{H_d(T_B) - h_d(T_B)}{w(T_B)} \quad . \quad . \quad (19)$$

From the point of view of energy costs, the optimum system is that set of values of T_n which minimises equation (18) subject to:

$$\sum_1^n q^*(T_i) \leqslant Q(T_n) - Q(T_F) \quad . \quad . \quad . \quad (20)$$

and:
$$f(T_1) \leqslant (1 + R)D\left[\frac{H_{do} - h_{do}}{H_{do} - h_d(T_1)}\right] \quad . \quad (21)$$

These relations ensure that the column is free from pinched conditions and that sufficient overhead vapour is available for compression.

From equation (20):

$$Q_B = Q(T_B) - Q(T_F) - \sum_1^N q^*(T_n),$$

so:
$$\text{T.E.C.} = C_3\left\{Q(T_B) - Q(T_F\}\right.$$
$$\left. - \sum_1^N q^*(T_n)\left[1 - \frac{C.w.(T_n)}{H_d(T_n) - h_d(T_n)}\right]\right\} \quad (22)$$

In a given problem, a diagram such as that shown in Fig. 8 can be constructed. The energy cost is proportional to the area of the hatched region. The area of the nest of rectangles shown by broken lines represents the savings in energy costs achieved when the system of Fig. 5 is used instead of the conventional systems where all the heat is supplied to the reboiler. The inequality (20) is represented by the vertical distance between the value of $Q(T)$ and the upper boundary of the nest of rectangles. Therefore the height of the steps can be chosen to provide a suitable compromise between energy savings and an excessively large number of plates. Negative values of the abscissa clearly represent the range of values of T_n where heat supply by condensing compressed overhead vapour is uneconomic. The diagram also shows the advantages obtained when the Q-curve has a shape similar to that shown in Fig. 2.

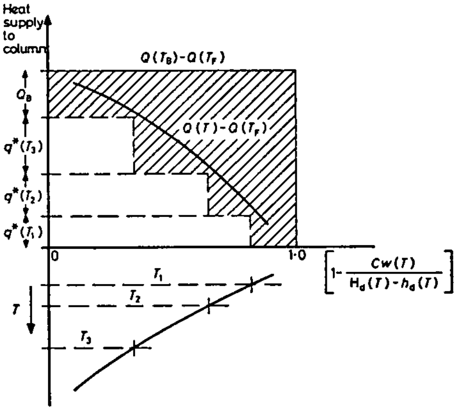


Fig. 8.—*Approximate determination of the optimum positions for the heat exchangers in the system of Fig. 5*

If N is small, as seems likely in most practical cases, estimates of the optimum values of T_n, corresponding to the maximum area of the nest of rectangles, can be found very easily by visual inspection of the diagram shown in Fig. 8. The condition imposed by inequality (21) can then be checked. The further saving obtained by using $N + 1$ exchangers can also be estimated very quickly. These estimates of the optimum values of T_n would then form the starting point of a further set of calculations which include the effect of the flashing condensate on the heat supply at each exchanger and introduce assigned capital costs for the extra equipment.

Symbols Used

B = flow of bottom product.
C = ratio of unit costs of energy as power and heat to reboiler.
c = specific heat of condensed overhead vapour.
C_1 = energy cost for conventional distillation with adiabatic column.
C_2 = cost of electrical energy in heat-pumping system.

C_3 = unit cost of heat supplied to the reboiler of the system in Fig. 5.
D = flow of distillate.
f = flow of condensate in optimum direct-compression system.
$\bar{f}$ = total flow of overhead vapour available for compression.
F = flow of feed.
G_p = flow of pth product stream.
$h_d(T)$ = unit enthalpy of compressed vapour after condensation.
h_{do} = unit enthalpy of condensed overhead vapour.
$H_d(T)$ = unit enthalpy of compressed vapour which delivers heat by condensation at temperature T.
H_{do} = unit enthalpy of overhead vapour.
H_F = unit enthalpy of feed stream.
H_p = unit enthalpy of pth product stream.
ΔH = enthalpy change accompanying the separation.
L = latent heat of vaporisation.
N = total number of exchangers in the system of Fig. 5.
q = heat required to operate heat pumps.
$Q(T)$ = ordinate of Q-curve at temperature T.
$Q^*(T_n)$ = heat supply to the exchanger where the temperature of the column is T_n.
Q_i = heat absorbed from reservoir at temperature T_i.
Q_B = heat supplied to boiler.
Q_D = heat removed at condenser.
Q_T = total heat taken from source at temperature T_B.
ΔQ = total amount of heat to be supplied below specified point in column if system is to be reversible.
$\Delta Q'$ = heat to be removed between specified point above feed and top of column if system is to be reversible.
R = reflux ratio at top of column.
S_F = unit entropy of feed.
S_p = unit entropy of pth product stream.
s_1 = slope of upper operating line.
s_2 = slope of lower operating line.
ΔS = entropy change accompanying the separation.
T = Absolute temperature.
T_B = Temperature in reboiler.
T_D = Temperature in condenser.
T_F = Temperature at feed plate.
T_i = temperature of ith heat reservoir.
T_n = temperature of contents of column at exchanger E_n.
T_o = highest temperature in column to which heat is supplied by compressed vapour.
T.E.C. = total cost of energy.
ΔT = Temperature drop across heat-transfer surface.
w = incremental work in direct compression systems.
W = total mechanical work performed on system.
$W^*(T_n)$ = work of compression associated with the heat supply at exchanger E_n.

The above quantities may be expressed in any set of consistent units in which force and mass are not defined independently.

References

[1] van Nuys, C. C. *Chem. Metall. Engng*, 1923, **28**, 207.
[2] van Nuys, C. C. *Chem. Metall. Engng*, 1923, **28**, 255.
[3] van Nuys, C. C. *Chem. Metall. Engng*, 1923, **28**, 311.
[4] van Nuys, C. C. *Chem. Metall. Engng*, 1923, **28**, 359.
[5] van Nuys, C. C. *Chem. Metall. Engng*, 1923, **28**, 408.
[6] Dodge, B. F., and Housum, C. *Trans. Amer. Inst. chem. Engrs*, 1927, **19**, 117.
[7] Hausen, H. *Z. techn. Phys.*, 1932, **13**, 271.
[8] Haselden, G. G. *Trans. Instn chem. Engrs*, 1958, **36**, 123.
[9] Haselden, G. G. *de Ingenieur*, 1962, **74**, "Chem. Tech.", **2**, 9.
[10] Ruhemann, M. "*The Separation of Gases*", 1940 (Oxford: Oxford University Press).
[11] Freshwater, D. C. *Trans. Instn chem. Engrs*, 1951, **29**, 149.
[12] Timmers, A. C. Contribution to discussion of reference 9.
[13] Beek, J. Contribution to discussion of reference 9.
[14] Plewes, A. C., Jardine, D. A., and Butler, R. M. *Canadian Journal of Technology*, 1954, **32**, 133.
[15] Lemlich, R., Gottslich, C., and Hoke, R. *Industrial and Engineering Chemistry* (*Data Series*), 1957, **2**, 32.
[16] *British Patent application No.* 7355/63.
[17] Vilbrandt, F. C., and Dryden, C. E. "*Chemical Engineering Plant Design*", 1959 (London: McGraw-Hill Book Co.).
[18] Sawistowski, H., and Smith, W. "*Mass Transfer Process Calculations*", 1963 (New York: Interscience Publishers).
[19] Freshwater, D. C. *British Chemical Engineering*, 1961, **6**, 388.

The manuscript of this paper was received on 6 *May*, 1964.