AN EFFECTIVE COMPUTATIONAL METHOD OF THE OPTIMAL DISCRETE CONTROL ACTION OF CHEMICAL PROCESS BASED ON SENSITIVITY ANALYSIS

TAKEICHIRO TAKAMATSU AND YOSHIAKI SHIMIZU
*Department of Chemical Engineering, Kyoto University,
Kyoto 606*

Piece-wise constant control policy is studied based on sensitivity analysis with a discontinuous function and an algorithm for solving optimal control problems in a practical manner is developed.

Compared with other numerical methods (simplex, Rosenbrock's and the conjugate gradient method), both effectiveness of the algorithm and technical feasibility of the decision obtained are examined by two examples of chemical reaction processes.

Introduction

In many chemical processes, some controlling actions are operated discontinuously along time or space co-ordinate from practical feasibility of the discontinuous policies such as relay or bang-bang control. A mathematical model for such cases is described by a set of differential equations with discontinuous arguments of independent or control variables. It is of practical significance to study the discontinuous system.

Though the concept of sensitivity has been effectively used in continuous systems, not many studies have been made of discontinuous ones. P. R. Belanger¹⁾ has studied sensitivity analysis with control function and has discussed the conditions under which state can reach the assigned target set even if both continuous and discontinuous deviations of control function may occur. E. N. Rozenvasser¹¹⁾ has proposed a general sensitivity equation for the system parameters, including the case when the system model may change discontinuously under a certain condition. Recently, noticing that bang-bang control frequently becomes the optimal control when control variables appear linearly in the system model, J. Rootenberg and P. Courtin⁹⁾ have worked on sensitivity analysis for the system parameters.

The first part of this paper describes sensitivity analysis along with the previous studies. Then the

foregoing consideration is applied to develop an effective method for solving optimal control problems in a practical manner. For this purpose the piece-wise constant control set was chosen as the admissible set. This restriction on control set did not result in critical deactivation of system performance since system performance is not so sensitive, in many cases, with control near optimal¹³⁾. This simple form of control action not only offered a higher technical feasibility but also made it easier to develop a new optimization technique.

1. Sensitivity Analysis with Discontinuous Function

Consider a system whose dynamics are described by

$$\dot{x} = f(x(t), w(t), t), \quad x(t_0) = x_0 \quad (1)$$

where $x(t)$ denotes an n -dimensional state vector which is continuous on $t \in [t_0, t_f]$, $w(t)$ is a scalar first-kind discontinuous function defined on $t \in [t_0, t_f]$,* $f(\cdot)$ is an n -dimensional vector function of C^1 class with its arguments and t is an independent variable.

Suppose that $w(t)$ is discontinuous at time $t_i \in (t_0, t_f)$ and has $N-1$ discontinuous points totally in $[t_0, t_f]$, and that at any instant between the sequential discontinuous points, $t \in [t_{i-1}, t_i)$ has a constant value, w_i .

$$w(t) = w_i = \text{constant} \quad t \in [t_{i-1}, t_i) \quad (i = 1, \dots, N) \quad (2)$$

In the following, this kind of function is called piece-wise constant and its small deviation, $\delta w(t)$ is defined by Eq. (3).

$$\delta w(t) = \tilde{w}(t) - w(t) \quad (3)$$

where $\tilde{w}(t)$ is

Received August 22, 1977. Correspondence concerning this article should be addressed to Y. Shimizu, Research Reactor Institute, Kyoto Univ., Kumatori-cho, Sennan-gun, Osaka 590-04.

* For simplicity the scalar case will be considered here, but extension to the vector case is straightforward.

Fig. 1 Small variation of piece-wise constant function

$$\tilde{w}(t) = \begin{cases} w_1 + \epsilon_1 \delta w_1 & t \in [t_0, t_1 + \epsilon_2 \delta t_1] \\ w_2 + \epsilon_1 \delta w_2 & t \in [t_1 + \epsilon_2 \delta t_1, t_2 + \epsilon_2 \delta t_2] \\ \vdots & \vdots \\ w_N + \epsilon_1 \delta w_N & t \in [t_{N-1} + \epsilon_2 \delta t_{N-1}, t_N] \end{cases} \quad (4)$$

where ϵ_1 and ϵ_2 are small positive real numbers.

Assuming that Eq. (1) has a unique solution, then the deviation of $x(t)$, $\delta x(t)$ due to $\delta w(t)$ can be derived from Eq. (5).

$$\delta \dot{x}(t) = (\partial f / \partial x) \delta x(t) + (\partial f / \partial w) \delta w(t) + \Delta f(t) \quad (5)$$

$$\delta x(t_0) = 0$$

where $\Delta f(t)$ is

$$\Delta f(t) = \begin{cases} f(x, w_i, t) - f(x, w_{i+1}, t), & t \in [t_i, t_i + \epsilon_2 \delta t_i], \text{ when } \delta t_i \geq 0 \\ f(x, w_i, t) - f(x, w_{i+1}, t), & t \in [t_i + \epsilon_2 \delta t_i, t_i], \text{ when } \delta t_i < 0 \\ 0 & \text{(otherwise)} \end{cases} \quad (6)$$

$(i=1, \dots, N-1)$

Though Eq. (5) is no more than the special case of the general sensitivity equation of Rozenvasser, its solution can be explicitly expressed as follows.

$$\delta x(t_f) = \epsilon_1 \sum_{i=1}^N \delta w_i \int_{t_{i-1}}^{t_i} \Phi(t_f, \tau) \left(\frac{\partial f}{\partial w} \right) d\tau + \epsilon_2 \sum_{i=1}^{N-1} \delta t_i \Phi(t_f, t_i + \theta_i^*) \{ f(x, w_i, t_i + \theta_i) - f(x, w_{i+1}, t_i + \theta_i) \} \quad (7)$$

Here $\Phi(t, \tau)$, called state transition matrix, is obtained from the following differential equation.

$$\dot{\Phi}(t, t_0) = \left(\frac{\partial f}{\partial x} \right) \Phi(t, t_0), \quad \Phi(t_0, t_0) = I \quad (8)$$

where I is an $(n \times n)$ -dimensional unit matrix. In these equations the terms $\partial f / \partial x$ and $\partial f / \partial w$ are calculated along the nominal value of $w(t)$.

2. Statement of the Problem and Fundamental Results

Based on the foregoing results, the following optimal

* It is assumed there exists such θ_i , $0 \leq |\theta_i| \leq \epsilon_2 \cdot |\delta t_i|$ that $|\theta_i / \epsilon_2| \rightarrow 0$ as $\epsilon_2 \rightarrow 0$.

** Extension to the vector case is also straightforward.

control problem will be considered under the restriction imposed on the control set from technical feasibility.

$$\begin{aligned} \text{Max}_{u \in U} J &= K(x(t_f), t_f) \\ \text{subject to } \begin{cases} \dot{x} = f(x(t), u(t)), & x(t_0) = x_0 \\ u \leq u(t) \leq \bar{u} \end{cases} \end{aligned}$$

where x is an n -dimensional state vector, u is a scalar** control function, f is an n -dimensional vector-valued function of C^1 class and a final time t_f is assumed to be fixed.

The admissible control set U is specified to the control set which is made of piece-wise constant functions with $N-1$ discontinuous points over $[t_0, t_f]$.

Now the control function $u(t)$ is uniquely prescribed by a pair of (v_i, τ_i) where v_i is the value of $u(t)$ at i -th interval and τ_i is the i -th discontinuous point. Then the above problem can be transformed into the following $2N-1$ dimensional optimization problem with v and τ , where v and τ are respectively the vector forms of the sequences of v_i and τ_i .

$$\begin{aligned} \text{Max}_{v, \tau} J &= K(x(t_f), \tau_f) \\ \text{subject to } \begin{cases} \dot{x} = f(x(t), v_i), & x(\tau_{i-1}^+) = x(\tau_{i-1}^-) \\ (\tau_{i-1} \leq t < \tau_i) \\ u \leq v_i \leq \bar{u} \quad (i=1, \dots, N) \end{cases} \end{aligned}$$

Here it should be noticed that $\tau_0 = t_0$ and $\tau_N = t_f$ are implied.

Following the usual approach, the inequality constraints on each v_i will be augmented to the original performance index by the Lagrangian multipliers, $\bar{\lambda}_i$ and $\underline{\lambda}_i$.

$$\tilde{J} = K(x(t_f), t_f) + \sum_{i=1}^N \bar{\lambda}_i (\bar{u} - v_i) + \sum_{i=1}^N \underline{\lambda}_i (v_i - \underline{u}) \quad (9)$$

Now the variation of $\tilde{J}$ due to the small deviation of $u(t)$, a performance index sensitivity, will be described by referring to Eq. (7) as follows.

$$\begin{aligned} \delta \tilde{J} &= \left[\frac{\partial K}{\partial x} \right]_{t_f} \delta x(t_f) + \epsilon_1 \sum_{i=1}^N (\bar{\lambda}_i - \underline{\lambda}_i) \delta v_i \\ &= \epsilon_1 \sum_{i=1}^N \delta v_i \left\{ \left[\frac{\partial K}{\partial x} \right]_{t_f} \int_{\tau_{i-1}}^{\tau_i} \Phi(t_f, \tau) \left(\frac{\partial f}{\partial u} \right) d\tau + \bar{\lambda}_i - \underline{\lambda}_i \right\} \\ &\quad + \epsilon_2 \sum_{i=1}^{N-1} \delta \tau_i \left[\frac{\partial K}{\partial x} \right]_{t_f} \Phi(t_f, \tau_i + \theta_i) \{ f(x, v_i, \tau_i + \theta_i) - f(x, v_{i+1}, \tau_i + \theta_i) \} \end{aligned} \quad (10)$$

Then the necessary conditions for optimality are given as follows.

$$\left[\frac{\partial K}{\partial x} \right]_{t_f} \int_{\tau_{i-1}}^{\tau_i} \Phi(t_f, \tau) \left(\frac{\partial f}{\partial u} \right) d\tau + \bar{\lambda}_i - \underline{\lambda}_i = 0 \quad (11a)$$

$$\bar{\lambda}_i (\bar{u} - v_i) = 0, \quad \bar{\lambda}_i \geq 0 \quad (i=1, \dots, N) \quad (11b)$$

$$\underline{\lambda}_i (v_i - \underline{u}) = 0, \quad \underline{\lambda}_i \leq 0 \quad (i=1, \dots, N) \quad (11c)$$

$$\left[\frac{\partial K}{\partial x} \right]_{t_f} \Phi(t_f, \tau_i) \{ f(x, v_i, \tau_i) - f(x, v_{i+1}, \tau_i) \} = 0 \quad (i=1, \dots, N-1) \quad (11d)$$

By using the well-known relation between the matrix $\Phi(t, t_0)$ and its adjoint $\Psi(t, t_0)$, Eq. (12), the above conditions will be rewritten as Eqs. (13) for convenience.

$$\Phi(t_1, t_2) = \Phi^{-1}(t_2, t_1) = \Psi^T(t_2, t_1) \quad (12)$$

$$\begin{cases} \left[\frac{\partial K}{\partial \mathbf{x}} \right]_{t_f} [\Psi^{-1}(t_f, t_0)]^T \int_{\tau_{i-1}}^{\tau_i} \Psi^T(\tau, t_0) \left(\frac{\partial \mathbf{f}}{\partial u} \right) d\tau \\ + \lambda_i - \bar{\lambda}_i = 0 \end{cases} \quad (13a)$$

$$\bar{\lambda}_i(\bar{u} - v_i) = 0, \quad \bar{\lambda}_i \geq 0 \quad (i=1, \dots, N) \quad (13b)$$

$$\lambda_i(v_i - \bar{u}) = 0, \quad \lambda_i \leq 0 \quad (13c)$$

$$\begin{cases} \left[\frac{\partial K}{\partial \mathbf{x}} \right]_{t_f} [\Psi^{-1}(t_f, t_0)]^T \Psi^T(\tau_i, t_0) \{ \mathbf{f}(\mathbf{x}, v_i, \tau_i) \\ - \mathbf{f}(\mathbf{x}, v_{i+1}, \tau_i) \} = 0 \end{cases} \quad (i=1, \dots, N-1) \quad (13d)$$

where $\Psi(t, t_0)$ is the solution of the following adjoint system.

$$\dot{\Psi}(t, t_0) = - \left(\frac{\partial \mathbf{f}}{\partial \mathbf{x}} \right)^T \Psi(t, t_0), \quad \Psi(t_0, t_0) = \mathbf{I} \quad (14)$$

3. Computational Aspects

The necessary conditions given by Eqs. (13) do not require solving the two-point boundary value problem like the modified discrete maximum principle²¹. This feature provides a striking computational advantage.

Although Eqs. (13) may be solved directly by numerical methods such as the Newton-Raphson or certain direct search methods, a certain kind of gradient method is available by regarding the left-hand side of Eqs. (13a) and (13d) as the variation rate of $\bar{J}$ with v_i and τ_i respectively. In the calculation of this gradient, the system equation and the adjoint equation, Eq. (14), can be simultaneously solved in forward time and such a troublesome procedure as the exchange of integration from forward to backward does not occur as it does in the usual gradient method.

Among the several sorts an algorithm has been developed by the steepest-descent method with quadratic interpolation to estimate the minimum point at each iteration. In this algorithm, the next searching point is determined by Eq. (15) or Eq. (16) at k -th step.

For v -mode:

$$\begin{cases} \bar{u} & (\text{if } {}^{(k)}v_i + {}^{(k)}\nabla_{v_i} J > \bar{u}) \\ {}^{(k+1)}v_i = {}^{(k)}v_i + {}^{(k)}\nabla_{v_i} J & (i=1, \dots, N) \\ \bar{u} & (\text{if } {}^{(k)}v_i + {}^{(k)}\nabla_{v_i} J < \bar{u}) \end{cases} \quad (15)$$

For τ -mode:

$$\begin{cases} {}^{(k+1)}\tau_i = {}^{(k)}\tau_i + {}^{(k)}\nabla_{\tau_i} J & (i=1, \dots, N-1) \\ {}^{(k+1)}v_i = {}^{(k)}v_i & (i=1, \dots, N) \end{cases} \quad (16)$$

where ${}^{(k)}h^*$, which gives the minimum value of $\bar{J}$ in k -th searching direction, is determined from quadratic interpolation and $\nabla_{v_i} J$ and $\nabla_{\tau_i} J$ are

$$\nabla_{v_i} J = \left[\frac{\partial K}{\partial \mathbf{x}} \right]_{t_f} [\Psi^{-1}(t_f, t_0)]^T \int_{\tau_{i-1}}^{\tau_i} \Psi^T(\tau, t_0) \left(\frac{\partial \mathbf{f}}{\partial u} \right) d\tau \quad (17)$$

$$\begin{aligned} \nabla_{\tau_i} J = & \left[\frac{\partial K}{\partial \mathbf{x}} \right]_{t_f} [\Psi^{-1}(t_f, t_0)]^T \Psi^T(\tau_i, t_0) \{ \mathbf{f}(\mathbf{x}, v_i, \tau_i) \\ & - \mathbf{f}(\mathbf{x}, v_{i+1}, \tau_i) \} \end{aligned} \quad (18)$$

Note that the term v -mode means that the search will be done with v , and τ -mode implies the same for τ .

4. Illustrative Examples

Two optimal control problems of chemical reaction process have been solved as illustrative examples. From the aspects of computational effectiveness as well as technical feasibility, comparison is made with two kinds of direct search method and the conjugate gradient method in functional space.

(A) The optimal temperature control of batch reactor

with the reaction $A \xrightarrow{k_1} B \xrightarrow{k_2} C$.

For this well known problem, the optimal control problem is formulated as follows.

$$\text{Max}_T J = x_2(t_f)$$

$$\text{subject to } \begin{cases} dx_1/dt = -k_1 x_1, & x_1(0) = x_{10} \\ dx_2/dt = k_1 x_1 - k_2 x_2, & x_2(0) = x_{20} \end{cases}$$

where x_1 and x_2 are the concentration of components A and B respectively and k_1 and k_2 are reaction rate constants of Arrhenius type.

$$k_i = k_{i0} \exp(-E_i/RT) \quad (i=1, 2) \quad (19)$$

(B) The optimal catalyst blending problem

This was studied originally by Jackson³⁾ for the singular control problem. The problem presented is as follows.

In an isothermal tubular reactor a catalytic reaction is carried out by the following reaction scheme.

The catalyst consists of a substance which catalyzes the two reactions $A \rightleftharpoons B$ and $B \rightarrow C$ differently. Representing this catalyst blend ratio by u , the system equations can be written as

$$\begin{cases} dx/dt = u(k_2 y - k_1 x), & x(0) = x_0 \\ dy/dt = u(k_1 x - k_2 y) - (1-u)k_3 y, & y(0) = y_0 \end{cases} \quad (20)$$

where x and y denote the mole fraction of components A and B respectively and t is the residence time from the instant of entry into the reactor. The aim of this process is to produce the maximum yield of desired product C at the reactor outlet. Now the

Fig. 2 Comparison of optimal control between equal and unequal division for problem (A)

Fig. 3 Comparison of optimal control among the methods for problem (A)

Table 1 Parameters and constants used for computation

Problem (A)	$t_f = 8.0$ [min]	$R = 2.0$ [cal/mole·°K]
	$k_{10} = 0.535 \times 10^{11}$	$k_{20} = 0.461 \times 10^{18}$ [min ⁻¹]
	$E_1 = 18000$	$E_2 = 30000$ [cal/mole]
	$x_{10} = 0.53$	$x_{20} = 0.43$ [mole/liter]
Problem (B)	$k_1 = k_3 = 1.0$	$k_2 = 10.0$ [hr ⁻¹]
	$x_0 = 1.0$	$y_0 = 0.0$ [—]
	$t_f = 1.0$	[hr]

optimal control problem is presented as

$$\text{Max}_{0 \leq u \leq 1} J = 1 - x(t_f) - y(t_f) \quad \text{subject to Eqs. (20)}$$

The parameters and constants used for computation are tabulated in Table 1.

5. Results and Discussion

If the control policy is confined to piece-wise constant control, a certain direct search method will be available. In the first example, computational effectiveness was compared among the proposed,

Deactivation ratio of PI is defined by $[J_{\text{cont.}}^* - J^*(N)]/J_{\text{cont.}}^*$.

Fig. 4 Deactivation ratio of PI value vs. division number

simplex, and Rosenbrock's methods in the value and the number of evaluations of performance index and the computation time. Calculations were made for

Table 2 Comparison of numerical results

	<i>N</i>	2	3	4	6	8
Proposed	(1)	/0.6792796 /82	/0.6793627 /76	/0.6793860 /31	/0.6794110 /22	/0.6794190 /19
	(2)*	/~6.5 /0.6792369 /19	/~6.0 /0.6793415 /19	/~5.5 /0.6793808 /19	/~2.5 /0.6794102 /19	/~3.0 /0.6794190 /19
		/~2.0	/~2.0	/~2.0	/~2.5	/~3.0
Simplex		/0.6792777 /42	/0.6793640 /93	/0.6793734 /109	/0.6794106 /212	/0.6794133 /279
		/~3.0	/~4.0	/~5.0	/~8.0	/~10.0
Rosenbrock		/0.6792675 /112	/0.6793424 /135	/0.6793795 /240	/0.6793995 /301	/0.6793888 /340
		/~4.5	/~5.0	/~8.0	/~9.5	/~11.0

Read the values in this table as [/value of PI/number of evaluation of PI/computation time [sec]].

* Only v -mode search has been done in equally divided interval.

several different numbers of partition and these results are shown in Figs. 2-4 and Table 2.

Although a general conclusion should not be deduced from this example alone, overall superiority of the proposed method to the others can be deduced from the results summarised in Table 2.

In addition, the following properties have been revealed.

Though the simplex method presented good performance for low-dimensional search, high dimensional search was extremely time-consuming, as is the case with all direct search methods. On the other hand, the computation time of the proposed method depended little on the partition number since it does not require additional computation for calculating the gradient by Eq. (17) or Eq. (18).

It may be asserted that for a large number of partitions the τ -mode search is not necessarily essential, but for a low number it becomes useful since a fairly large portion left "unoptimamized" is expected to be recovered.

In the second example, point of view was focussed on technical feasibility, which is another aspect of the proposed method.

For this, the solution of the proposed method was compared with the continuous one determined by the conjugate gradient method of Fletcher and Reeves⁵⁾. These results are shown in Figs. 5 (a) and (b).

Apparently, blending the catalyst to meet the continuous policy is actually impossible. Against this the piece-wise constant control policy can be easily established in actual processes. In addition, since nearly the same performance was attained with much less computation time than with the continuous method, the proposed approach is considered a reasonable choice for practical use.

Computation was made by using FACOM M-190 at Data Processing Center of Kyoto University.

(a) Proposed method
[PI=0.048044, Computing time = ~4 sec]

(b) Conjugate gradient method (Fletcher and Reeves)
[PI=0.048053, Computing time = ~11 sec]

Fig. 5 Optimal blending policy of catalyst

Conclusion

Through sensitivity analysis for a discontinuous function a method has been presented for computing optimal control in a practical manner. From this viewpoint, selecting the admissible control set in the piece-wise constant set, a rational method was developed.

Although an algorithm has been presented only for the steepest-descent method in this paper, the conditions shown here are amenable to other methods, such as another kind of gradient method or the maximum sensitivity method¹²⁾.

Nomenclature

A, B, C	=	chemical species
E_i	=	activation energy ($i=1, 2$) [cal/mole]
f	=	n -dimensional function
Δf	=	n -dimensional function defined by Eq. (6)
h	=	step width of one-dimensional search
I	=	$n \times n$ -dimensional unit matrix
$J, \tilde{J}$	=	performance index
K	=	scalar function
k_1, k_2, k_3	=	reaction rate constants [min ⁻¹], [hr ⁻¹]
k_{i0}	=	frequency coefficient ($i=1, 2$) [min ⁻¹], [hr ⁻¹]
N	=	partition number
R	=	gas constant [cal/mole·°K]
T	=	temperature [°K]
t	=	independent variable of the system [min], [hr]
U	=	admissible control set
u	=	control variable
v	=	value of constant control variable
w	=	a scalar discontinuous function of first-kind
x	=	n -dimensional state vector
x, y	=	concentration or mole fraction [mole/liter], [—]
ϵ_1, ϵ_2	=	small positive real numbers
θ	=	real number
τ	=	independent variable of the system
Φ	=	transition matrix
Ψ	=	adjoint matrix of Φ
$\lambda_i, \bar{\lambda}_i$	=	Lagrangian multipliers ($i=1, \dots, N$)

<Subscripts>

conti.	=	continuous
f	=	final
i	=	i -th value
0	=	initial

<Superscripts and symbols>

T	=	transpose of vector or matrix
*	=	optimal
-1	=	inverse of matrix
(k)	=	k -th step
$\sim$	=	deviated value
δ	=	deviated amount
$\nabla_x J$	=	gradient of J with x

Literature Cited

- 1) Belanger, P. R.: *IEEE. Trans. AC.*, **AC-11**, 77 (1966).
- 2) Crescitelli, S. and B. Nicoletti: *Chem. Eng. Sci.*, **28**, 463 (1973).
- 3) Jackson, R.: *J. Opt. Theory Appl.*, **2**, 27 (1968).
- 4) Kuwabara, M., Y. Kotera and J. Tsukamoto: *System and Control*, **17**, 62 (1973).
- 5) Lasdon, L. S., S. K. Mitter and A. D. Waren: *IEEE Trans. AC.*, **AC-12**, 132 (1967).
- 6) Lee, T. S.: "Quasilinearization and Invariant Imbedding", Chap. 5, Academic Press, New York (1968).
- 7) Nelder, J. A. and R. Mead: *Computer J.*, **7**, 308 (1965).
- 8) Ngo, N. T.: *Auto. Remote Control*, **31**, 735 (1971).
- 9) Rootenberg, J. and P. Courtin: *Int. J. Control*, **18**, 537 (1973).
- 10) Rosenbrock, H. H.: *Computer J.*, **3**, 175 (1960).
- 11) Rozenvasser, E. N.: *Auto. Remote Control*, **28**, 400 (1967).
- 12) Takamatsu, T., I. Hashimoto, H. Nishitani and S. Tomita: *Chem. Eng. Sci.*, **31**, 705 (1976).
- 13) Weber, A. P. J. and L. Lapidus: *AIChE J.*, **17**, 649 (1971).

(Presented at the 10th Autumn Meeting of The Soc. of Chem. Engrs., Japan, at Nagoya, October, 1976).