3-1975

A General Solution for Linear Decision Rules: An Optimal Dynamic Strategy Applicable Under Uncertainty

George A. Hay
Cornell Law School, george.hay@cornell.edu

Charles C. Holt

Recommended Citation

A GENERAL SOLUTION FOR LINEAR DECISION RULES: AN OPTIMAL DYNAMIC STRATEGY APPLICABLE UNDER UNCERTAINTY[1]

By George A. Hay and Charles C. Holt[2]

Linear decision rules for controlling complex systems are often obtained by matrix inversion, but transform methods offer an alternative approach that yields insights into the structure of the decision problem of maximizing expected payoffs under constraints.

1. INTRODUCTION

PROBLEMS OF DECISION MAKING and control involving dynamic relationships and uncertainty have in recent years increasingly been quantified and analyzed mathematically to obtain optimal solutions. This paper presents one such analysis which is applicable to large and complex decision problems and yet is readily computable. The combination of flexibility and simplicity is attained by using quadratic functions to approximate the decision maker's welfare function that expresses his objectives, and by using linear difference equations in discrete time to approximate the dynamic relationships of the system being controlled and any equality constraints on the controlled variables. Since the first order conditions for an optimum are linear equations, the solutions can be obtained by using the well-developed body of linear mathematics. The paper develops a general method of solution utilizing a technique known as the z-transform.

Although this decision analysis will yield a decision rule that is optimal under conditions of uncertainty in the sense of minimizing (or maximizing) the expected value of the decision criterion, the analysis is carried out entirely in terms of certainty as if the future were known exactly. This seeming paradox is explained by the fact that linear decision rules have been proved to have the property of certainty equivalence [26, 28, and 29]. In brief, this means that the decision rules obtained for the certainty problem are also optimal for the uncertainty problem. Expected value forecasts of the uncontrolled variables are substituted for perfect forecasts in the decision rules.[3] In the following analysis we treat the future uncontrolled variables as if they were "known," but of course in the final application of the decision rules these variables will be replaced by forecasts. The ability to separate out the probability analysis and treat it by a separate proof *greatly* simplifies the present analysis. For contrast, see some of the automatic control literature, for example [19 or 32].

---

[1] An earlier version of this paper was presented and discussed at the September, 1971 European Congress of the Econometric Society in Barcelona, Spain.

[2] Charles C. Holt's research for this paper was supported by the National Science Foundation.

[3] Durbin [6] has considered the problem that is posed when sampling or other errors occur in estimating the expected values, and Holt has considered optimal methods for estimating forecast relationships [12].

This paper presents a general method which can be applied in a straightforward manner to a class of decision problems that involve complicated objectives, many uncontrolled variables impinging on the system, some of them random, many control (or instrument) variables under the direction of the decision maker (or automatic controller), many partly controlled variables, and many complex relationships linking the controlled, the partly controlled, and the uncontrolled variables.[4]

This type of decision analysis has already been applied to a rather wide range of problems, and in the industrial area it has already reached the stage of practical applications, some of which are reported in [13] together with related analyses. In the area of positive economic theory, several authors [4, 5, 9, and 20] have used the analysis to derive the implications of profit maximizing behavior for firms' decisions regarding variables such as investment, wages, vacancies, output, inventories and price. The models generated predictions of the coefficients in regression equations describing industry aggregates of behavioral relations. These predictions have turned out to be reasonably accurate.

Similar decision models have been obtained independently by three research groups: H. Theil, Kalman and associates, and Holt and associates. Theil developed his analysis working on problems of economic policy using matrix inversion for obtaining optimal decisions [27 and 28]. Kalman, Lapidus, and Shapiro developed their analysis for the optimization of industrial process controls [18] using Bellman's dynamic programming approach [2]. Holt, Modigliani, Muth, and Simon developed their analysis, an earlier version of the one presented in this paper, working on business decision problems in the area of production, inventory control, and employment [13, 14, 15, and 16]. Subsequent work on industrial decisions using continuous time has been done by Schild [25], on railroad scheduling by Mansfield and Wein [21], and on economic stabilization by Theil [28], van den Bogaard and Barten [30], van den Bogaard and Theil [31], and Holt [11].

Although a great deal of work has been done on optimization problems in the fields of operations research and even more in automatic control engineering, the knowledge base is not yet fully consolidated even for the relatively simple case in which the criterion function is quadratic and the constraints are linear equations. With the work listed in the above paragraph by Bellman, Theil, Simon, Kalman, Muth, and Holt, the basic work on the linear decision analysis was essentially complete for optimal strategies under uncertainty. However, the solution for optimal strategies is inherently complex and expositional generalization and clarity are needed. In addition, the refinement of solution methods and the writing of general purpose computer programs are needed.[5]

---

[4] Economists have commonly drawn a dichotomy between exogenous and endogenous variables. In this paper we need three classifications, but they can readily be related to the dichotomy. The uncontrolled and the controlled variables are exogenous to the system, and the partly controlled variables are endogenous. The analysis yields decision rules (the compensating networks in control parlance), which are the dynamic relations that govern the controlled variables.

[5] Computer programs for the analysis in this paper have been designed by the authors, but Hay's program has a small capacity and is not documented, and Holt's large and general program has not been quite finished (in its treatment of zero roots).

Some sophisticated textbooks on control engineering still neglect to present Simon's certainty equivalence theorem [1] or imply that it applies only to Gaussian distributions [3]. Also the engineering orientation toward servomechanisms and regulators is, in some respects, less general than the decision analysis formulation.

Much of the engineering control literature is developed in terms of linear differential equations, but economists and operations researchers are more interested in discrete time and difference equations. Even though the z-transform is well known to control engineers, it is less so to economists and operations research people, who could use it for solving decision problems in discrete time.

The formulation that is presented here makes exogenous variables explicit and hence allows the isolation and general treatment of the forecasting problem. This has not yet been done in the control literature, although it is straightforward. By using the z-transform, the inversion of large matrices, à la Theil, is avoided, but this comes at the price of searching numerically for the roots of a polynomial. The Kalman matrix iteration approach has been advantageously applied to the decision problem posed here. Since this paper deals with a decision analysis formulation with time discrete and uses a z-transform solution, we hope that it will contribute to the interdisciplinary consolidation of knowledge that is badly needed.

The analysis is presented as follows. Section 2 states the formal problem, and Section 3 gives the solution for the decision rules. Section 4 illustrates the technique with a simple example, whose concreteness should serve to illuminate both the notation and the solution procedure of the general derivation which, admittedly, is a bit hard to follow. Section 5 briefly supplements and extends the analysis, and Section 6 concludes.

2. THE DECISION PROBLEM

The decision problem is to find the values of the control variables that minimize (or maximize) the decision maker's welfare taking into account the constraint relationships that exist between the controlled, partly controlled, and uncontrolled variables. Welfare is measured by a function which serves as the criterion for making the decisions. The uncontrolled variables constitute the decision maker's external environment in response to which he seeks the best adjustment by setting the value of the controlled variables. The partly or indirectly controlled variables that occur in the criterion function are determined by certain linear relationships with the controlled and uncontrolled variables. Other linear relationships can apply constraints to the decision maker in setting his control variables. The decision problem is to set the control variables so that through their interactions with the partly controlled and uncontrolled variables the best possible results are obtained as measured by the criterion function. The decision model can now be stated precisely.[6]

[6] The example of Section 4 parallels this theoretical exposition including the use of identical equation numbers. The example should aid in interpreting both the notation and operations.

The variables are of three types:

x_{it} $(i = 1, 2, \ldots, N_x)$: the ith controlled (or instrument) variable in the tth period. These variables are under the control of the decision maker.

y_{jt} $(j = 1, 2, \ldots, N_y)$: the jth partly controlled (or endogenous) variable in the tth period. These variables can be influenced by the decision maker through their relationships with the controlled variables and similarly are influenced by the uncontrolled variables.

v_{kt} $(k = 1, 2, \ldots, N_v)$: the kth uncontrolled (or exogenous) variable in the tth period. These variables are determined outside the system under consideration. They may be forecast by the decision maker but cannot (for purposes of this analysis) be influenced by him.

The variables should be chosen in such a way as to simplify as much as possible the criterion function and the relationships. Usually this will indicate avoiding variables that appear in continuing sums because this introduces a large number of terms. Such a sum can be defined as a new variable, thereby greatly decreasing the number of terms.

The criterion for making decisions is assumed to be a function that puts weight (or measure) on the possible outcomes indicating their desirability or undesirability. This common measure of welfare, utility, cost, profit, output, quality, or other objective enables quantitative comparisons to be made between alternative courses of action. This statement of the decision maker's value judgments enables a meaningful choice to be made of the best course of action. Weight may attach directly to the values given the control variables as well as to the state of the system which is the joint consequence of the decision maker's actions and the uncontrolled environment.

We assume that the criterion function can be adequately approximated in the region of its constrained optimum by a quadratic function of the controlled, partly controlled, and uncontrolled variables:

$$(1) \qquad C \equiv \sum_{t=0}^{T} C_t$$

where T is a large finite integer and[7] where

$$C_t \equiv \sum_{P \in \bar{P}} \sum_{\gamma} \sum_{Q \in \bar{Q}} \sum_{\delta} (K_{PQ\gamma\delta} P_{t-\gamma} Q_{t-\delta}),$$

$$\bar{P}, \bar{Q} \equiv x_1, x_2, \ldots, x_{N_x}, y_1, y_2, \ldots, y_{N_y}, v_1, v_2, \ldots, v_{N_v},$$

$$(x_i)_t = x_{it}, \quad (y_j)_t = y_{jt}, \quad (v_k)_t = v_{kt},$$

$$\gamma, \delta = 0, 1, 2, \ldots, L.[8]$$

[7] Alternatively, a discount factor can be added and T allowed to go to infinity. This is treated in Section 5.

[8] For specific solutions, computer storage space can be conserved by specifying the maximum lag for each of the variables.

The component of the criterion from the tth period, C_t, is a quadratic function including square, cross product, linear, and constant terms, all with constant coefficients $K_{PQ\gamma\delta}$. The linear and constant terms are admitted by holding one of the uncontrolled variables constant, i.e., $V_{1t} = 1$ ($t = 0, 1, 2, \ldots, T$). At this stage we assume that the C_t function is identical for all time periods. This assumption will be somewhat relaxed later.

Note that the welfare in the tth period depends only on current and those past values up to the maximum lag L of the variables, not on future values.

The constraint relations between the variables are assumed to be adequately approximated locally by a set of linear functions:

$$(2) \qquad F_{ft} = 0$$

where

$$F_{ft} \equiv \sum_{P \in \bar{P}} \sum_{\gamma} (H_{fP\gamma} P_{t-\gamma}) = 0,$$

$$\bar{P} = x_i, x_2, \ldots, x_{N_x}, y_1, y_2, \ldots, y_{N_y}, v_1, v_2, \ldots, v_{N_v},$$

$$f = 1, 2, \ldots, N_f,$$

$$\gamma = 0, 1, 2, \ldots, L,$$

$$t = 0, 1, 2, \ldots, T.$$

These relationships reflect the fact that the partly controlled variables of a period are jointly and uniquely determined by the controlled and uncontrolled variables of the current and previous periods. (Thus the lag γ is limited to positive values.) To ensure this uniqueness, we require that a non-singular square matrix of rank N_y can be made up from the coefficients of the partly controlled variables. Thus, we cannot have fewer equations in (2) than partly controlled variables, $N_f \geqslant N_y$.

All of the control variables may not be free to move independently; specifically, there may be additional linear constraints governing the control variables so that there are not N_x degrees of freedom open to the decision maker. This situation would appear as a larger number of constraint relations than there are partly controlled variables. However, there is an upper limit to the number of equations that is reached when the decision maker has no degrees of freedom left and hence no decision problem $N_x + N_y > N_f$. The constraint relations must be independent and not inconsistent. This requirement combined with the determinantal requirement above means from (2) for $t = 0$ a matrix of rank N_f can be made up from the coefficients of all N_y of the partly controlled variables y_{j0} ($j = 1, \ldots, N_y$) and ($N_f - N_y$) different controlled variables x_{i0}.

At this point we assume that the same relations apply to each period but some relaxation of this will be discussed later.

Constant terms can appear in the relations by holding one of the uncontrolled variables constant as before.

We assume that the decision analysis is made at the beginning of period $t = 0$. Thus negative values of t designate the past which is known, and the periods $(t = 0, 1, 2, \ldots, T)$ constitute the future.

The formal statement of the decision problem is now complete except for conditions that will ensure the existence of a unique solution, but because of their complexity they will be discussed later. The problem can now be summarized as follows: given the initial conditions of a dynamic system and perfect forecasts of the future values of the uncontrolled variables, find the values of the controlled variables that minimize (maximize) the criterion subject to the constraint relations.

Since in the uncertain dynamic case we would take action only in the next period on the basis of current knowledge, we need in this analysis to find the solution only for the actions in the period $t = 0$. The solution that is desired is a general one in the form of an optimal decision rule that can be applied repeatedly to find the best values for the control variables period after period.

We might remark in passing that the above model is suitable for dealing with decision problems that can be characterized as being "smooth" or "continuous." If sharp discontinuities or inequality constraints are important in a decision problem, the above model would be likely to give a poor approximation. However, decision problems that involve substantial departures from linearity and quadraticity in the large may not raise serious difficulties, if in spite of disturbances to the system the control decisions that have been based on the approximations are able to avoid extreme fluctuations in the variables, and hence do not encounter the discontinuities or violate the constraints.[9]

3. THE SOLUTION FOR THE DECISION RULES

In this section we carry through the main development of the solution leaving aside in the interests of simplicity certain miscellaneous technical points for consideration later in Section 5.

We can summarize the decision problems as one of finding the minimum (or maximum) of the criterion (1) subject to the constraint of the relations (2). This can be converted into an unconstrained minimization (maximization) problem by the use of Lagrange multipliers.[10] We desire to minimize (maximize)

$$(3) \qquad C(\lambda) \equiv \sum_{t=0}^{T} \left(C_t + \sum_{f=1}^{N_f} \lambda_{ft} F_{ft} \right)$$

over x_{it}, y_{jt}, λ_{ft} $(i = 1, 2, \ldots, N_x; j = 1, 2, \ldots, N_y; f = 1, 2, \ldots, N_f)$ where λ_{ft} is the Lagrange multiplier corresponding to the constraint $F_{ft} = 0$. In this certainty analysis all the uncontrolled variables v_{kt} $(k = 1, 2, \ldots, N_v; t = 0, 1, 2, \ldots, T)$ are treated as if they were known and, of course, past values of x_{it}, y_{jt}, v_{kt} are known initial conditions for $t < 0$.

[9] Although it is impossible within the model to impose inequality limits on the excursions of variables, the effect of a limit can be approximated by introducing a quadratic cost term that rises steeply as the excursion becomes excessive. In this way excessive excursions can be minimized.

[10] See [10] or a textbook on advanced calculus.

First Order Conditions

First order conditions for finding the minimum (maximum) of (3) are obtained by equating to zero the partial derivatives of $C(\lambda)$ with respect to the unknown x and y components of P and Q. We consider derivatives with respect to these variables in a generalized future period t.

$$(4) \qquad \frac{\partial C(\lambda)}{\partial P_{mn}} = \sum_{Q \in \bar{Q}} \sum_{\delta} \sum_{\gamma} (K_{PQ\gamma\delta} + K_{QP\delta})Q_{t+\gamma-\delta}$$

$$+ \sum_{f} \sum_{\gamma} (H_{fP\gamma}\lambda_{f,t+\gamma}) = 0 \quad (P_{mn} \in \bar{\bar{P}})$$

where

$$\bar{\bar{P}} = (x_1, x_2, \ldots, x_{N_x}, y_1, y_2, \ldots, y_{N_y})_t \qquad\qquad (t = 1, 2, \ldots, T),$$

$$\bar{Q} = x_1, x_2, \ldots, x_{N_x}, y_1, y_2, \ldots, y_{N_y}, v_1, v_2, \ldots, v_{N_v},$$

$$\delta, \gamma = 0, 1, 2, \ldots, L,$$

$$f = 1, 2, \ldots, N_f,$$

$$t = 0, 1, 2, \ldots, T,$$

and where

$$Q_{t>T} = \lambda_{f,t>T} = 0 \quad \text{for all } Q \in \bar{Q} \text{ and } f.$$

These conditions together with the constraint relations (2) constitute the first order conditions for an optimum. They can be written in matrix form as

$$(5) \qquad A_{1t}[V_{tL}] + A_{2t}\begin{bmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}_{tL} \\ \lambda_{tL} \end{bmatrix} = [0] \qquad\qquad (t = 0, 1, 2, \ldots, T)$$

where

$$V_{tL} = \begin{bmatrix} V_{t-L} \\ \vdots \\ V_{t+L} \end{bmatrix}, \quad \begin{pmatrix} X \\ Y \end{pmatrix}_{tL} = \begin{bmatrix} X_{t-L} \\ Y_{t-L} \\ \vdots \\ X_{t+L} \\ Y_{t+L} \end{bmatrix}, \quad \lambda_{tL} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \lambda_t \\ \vdots \\ \lambda_{t+L} \end{bmatrix},$$

$$V_T = \begin{bmatrix} v_{1t} \\ \vdots \\ v_{N_v t} \end{bmatrix}, \quad X_T = \begin{bmatrix} x_{1t} \\ \vdots \\ x_{N_x t} \end{bmatrix}, \quad Y_T = \begin{bmatrix} y_{1t} \\ \vdots \\ y_{N_y t} \end{bmatrix}, \quad \lambda_t = \begin{bmatrix} \lambda_{1t} \\ \vdots \\ \lambda_{N_f t} \end{bmatrix}$$
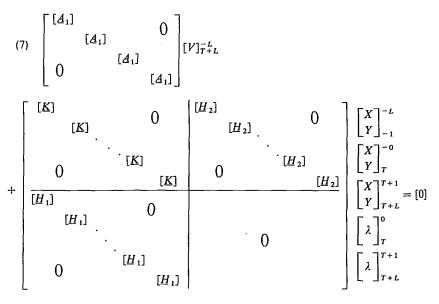
and all of the matrices A_{1t} and A_{2t} would be identical respectively for all t except that no variables from the periods beyond T occur in (3), and hence the corresponding matrix elements are zero. Without actually changing the equations, we can assign zero values to all variables in the periods beyond T. By so doing, we are free to assign arbitrary values to the corresponding matrix elements. We chose to assign values to these elements so that A_{1t} $(t = 0, 1, 2, \ldots, T) = A_1$ and A_{2t} $(t = 0, 1, 2, \ldots, T) = A_2$. The matrix A_2 can be expanded to show its structure more clearly:

$$(6) \qquad A_2 = \begin{bmatrix} K & H_2 \\ H_1 & 0 \end{bmatrix}$$

where the matrices K, H_1, and H_2 are defined by reference to (2), (4), and (5). The elements of K come from the criterion function, and the elements of H_1 and H_2 come from the constraint relations.

We desire the solution of the system of equations (5) for X_0, the optimal action vector of the period $t = 0$.

The set of matrix equations (5) can now be written as a single matrix equation:

$$(7) \quad \begin{bmatrix} [A_1] & & & \\ & [A_1] & & 0 \\ & & [A_1] & \\ 0 & & & [A_1] \end{bmatrix} [V]_{T+L}^{-L}$$

$$+ \begin{bmatrix} [K] & & & 0 & [H_2] & & & 0 \\ & [K] & & & & [H_2] & & \\ & & \cdot & [K] & & & \cdot & [H_2] \\ 0 & & & [K] & 0 & & & [H_2] \\ \hline [H_1] & & & 0 & & & & \\ & [H_1] & & & & & 0 & \\ & & \cdot & [H_1] & & & & \\ 0 & & & [H_1] & & & & \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix}_{-1}^{-L} \\ \begin{bmatrix} X \\ Y \end{bmatrix}_{T}^{-0} \\ \begin{bmatrix} X \\ Y \end{bmatrix}_{T+L}^{T+1} = [0] \\ [\lambda]_{T}^{0} \\ [\lambda]_{T+L}^{T+1}$$

where there are T (one for each time period) diagonal submatrices with overlapping columns in each block, and time limits have been added to the vectors defined in (5) to indicate the number of time periods included in each vector. The right vector has been partitioned to separate the known from the unknown elements.

The block of diagonal H_1's will be found to be the transpose of the diagonal block of H_2's.

Second Order Conditions

The vectors

$$\begin{bmatrix} X \\ Y \end{bmatrix}_T^0 \quad \text{and} \quad \begin{bmatrix} \lambda \end{bmatrix}_T^0$$

contain all of the unknowns, and we desire to determine whether a unique and optimal solution for them exists. By selecting those columns in (7) that multiply the above vectors we obtain a matrix to which we might refer as the test matrix. Uniqueness of solution would be ensured simply by the nonsingularity of the test matrix. However, it is quite possible in a complex situation through errors of statistical estimation or an inadequately comprehensive criterion function to set up a nonsense decision problem that has no unique optimum. Hence, it is highly desirable to have a method for detecting this situation at the outset. A unique optimal solution would be ensured by the positive (negative) definiteness of the block of the test matrix having a diagonal of K matrices. This is a sufficient condition for the existence of an optimal solution, but it is not necessary. Even though this test fails, it is still possible by considering the constraint relations to find that a unique optimum exists.

The relevant test involves sign conditions on the determinants of certain matrices made up in a rather complex way from the test matrix. We will avoid discussing this test further by citing Samuelson,[11] who has analyzed the general case.

Satisfying the second order conditions ensures us that we have not framed a decision problem in which an infinite number of equally good decisions exist[12] or one that allows the bliss of infinite welfare to be obtained by giving some of the control variables infinite values. The latter would clearly imply a nonsense decision problem but not necessarily the former.

Assuming that an optimal solution exists, we return to the problem of obtaining a solution[13] for the first order conditions.

Transform methods offer an approach to the solution which as we shall see is actually simplified by having a large number of periods in the criterion function. These methods enable us to concentrate on solving for the initial actions without considering explicitly most of the other unknowns.

[11] See [24, pp. 376–379]. If necessary, the K block of the test matrix may be readily put in symmetric form.

[12] This case would give us fewer equations than unknowns and will not be considered here.

[13] Usual matrix inversion methods could be used for obtaining the solution (see [28 and 30]) except that the system of equations may be of unwieldy size. Also, if we think of equations (5) without time subscripts on the matrices as difference equations representing the dynamics of a system, it would be an inherently unstable system. This may cause difficulty for numerical matrix inversion methods by producing a tendency for round-off errors to cumulate excessively.

Initial Equations

If we consider in (5) the first order conditions obtained for the initial period ($t = 0$), we obtain equations that include the controlled variables in the initial period for which we desire a solution:

$$(8) \qquad A_1[V_{01}] + A_2 \begin{bmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}_{0L} \\ \lambda_{0L} \end{bmatrix} = [0].$$

Unfortunately, we would have only ($N_x + N_y + N_f$) equations but a larger number of unknown X's, Y's, and λ's since each of the variables will appear not only with the subscripts of $t = 0$ and $t < 0$ (and therefore already known) but also with subscripts of $t > 0$ as a result of the intertemporal nature of some of the constraints and the existence of lags in the criterion, C_t.

To move toward a complete solution, we note that the repetitive nature of these equations for all t ($t = 0, 1, 2, \ldots, T$) suggests that we might consider a set of equations relating *sequences* of the time series variables. Indeed, because the first order conditions are identical for every time period, there is, for each set of T equations, an additional relation among the sequences of variables. Moreover, the restrictions placed on the decision variables for $t = 0$ by the set of additional relations among the sequences of variables will be precisely enough to make up the remainder of the ($N_x + N_y + N_f$) equations needed to solve for the initial period's decision variables.

We turn to transform methods to obtain the missing equations. The method we shall use is to transform the set of matrix equations (5) into a single matrix equation. Then, reducing a certain determinant to zero by setting the transform parameter equal to the roots of the determinant, we eliminate all of the unknown transform variables. This leaves us with a set of equations sufficient to complete (8).

z-Transform[14]

Consider a time series variable M_t, for example, that takes a sequence of real bounded values at successive points of time ($\ldots, M_{-2}, M_{-1}, M_0, M_1, M_2, \ldots$). Its one-sided truncated generating function transform is defined as follows:

$$(9) \qquad G[M_t(t = 0, 1, 2, \ldots, T), z] \equiv GM_t \equiv \sum_{t=0}^{T} M_t z^t$$

where z is a finite real or complex number. In the latter case, GM_t is complex also. We will refer to GM_t (which is a function of the transform parameter z) as a transform variable to distinguish it from M_t which is a time series variable.

---

[14] For a mathematical discussion of the z-transform, see [8 and 17].

Variables with leads and lags are transformed similarly:

$$(10) \qquad GM_{t-p} \equiv \sum_{t=0}^{T} M_{t-p} z^t = \sum_{t=-p}^{T-p} M_t z^{t+p}$$

$$= z^p \left[\sum_{t=-p}^{-1} M_t z^t + GM_t - \left(\sum_{t=T-p+1}^{T} M_t z^t \right) \right]$$

and

$$(11) \qquad GM_{t+p} \equiv \sum_{t=0}^{T} M_{t+p} z^t = \sum_{t=p}^{T+p} M_t z^{t-p}$$

$$= \frac{1}{z^p} \left[-\sum_{t=0}^{p-1} M_t z^t + GM_t + \left(\sum_{t=T+1}^{T+p} M_t z^t \right) \right]$$

where p is a positive integer which indicates the number of periods of lag or lead.

In this analysis the last term of (11) can be neglected because the variables with subscripts greater than T are set equal to zero. The last term in (10) will be very small if T is large and $|z| < 1$. It is quite reasonable to assume that the criterion contains a large number of periods and we shall see later that the latter condition also is satisfied. Hence, we will simplify the analysis at the cost of introducing only a negligibly small error by dropping the last terms of both (10) and (11).

The V's and the other variables can grow exponentially or otherwise in time provided that they do not become infinite within the T time periods and provided they are not growing faster than can be attenuated by the factor z^t. Otherwise, the last terms of the transforms cease to be negligible.

It is clear from (9) that $GKM_t = KGM_t$ where K is a constant.

Transform Equations

We proceed by multiplying each of the first order equations (5) for time period $t(t = 0, 1, 2, \ldots, T)$ by z^t where z is an arbitrary number within the context of certain restrictions to be discussed below. The set consisting of the T versions of any particular equation can then be compressed into a single equation by converting each of the sequences of time series variables into a transform variable, which is a function of z, and a few additional time series terms. Thus by applying the generating function transform to each of these sequences of variables we can obtain $(N_x + N_y + N_f)$ equations relating the transformed variables and a few time sequence variables. We can write the transform equations in matrix form using the vectors of (5) except that we need also to provide for vectors of transform variables. Since the V's are known, future as well as past, their transforms can be evaluated, but the other transform variables are unknown. Reflecting this we separate the known from the unknown variables in the transform equations:

$$(12) \qquad \underline{B}_0(z)GV_t + \underline{B}_1(z)V_{tL} + \underline{B}_2(z) \begin{bmatrix} \begin{pmatrix} X \\ Y \\ \lambda_{tL} \end{pmatrix}_{tL} \end{bmatrix} + \underline{B}_3(z)G \begin{bmatrix} X_t \\ Y_t \\ \lambda_t \end{bmatrix} = [0]$$

where the matrices $\underline{B}_0(z), \ldots, \underline{B}_3(z)$ have elements that are polynomials in z and $1/z$ arising from the application of the transform formulas (9), (10), and (11), and the transform operator G is generalized to apply to vectors of time series variables.

Eliminating the Unknown Transform Variables

Since the last term of (12) includes the unknown transform variables which in turn depend on whole sequences of unknowns, it is almost essential to supress this term. Fortunately this can be done. Premultiply (12) by the adjoint of $\underline{B}_3(z)$ to obtain the determinant $|\underline{B}_3(z)|$ times the unit matrix I as the coefficient of the last term, i.e.,

$$(13) \qquad \underline{B}_3(z)^a \left\{ \underline{B}_0(z)GV_t + \underline{B}_1(z)V_{tL} + \underline{B}_2(2) \begin{bmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}_{tL} \\ \lambda_{tL} \end{bmatrix} \right\} + |\underline{B}_3(z)|IG \begin{bmatrix} X_t \\ Y_t \\ \lambda_t \end{bmatrix} = 0.$$

Since the relations among the sequences of variables must hold for any arbitrary value of z, they must hold for the particular values of z which make $|\underline{B}_3(z)| = 0$. Thus we find the values of z which satisfy the determinantal equation

$$(14) \qquad |\underline{B}_3(z_r)| = 0$$

where z_r $(r = 1, 2, \ldots, 2R)$ are the roots, and $2R$ is the total number of roots.[15]

By substituting z_r into (13) the determinant will go to zero, and this will cause the last term to vanish if its vector of transform variables contains no infinite elements, i.e., all z_r, x_{it}, y_{jt}, and λ_{ft} are bounded. To ensure this, we need to impose the restriction that all v_{kt}'s are bounded for all time periods $(t = 0, 1, 2, \ldots, T)$. The unique solution of the complete system of equations (7) then ensures the boundedness of the X's, Y's, and λ's. The boundedness of the relevant z_r's will be demonstrated later.

The substitution of z_r into (13) has the additional effect of reducing the remaining product matrices to unit rank (see [7, Theorem D, p. 61]) so that the matrix equation contains but a single independent equation. This equation can be found by evaluating one row of $\underline{B}_3(z)^a$ (it makes no difference which is selected) and carrying out the matrix multiplications for this one row. The equation thus obtained after substitution of the root z_r is:

$$(15) \qquad c'_{0r}G_rV_t + c'_{1r}V_{0L} + c'_{2r} \begin{bmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}_{0L} \\ \lambda_{0L} \end{bmatrix} = [0]$$

---

[15] The procedure that the Holt program uses for finding the roots of (14) is to substitute a trial value of z_i in (12) and eliminate all of the transform variables but one. Its coefficient is zero when z_i is a root. Numerical gradient methods are used to find a root and a transformation of the equation is made to avoid finding the same root again.

where c_{0r}', c_{1r}', and c_{2r}' are now row vectors with constant elements and $G_r V_t \equiv G[V_t(t = 0, 1, 2, \ldots, T), z_r]$.

Of the $2R$ roots we select half of them by the criterion $|z_r| < 1$ to be discussed later. Thus we obtain R equations like (15) which can be written in matrix form

$$(16) \qquad I[c_{0r}' G_r V_t] + \underline{C}_1 V_{0L} + \underline{C}_2 \begin{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix}_{0L} \\ \lambda_{0L} \end{bmatrix} = [0]$$

where $\underline{C}_1$ and $\underline{C}_2$ are matrices whose rows are c_{1r}' and c_{2r}' respectively and the vector

$$[c_{0r}' G_r V_t] \equiv \begin{bmatrix} c_{01}' G_1 V_t \\ \vdots \\ c_{0R}' G_R V_t \end{bmatrix}.$$

Complex roots which will be considered later require a slight modification of this vector.

Solution for the Decision Rules

We can now combine (8) and (16) into a single matrix equation, the latter supplying the exact number of rows needed to complete the system

$$\begin{bmatrix} c_{0r}' & G_r & V_T \end{bmatrix}$$
$$[V]_{+L}^{-L}$$
$$\begin{bmatrix} X \\ Y \end{bmatrix}_{-1}^{-1}$$

$$(17) \qquad \begin{bmatrix} 0 & A_1 & A_2 \\ I & \underline{C}_1 & \underline{C}_2 \end{bmatrix} \quad --\begin{bmatrix} X_0 \\ Y_0 \end{bmatrix}--- = [0]$$

$$\begin{bmatrix} X \\ Y \end{bmatrix}_{+L}^{1}$$

$$[\lambda]_{+L}^{0}$$

where we have partially expanded the vector to show its elements.

The final solution is now obtained by eliminating the unknowns that we do not need—namely, those below the dotted partition in the vector. We then solve the remaining equations for X_0 by matrix inversion to obtain decision rules of the form

$$(18) \qquad x_{i0} = \sum_{Q \in \bar{Q}} \sum_{\gamma} W_{iQ\gamma} Q_{-\gamma} + \sum_{t=0}^{T} W_{kt} v_{kt},$$

$$L = 1, 2, \ldots, N_x, \quad k = 1, 2, \ldots, N_v,$$

$$\gamma = 1, 2, \ldots, L,$$

$$\bar{Q} = x_1, x_2, \ldots, x_{N_x}, y_1, y_2, \ldots, y_{N_y}, v_1, v_2, \ldots, v_{N_v},$$

where the W_{iQ_γ} and W_{kt} are constants. In this way we obtain for each of the control variables a decision rule which is a linear function of known past values of the control, partly controlled, and uncontrolled variables, and future values of the uncontrolled variables which we have assumed were known but actually will need to be forecast.

Plans for actions in later periods are implicit in the analysis and can be found if needed by applying the above decision rules successively using the dynamic relations in (2) to obtain the initial conditions for the future periods. Of course, in the uncertainty case these plans are tentative and subject to revision each period.

4. AN EXAMPLE

Before going on to consider some more detailed points and extensions, we will illustrate the foregoing analysis with a specific example. The numbering of equations corresponds to the related equation in the preceding sections.

A firm for which sales are determined exogenously is attempting to minimize costs over a T-period decision horizon. The decision variables for period t are production, X_t, and end-of-period inventories, H_t. Sales are denoted as S_t.

The cost structure for the firm is as follows: There is assumed to be a desired level of inventories, H_t^*, which is proportional to sales, i.e.,

$$H_t^* = c_{11}S_t.$$

Moreover, the firm incurs a cost for departing from the desired level which is assumed to be quadratic, i.e.,

$$c_1(H_t - H_t^*)^2.$$

There is also a cost of changing production rate, i.e.,

$$c_2(X_t - X_{t-1})^2.$$

Unit costs of production within a period are assumed constant and are equal to C_{20}.

Thus the cost function for any period t is

$$
\begin{aligned}
(1) \quad C_t &= c_{20}X_t + c_1(H_t - c_{11}S_t)^2 + c_2(X_t - X_{t-1})^2 \\
&= c_{20}X_t + c_1H_t^2 + c_1c_{11}^2S_t^2 - 2c_1c_{11}H_tS_t + c_2X_t^2 + c_2X_{t-1}^2 \\
&\quad - 2c_2X_tX_{t-1}.
\end{aligned}
$$

There is a single constraint:

$$(2) \quad X_t - S_t = H_t - H_{t-1}.$$

The Lagrangean to be minimized is

$$
\begin{aligned}
(3) \quad C(\lambda) = \sum_{t=0}^{T} [(&c_{20}X_t + c_1H_t^2 + c_1c_{11}^2S_t^2 - 2c_1c_{11}H_tS_t + c_2X_t^2 \\
&+ c_2X_{t-1}^2 - 2c_2X_tX_{t-1}) + \lambda_t(X_t - S_t - H_t + H_{t-1})]
\end{aligned}
$$

(4) $\quad \dfrac{\partial C(\lambda)}{\partial H_t} = 2c_1 H_t - 2c_1 c_{11} S_t - \lambda_t + \lambda_{t+1} = 0.$

(The last term comes from differentiating the criterion function in period $t + 1$ with respect to H_t.)

$$\frac{\partial C(\lambda)}{\partial X_t} = c_{20} + 2c_2 X_t - 2c_2 X_{t-1} + \lambda_t + 2c_2 X_t - 2c_2 X_{t+1} = 0$$

$$(t = 0, 1, 2, \ldots, T),$$

(5)

$$
\begin{bmatrix} -2c_1 c_{11} & 0 \\ 0 & c_{20} \\ -1 & 0 \end{bmatrix}
\begin{bmatrix} S_t \\ 1 \end{bmatrix}
+
\left[\begin{array}{ccccc:cc} 0 & 0 & 2c_1 & 0 & 0 & -1 & 1 \\ \hdashline 0 & -2c_2 & 0 & 4c_2 & -2c_2 & 1 & 0 \\ \hdashline 1 & 0 & -1 & 1 & 0 & 0 & 0 \end{array}\right]
$$

$$
\times \begin{bmatrix} H_{t-1} \\ X_{t-1} \\ H_t \\ X_t \\ X_{t+1} \\ \hdashline \lambda_t \\ \lambda_{t+1} \end{bmatrix} = [0].
$$

Applying the transform relation to (4) and the constraint equation (2) yields

$$2c_1 G(H) - 2c_1 c_{11} G(S) - G(\lambda) + \frac{[G(\lambda) - \lambda_0]}{z} = 0,$$

$$\frac{c_{20}}{1 - z} + 2c_2 G(X) - 2c_2[X_{-1} + z G(X)] + G(\lambda) + 2c_2 G(X)$$

$$- 2c_2 \frac{[G(X) - X_0]}{z} = 0,$$

$$G(X) - G(S) - G(H) + [H_{-1} + z G(H)] = 0.$$

Rearranging terms yields:

(12)

$$
\begin{bmatrix} -2c_1 c_{11} \\ 0 \\ -1 \end{bmatrix} [G(S)]
+
\begin{bmatrix} 0 & 0 \\ 0 & \dfrac{c_{20}}{1 - z} \\ 0 & 0 \end{bmatrix}
\begin{bmatrix} S_0 \\ 1 \end{bmatrix}
+
$$

$$+ \begin{bmatrix} 0 & 0 & 0 & 0 & \vdots & \dfrac{-1}{z} \\ 0 & -2c_2 & 0 & \dfrac{2c_2}{z} & \vdots & 0 \\ 1 & 0 & 0 & 0 & \vdots & 0 \end{bmatrix} \begin{bmatrix} H_{-1} \\ X_{-1} \\ H_0 \\ X_0 \\ \lambda_0 \end{bmatrix}$$

$$+ \begin{bmatrix} 2c_1 & 0 & -\left(1 - \dfrac{1}{z}\right) \\ 0 & 2c_2\left(2 - z - \dfrac{1}{z}\right) & 1 \\ -(1 - z) & 1 & 0 \end{bmatrix} G \begin{bmatrix} H \\ X \\ \lambda \end{bmatrix} = [0].$$

The roots are located by solving the determinantal equation:

$$(14) \qquad \begin{vmatrix} 2c_1 & 0 & -\left(1 - \dfrac{1}{z}\right) \\ 0 & 2c_2\left(2 - z - \dfrac{1}{z}\right) & 1 \\ -(1 - z) & 1 & 0 \end{vmatrix} = 0.$$

Expansion of the determinant yields the following equation:

$$(14') \qquad c_2\left(z^2 + \left(\dfrac{1}{z}\right)^2 - 4z - \dfrac{4}{z}\right) + (c_1 + 6c_2) = 0.$$

Solution of the quadratic will yield four roots (complex, if both c_1 and c_2 are positive). Furthermore, it is obvious from the form of $(14')$ that if z is a root, then $1/z$ is also a root. Therefore there will be in general exactly two roots that satisfy both $(14')$ and the convergence requirement $|z| < 1$ (see Section 5, below). Since in this case the two roots will be complex conjugates, we carry forward only one of the pair.

The adjoint of the matrix premultiplying $G[H, X, \lambda]$ equals:

$$\begin{bmatrix} -1 & -\left(1 - \dfrac{1}{z}\right) & 2c_2\left(2 - z - \dfrac{1}{z}\right)\left(1 - \dfrac{1}{z}\right) \\ -(1 - z) & -\left(1 - \dfrac{1}{z}\right)(1 - z) & -2c_1 \\ 2c_2\left(2 - z - \dfrac{1}{z}\right)(1 - z) & -2c_1 & 4c_1 c_2\left(2 - z - \dfrac{1}{z}\right) \end{bmatrix}.$$

Since when z is a root, the adjoint matrix has unit rank (as can be seen by evaluating any of the 2×2 sub-determinants), we can carry out the multiplication using only one row (say the first) to yield:

(15, 16)
$$\left[2c_1c_{11} - 2c_2\left(2 - z - \frac{1}{z}\right)\left(1 - \frac{1}{z}\right)\right]G_z(S) + \left[0 \; \middle| \; -c_{20}\frac{(1 - (1/z))}{(1 - z)}\right]\left[\begin{matrix} S_0 \\ 1 \end{matrix}\right]$$

$$+ \left[2c_2\left(2 - z - \frac{1}{z}\right)\left(1 - \frac{1}{z}\right) \; \middle| \; 2c_2\left(1 - \frac{1}{z}\right) \middle| 0 \; \middle| \; \frac{-2c_2(1 - (1/z))}{z} \middle| 1 \; \middle| \frac{1}{z}\right]$$

$$\times \begin{bmatrix} H_{-1} \\ X_{-1} \\ H_0 \\ X_0 \\ \lambda_0 \end{bmatrix} = 0$$

where the vertical bars simply separate matrix elements.

However, since z is complex, (16) actually contains two independent equations, since both the real and the imaginary parts of the left-hand side of (16) must equal zero. (See Section 5 below.) Hence, combining (5) (for $t = 0$) with the two equations in (16) yields a single matrix equation with the exact number of rows to complete the system.

(17)
$$\begin{bmatrix} 0 & -2c_1c_{11} & 0 & 0 & 0 & 2c_1 & 0 & 0 & -1 & 1 \\ 0 & 0 & c_{21} & 0 & -2c_2 & 0 & 4c_2 & -2c_2 & 1 & 0 \\ 0 & -1 & 0 & 1 & 0 & -1 & 1 & 0 & 0 & 0 \\ 1 & 0 & \dfrac{-c_{20}b_R}{a_R} & 2c_2a_Rb_R^2 & 2c_2b_R & 0 & \dfrac{-2c_2b_R}{z_R} & 0 & \dfrac{1}{z_R} & 0 \\ 1 & 0 & \dfrac{-c_{20}b_I}{a_I} & 2c_2a_Ib_I^2 & 2c_2b_I & 0 & \dfrac{-2c_2b_I}{z_I} & 0 & \dfrac{1}{z_I} & 0 \end{bmatrix}$$

$$\times \begin{bmatrix} \displaystyle\sum_{t=0}^{T} W_{R,I}z_{R,I}S_t \\ S_0 \\ 1 \\ H_{-1} \\ X_{-1} \\ H_0 \\ X_0 \\ X_1 \\ \lambda_0 \\ \lambda_1 \end{bmatrix} = [0]$$

where

$$a_{R,I} = (1 - z_{R,I}), \quad b_{R,I} = \left(1 - \frac{1}{z_{R,I}}\right),$$

$$W_{R,I} = 2c_1 c_{11} - 2c_2\left(2 - z_{R,I} - \frac{1}{z_{R,I}}\right)\left(1 - \frac{1}{z_{R,I}}\right),$$

and the subscripts R, I denote whether the real or the imaginary part of the complex expression is to be used. The expression $\Sigma_{t=0}^{T} W_{R,I} z_{R,I} S_t$ uses the real parts of W and z for multiplication with the fourth row of the premultiplying matrix, and the imaginary parts for multiplication with the fifth row. Since the i's drop out, the matrices contain only real elements.

To obtain the final solution we isolate the last five elements of the column matrix and solve by matrix inversion:

$$(17') \quad \begin{bmatrix} H_0 \\ X_0 \\ \hline X_1 \\ \lambda_0 \\ \lambda_1 \end{bmatrix} = - \begin{bmatrix} 2c_1 & 0 & 0 & -1 & 1 \\ 0 & 4c_2 & -2c_2 & 1 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & \dfrac{-2c_2 a_R b_R}{z_R} & 0 & \dfrac{1}{z_R} & 0 \\ 0 & \dfrac{-2c_2 a_I b_I}{z_I} & 0 & \dfrac{1}{z_I} & 0 \end{bmatrix}^{-1}$$

$$\times \begin{bmatrix} 0 & -2c_1 c_{11} & 0 & 0 & 0 \\ 0 & 0 & c_{21} & 0 & -2c_{20} \\ 0 & -1 & 0 & 1 & 0 \\ 1 & 0 & \dfrac{-c_{20} b_R}{a_R} & 2c_2 a_R b_R^2 & 2c_2 b_R \\ 1 & 0 & \dfrac{-c_{20} b_I}{a_I} & 2c_2 a_I b_I^2 & 2c_2 b_I \end{bmatrix} \begin{bmatrix} \displaystyle\sum_{t=0}^{T} W_{R,I} z_{R,I} S_t \\ S_0 \\ 1 \\ H_{-1} \\ X_{-1} \end{bmatrix}$$

We can then eliminate the variables which we do not need, those below the dotted line; we are left with decision rules for H_0 and X_0 which are linear in the lagged values of H and X, a constant term, and weighted forecasts of current and future sales. Since the weights on future sales normally approach zero rapidly,[16] the decision rules can be truncated after a few periods, if necessary adjusting the weight on the last included term to make up for the omitted terms.

---

[16] Even though the derivation has utilized the assumption that the criterion function extends for a great many periods, T, the solution for the decision rules usually indicates that forecasts beyond a few periods are given so little weight as to be practically irrelevant. Thus the forecast analysis itself determines the relevant horizon.

5. ANCILLARY ANALYSIS AND EXTENSIONS

With the main development complete and illustrated, we turn to some points which have been set aside and others which are important qualifications or extensions of the main development. These include a variety of issues relating to the roots of the characteristic equation, problems posed by estimated rather than known relationships, changes through time in the criterion and the constraints, and time discounting.

Choosing the Appropriate Roots

In solving the characteristic equation (14) to eliminate the unknown transform variable, we obtained a total of $2R$ roots. From these we selected exactly R to supply the number of additional equations to complete the system and permit us to derive the optimal decision rules. We now demonstrate that the roots occur in reciprocal pairs and that only the R "small" roots, i.e., those whose absolute values are less than one, are relevant. Subsequently we show why this procedure will always yield exactly the right number of roots and the corresponding additional equations to complete the system.

Consider a decision problem accurately described by (1) and (2) in which T is even and very large and the second order conditions are satisfied. Imagine that the solution of the first order conditions (7) has been obtained for all time periods, and then that we lift part of the solution for the periods $(T/2, \ldots, T)$. This part is also an optimal solution since any part of an optimal solution is itself an optimal solution for the smaller problem. Now if one of the uncontrolled variables were changed by a unit pulse disturbance lasting one period at time $t = T/2$, what would be the effect on the part solution?[17] Since (7) is a linear system, we have the property of super position so that we may answer this question in terms of deviations from the original solution without considering it explicitly. Since (5) is a difference equation, it has a solution for the pulse disturbance that is of known form in the succeeding periods:

$$(19) \qquad g_{\alpha t} = \sum_i W_{\alpha i} z_i^{(t - T/2)} \qquad\qquad (\alpha = X, Y, \lambda; t = T/2 + 1, \ldots, T),$$

where the $g_{\alpha,t}$ are deviations from the original solution, the $W_{\alpha i}$ are constants, and the z_i are the roots of the characteristic equation.

If any of the roots satisfied the conditions $|z_i| \geqslant 1$ and $W_{\alpha i} \neq 0$ for some α, the deviations would grow with the passage of time or persist through all the remaining periods. Because (1) is a quadratic form, such deviations from the original solution would be penalized with the square of their sizes, and the penalties would cumulate over all the periods from $(T/2 + 1)$ to T. Performance as measured by the criterion function would certainly be improved if the deviations from the original solution occurred only in the periods near the disturbance and decreased as you moved away in time. Clearly the response that would occur if the above conditions were

---

[17] No restrictions on the above conclusions arise from considering a pulse disturbance since through super position any disturbance can be built up from pulses. Also selecting the period $T/2$ for the pulse or assuming T to be large and even can have no effect on the roots of the determinant.

satisfied would be non-optimal, and we can conclude that they would *not* be allowed by a decision analysis that guarantees an optimum. It follows that after the pulse occurred, the deviations would decay with the passage of time. This means that the roots satisfy the condition $|z_i| < 1$ and/or $W_{\alpha i} = 0$ for all i and α.

Now imagine that we lift the first part of the original solution for the periods $(0, \ldots, T/2)$, again apply a unit pulse for the period $T/2$, and consider the deviations from the original solution. In the foregoing discussion we have treated time as going forward from the time of the disturbance. But the analysis of Sections 2 and 3 is "static" in the sense that time plays no unique role because the future uncontrolled variables including the pulse disturbance are assumed to be known perfectly in advance of their occurrence. We obtain no new information with the passage of time from $t = 0$ to $t = T/2$; hence there is nothing unidirectional about time. We could as well have run time backwards from the time of the pulse and obtain the following difference equation solution:

$$(20) \qquad g_{\alpha t} = \sum_i W_{\alpha i}(z_i^{-1})^{(T/2 - t)} \qquad\qquad (\alpha = X, Y, \lambda; t = T/2 - 1, \ldots, 0).$$

Now the requirement $|z_i| < 1$ implies $|z_i^{-1}| > 1$ which would cause the deviations to grow with the backward passage of time, but again this cannot occur because of its non-optimality. Furthermore since time is in no way unidirectional under certainty, the deviations (19) and (20) will necessarily be symmetrical about $t = T/2$. We conclude therefore that the characteristic equation must contain pairs of roots, one root in each pair being the reciprocal of the other. One root is stable, i.e., converges for the forward passage of time, and the other is stable for the backward passage of time. Also necessarily the roots that have non-zero weights and are stable for the forward passage of time will have zero weights for the backward passage of time. The same is true for the roots that are stable with the backward passage of time. From symmetry the non-zero weights of the reciprocal roots will be equal.[18]

Since a pulse disturbance contains frequency components of all harmonics, the weights attached to each of the stable roots will be non-zero; i.e., half of the roots will be excited for the forward passage of time and the other half for the backward passage of time. Hence *all* roots will occur in reciprocal pairs.[19] Since the determinantal equation (14) is the characteristic equation of the system, the foregoing conclusions apply to the roots z_r.

---

[18] An earlier approach to some of the problems treated in this paper is found in [22] in which the solution for decision rules is treated as a classical problem in the solution of difference equations.

[19] Mechanically, the presence of a z in the determinant $B_3(z)$ arises from a term in the objective function in which a current controlled or partly controlled variable is multiplied by a lagged controlled or partly controlled variable, e.g., $kx_t y_{t-\gamma}$ ($t = 0, 1, 2, \ldots, T$) where k is a constant and x and y are any controlled or partly controlled variables. Differentiation with respect to x_t leaves $ky_{t-\gamma}$ which after passing through the generating function transform yields kz^γ as the element in the xth row, yth column of $B_3(z)$. Differentiation with respect to y_t leaves $kx_{t+\gamma}$ which yields $k(1/z)$ as the element in the yth row, xth column of $B_3(z)$. Upon expansion, a term $k(z^\gamma + (1/z))$ will appear in the polynomial so that if z is a root, so is $(1/z)$. The same analysis holds if the lag originates in a constraint equation since there will be a term $\lambda_t y_{t-\gamma}$ in the Lagrange function where λ is the Lagrange multiplier. Both cases appear in the example of Section 4 above.

The conclusions about roots can be summarized by suitably arranging the roots so that

(21) $|z_r| < 1$ $(r = 1, 2, \ldots, R)$

and

(22) $z_{R+r} = \dfrac{1}{z_r}$ $(r = 1, 2, \ldots, R).$

Thus we have established the presence of $2R$ roots, half of which are "small," i.e., satisfy (21). Note that by the above argument we have excluded $|z_r| = 1$ for any r, since this would cause infinitely distant future events to affect the decisions made at $T/2$.

For the forward passage of time the roots satisfy (21), and our neglect of the last term in the transform formula (11) is justified for large T.

When the decision rule (18) is used in real time under conditions of uncertainty, it will be subjected continually to disturbances so its dynamic stability is of considerable interest and importance. In this case time is very much unidirectional in the forward direction so we should use the roots z_r $(r = 1, 2, \ldots, R)$. Regardless of the stability or instability of the original dynamic system (2), the introduction of the optimal decision rules to govern the controlled variables ensures that the system will be inherently stable, i.e., any disturbances will die out with the passage of time. This neatly handles one of the servo engineer's headaches. Not only is the system inherently stable, but further it has the optimal degree of stability.

Number of Equations and Unknowns

When we set up the final matrix equation in the unknown time series variable (17) we expect to obtain the final decision rules (18) by elimination and matrix inversion. This can be done only if we have a complete system of equations, i.e., as many linear independent relations between the unknown variables as there are unknowns. To test whether or not this condition is satisfied we need to determine the number of unknowns and equations.

The transformation of an equation containing a leading variable, M_{t+p} for example, introduces by use of (11) the following additional time series variables: $M_0, M_1, \ldots, M_{p-1}$, making a total of $(p + 1)$ time series variables. Even if there is no lead $p = 0$, one variable, M_0, occurs. Since the past time series variables (ones with negative time subscripts) in (17) are known, we need count as unknown only those with time subscripts between zero and the maximum lead L.

The greatest lead for each variable L_α^+ $(\alpha = X, Y, \lambda)$ can be determined from (5) where L is the max of L_α^+. Thus the total number of unknown x_{it}, y_{jt}, and λ_{ft} in (17) is given by

(23) $\text{Unknowns} = \displaystyle\sum_{\alpha = \left\{\begin{smallmatrix} x_1, x_2, \ldots, x_{N_x} \\ y_1, y_2, \ldots, y_{N_y} \\ \lambda_1, \lambda_2, \ldots, \lambda_{N_f} \end{smallmatrix}\right.} (L_\alpha^+ + 1) = \sum_{a = X, Y, \lambda} (L_\alpha^+) + (N_x + N_y + N_f).$

The initial equation (8) contains $(N_x + N_y)$ rows from differentiation with respect to the controlled and uncontrolled variables in the period $t = 0$, with an additional N_f rows from the corresponding constraints.

This gives us a total of $(N_x + N_y + N_f)$ linear independent equations whose independence follows from satisfying the second order conditions.

As we found in Section 3, an equation (15) is found for each of the R small roots of the determinantal equation, so we need to find what determines R.

Now consider the test matrix which is the matrix obtained from (7) that multiplies the complete vector of unknowns. Imagine the columns of the matrix and the rows of the vector rearranged so that the variables corresponding to the greatest lead are grouped together in the vector

$$\begin{bmatrix} \vdots \\ \vdots \\ X_{(t+L_{\bar{x}})} \\ Y_{(t+L_{\bar{y}})} \\ \lambda_{(t+L_\lambda^+)} \\ \vdots \\ \vdots \end{bmatrix}$$

where T runs over the range $(-L, \ldots, 0, \ldots, L)$. By performing row operations on this matrix, we can reduce all elements below the diagonal to zeros so that it is triangular. We know that this matrix is non-singular from the satisfaction of the second order conditions so its determinant is non-zero. Since the determinant of the large triangulated matrix is equal to the product of its diagonal elements, we conclude that all of the coefficients on the diagonal are non-zero. Thus the greatest lead variable for $t = 0$ is known to occur with a non-zero coefficient in a particular row, another in the following row, and so on.

The block diagonal character of (7) will carry through to the diagonalized matrix so that repeating sub-matrices similar to (5) occur. We may take the transform of this set of equations to obtain an equation comparable to (12). The vector of unknown transform variables in this equation has a matrix which corresponds to $\underline{B}_3(z)$ and will necessarily have the same determinant since the triangularization process and rearrangement of columns does not affect the fundamental structure of the equations. The transform process as shown by (11) converts leads to powers of $(1/z)$ in the coefficients of the transform variables so greatest lead terms become greatest power terms.

The determinantal polynomial is the sum of all possible different product terms where each term is the product of matrix elements chosen so that each row and column is represented only once. Because each of the greatest lead terms for $t = 0$ occurred in different rows with non-zero coefficients, each of the largest powers of $(1/z)$ will occur in different rows with non-zero coefficients and (because they relate to different transform variables) in different columns.

We conclude that the largest power of $(1/z)$ in the determinant is the product of all the largest powers, one from each row and column, and hence is equal to the sum of the greatest leads.

Since the leads are matched by lags, corresponding powers of z occur to give an equal number of reciprocal roots. Since the degree of the polynomial is $2R$ and half of its roots are small (modulus less than one) we have

$$(24) \qquad R = \sum_{x=X,Y,\lambda} L_x^+ = \sum_{x=X,Y,\lambda} L_x^-$$

where L_x^- is the greatest lag in (5).

Thus we have a number of small roots so that the number of equations $(R + N_x + N_y + N_f)$ equals exactly the unknowns given by (23). Independence of the R equations is ensured by the second order conditions as follows. In (17) we have a set of linear equations that are known (if the terminal terms from (11) are negligible) to relate a certain set of time series variables. Equation (7) which involves the set of variables for *all* time periods and which is known from the second order conditions to have a unique solution could be solved to eliminate all variables beyond the first period, giving us a second complete set of linear independent equations in this same set of variables. Since the second set of equations has a unique solution, the first set must have the same unique solution, and because its solution is unique, its equations are independent.

Complex Roots

When the determinantal polynomial (14) is factored and the R small roots selected, some of the z_r are likely to be complex numbers in conjugate pairs; for example,

$$z_r = .5 + .2i,$$
$$z_{r+1} = .5 - .2i.$$

Substitution of a complex z_r in the solution for equation (15) will yield an equation with complex coefficients which incorporates two linear equations with real coefficients, one equation relating the real components and the other relating the imaginary ones. If we repeat this process for the complex z_{r+1}, we only obtain the same two equations over again. Hence when roots are complex, we should disregard one root of each conjugate pair and obtain two equations from the remaining root.

Multiple Roots

If two or more, say n, roots are equal, i.e., $z_1 = z_2 = \ldots = z_n$, then their substitution in (13) will not yield n independent equations (15) but only one. However, n independent equations can readily be obtained by taking derivatives of (13):

$$(25) \qquad \left| \frac{\partial^k}{\partial_z k} \left| B_3(z)^a \{ \ldots \} + (z - z_n)^n \prod_{r=n+1}^{R} (z - z_r)\left(z - \frac{1}{z_r}\right) IG \left[\begin{pmatrix} X \\ Y \\ \lambda_{tL} \end{pmatrix}_{tL} \right] \right| \right|_{z=z_n} = [0]$$

$$(k = 1, 2, \ldots, n - 1)$$

where we have expanded the determinant to show its roots. We can take $n - 1$ derivatives of the factor $(z - z_n)^n$ before it disappears. Since this factor goes to zero when we let $z = z_n$, it will knock out the unknown transform terms on the right in each of the derivative equations. Equation (13) plus the $(n - 1)$ equations (25) give us the needed equations corresponding to the n roots. Having established this fact, it is more convenient to perform the $(n - 1)$ derivatives on (12) and obtain the equations (15) by elimination.

Insufficient Roots

There is one further instance where the initial solution may appear to generate an insufficient number of roots.[20] If one of the first order or constraint equations is such that all the controlled and partly controlled variables appear with the same subscript, *and* each of the same variables appears with a lag elsewhere in the Lagrangean function $C(\lambda)$ so that future values of those variables appear in (5) when $t = 0$. the transform technique will not generate enough additional equations to complete the system. The trick is that the equation in question is valid for every time period $(t = 0, 1, 2, \ldots, T)$, and we may re-use it with all time subscripts increased by one or more periods so long as the new equation contains only time-subscripted variables which are already in (5), thereby adding enough additional equations (and no additional unknowns) to complete the system.

In light of the above, it will often prove useful to check the original set of relationships to determine whether any partly controlled variables can be conveniently eliminated from the relation (2) and the criterion (1). If so, we can avoid carrying those variables through the balance of the analysis. Because the elimination of a partly controlled variable reduces by one the number of equations in (2), we need one less Lagrange multiplier. Hence there is a double saving. Any simultaneous relationships (i.e., all variables in the same time period) are especially suitable candidates for elimination. Variables that occur with various leads and lags often cannot be conveniently eliminated because an infinite regress occurs.

It is even possible to eliminate controlled variables if fairly simple relations—especially simultaneous ones—exist between variables from past periods and controlled variables in the period $t = 0$. The decision rules for the eliminated control variables are found at the end by making substitutions into the relations that were used for their elimination.

Estimated Constraint Relations

Often relationships will be estimated statistically with residual random error variables. These enter the decision analysis as uncontrolled variables which are to be forecast. If, however, their conditional expected values are zero,[21] there is no point in putting them in the analysis explicitly. If they are included, they will, of course, appear in the decision rules but when zero forecasts are substituted their terms disappear anyway.

[20] This is sometimes called the zero root case.
[21] This excludes autocorrelated error terms whose conditional forecasts will in general not be zero.

Changes in the Criterion and Relationships Through Time

We assumed above that the criterion function and the relationships were the same in every period. Specifically we assumed that $K_{PQ;\delta}$ in (1) and $H_{fP\gamma}$ in (2) were constants. This restriction of constancy is now reviewed to determine to what extent it can be relaxed.

The large matrix equation (7) shows the structure of the first order conditions where the right matrix and vector may be partitioned into two terms to separate the knowns from the unknowns. The foregoing derivation in Section 3 is a method for obtaining a solution of this system for certain X's as linear functions of the V's and certain past X's and Y's. This involves in effect obtaining part of the inverse of the block diagonal matrix that multiplies the unknown vector and pre-multiplying the other matrix. For this solution to be performed and the resulting decision rules used in succeeding periods, it is necessary that all of the matrices be composed of constants (i.e., the coefficients of the matrices do not change over time) and further that the diagonal blocks be identical, because our derivation has used this property in obtaining the transform equations. However, since the V's appear explicitly in the decision rules, they should allow some flexibility. By tracing the origins of the matrices and vectors we can determine the necessary restrictions on the decision model, (1) and (2).

The required constancy of the matrices in (7) clearly indicates that the $H_{fX\gamma}$ in (2) must be constants as had been assumed. However, the uncontrolled V's contained in these relations or in the criterion function have the following kind of flexibility. The uncontrolled variable v_{kt} may be the product of a coefficient w_{kjt} that takes different values at different points of time and another uncontrolled variable v'_{jt}:

(26) $v_{kt} = w_{kjt} v'_{jt}.$

This would allow the uncontrolled v'_{jt} to have a time varying impact on the relations in (2). We can also accommodate the even more general situation in which:

(27) $v_{kt} = f_{V_t}(\ldots)_{Vt}$

where f_{V_t} indicates that the functional relationship changes with time and $(\ldots)_{Vt}$ indicates that the set of variables that are arguments of the function change with time. The functions must, of course, give unique values of v_{kt}. Thus we see that the impact of uncontrolled variables on the relations (2) may be very flexible indeed.

The controlled variables may not be determined directly by the decision maker but through some complex and time varying function

(28) $x_{it} = f_{Xt}(\ldots)_{Xt}$

provided that knowing an optimal decision x_{it} gives sufficient indication of the appropriate values of its arguments $(\ldots)_{Xt}$ that are causal variables by means of which the decision maker has control of x_{it}.

In the criterion function (1) the square and cross product terms of, and between, partly controlled and controlled variables contribute terms to the submatrix K

in the $\underline{A}_2$ matrix of (7). We conclude that their coefficients must be constants with the same values in every period.

The terms in the criterion function that depend *only* on the uncontrolled variables disappear on differentiation. Indeed arbitrary additive functions of the V's can be included in the criterion function since they drop out on differentiation.

The cross product terms between controlled and partly controlled variables on the one hand and uncontrolled variables on the other in (1) contribute linear terms in the V's upon differentiation by the X's and Y's in (4). Their coefficients appear in $\underline{A}_1$ matrix of (7) and must be constants with the same values in all time periods. These components of the criterion reflect the dependence of the decision maker's welfare on the interaction between the controlled and partly controlled variables on one hand and the uncontrolled environment on the other. However, they can be given another interpretation in which the V's reflect changes in the weight given in the criterion function to linear terms in the controlled and partly controlled variables. Thus some of the V's can be given the role of changing from period to period *some* of the parameters of the criterion function.

In summary we see that the uncontrolled variables can play the following three distinct functions in the decision model. They can

(i) impinge as uncontrolled disturbance variables (random or otherwise) on the dynamic system;

(ii) interact as uncontrolled environmental variables with the controlled and partly controlled variables in jointly determining the decision maker's welfare which is measured by his criterion function; or

(iii) enable the linear contributions of the controlled and partly controlled variables to the decision maker's welfare to be given different weights at different times.

When we consider the problem of forecasting these three types of uncontrolled variables, it is interesting that for the first two the decision maker must forecast his environment, but for the third he must forecast his own preferences.

We need to add one final qualification to the constancy of coefficients. The decision rule incorporates weighted sums of the future uncontrolled variables, but these weights have the form z_r^t where t indicates distance in the future. Since $|z_r| < 1$, these weights approach zero indicating the practical irrelevance of the future beyond a certain horizon in the future. For the analysis to be applicable to current decisions, we need only assume that the relevant coefficients will be constant to this forecast horizon and it will be determined by the weights in the decision rule.

Systems and criterion functions that are changing through time may, of course, be handled by matrix inversion to obtain decisions. However, there would be no decision rules that could be used repeatedly.

Time Discounting

In some decision problems the decision maker may desire to put decreased weight on future welfare in making current decisions. This can be done by modify-

ing the criterion function (1) as follows:

$$(29) \qquad C = \lim_{T \to \infty} \sum_{t=0}^{T} C_t \rho^t / T$$

where the real constant ρ is less than unity. Here we can let the criterion include an infinite number of time periods because the discount factor ensures convergence provided, of course, that the uncontrolled variables do not grow at an exponential rate greater than ρz_r. Letting T be infinity nicely disposes of the terminal transform terms which appeared in (11).

The analysis with the discount modification carries through much as before except that we obtain the transform equations by using the transform for power functions

$$(30) \qquad G(M_t \rho^t) = \sum_{t=0}^{\infty} M_t (\rho z)^t = G(M_t, \rho z).$$

In effect z is replaced by ρz. The roots will no longer occur in simple reciprocal pairs since the direction of time is now unique. Because $|\rho z| < |z|$, another effect of time discounting will be the shortening of the forecast horizon, i.e., the time span beyond which the weights are negligible and hence forecasts are not needed.

6. CONCLUSION

This paper has presented, illustrated, and proved the sufficiency of the z-transform approach to the general solution of the linear decision analysis in discrete time. Because the solution for optimal strategies under dynamic uncertainty is, in general, so complex that in many practical cases it is, or borders on, the uncomputable, the relative simplicity[22] of the solution for this special case is especially important. Even where the criterion is not quadratic and the constraints are not linear, the linear decision analysis often is useful in obtaining a first approximation to the optimal decision (or control) rules.

While matrix iteration or matrix inversion are likely to be computationally superior approaches to obtaining the numerical solution for linear decision rules,[23] the analysis presented in this paper shows the mathematical structure of the solution more clearly for the case considered here in which the cost and constraint structures are nearly identical in all time periods. Then the partial solution of a large set of simultaneous linear equations is required in which the matrix is characterized by recurring diagonal blocks. The large (nearly infinite) size of the equation set allows the z-transform to be used to obtain the only part of the solution that is needed, namely the decisions for the first period. Thus, the nearly infinite

---

[22] The reader who needs convincing that this analysis is *relatively* simple might try finding the optimal strategy under uncertainty for, say, a cubic criterion with quadratic constraints. Since certainty equivalence is a property only of the linear decision analysis, it no longer applies. As the solution proceeds backwards from the future, the function to be minimized involves a conditional joint probability distribution, and becomes successively more complex at each time period. See Richard Porter's thesis [23].

[23] See [18, 23] or Gregory C. Chow, "Optimal Control of Linear Econometric Systems with Finite Time Horizon," *International Economic Review*, 13 (1972), 16–25.

set of equations is thereby reduced to the solution of a small set of linear equations, but a numerical search process is required to find the characteristic roots of the system.

While simpler and more rigorous proofs undoubtedly can be devised for the solution of the simultaneous difference equations that constitute the first order conditions for the optimum, this paper should contribute to crystalizing the issues involved. This method of solution already has proven useful in a number of economic and operations research analyses, and also may prove useful for engineering control problems as well. Interestingly, the applications have been as much in contributing to rigorous specifications for empirical estimation as for normative prescriptions.

Yale University
    and
The Urban Institute

REFERENCES

[1] ATHANS, M., AND P. FALB: *Optimal Control*, New York: McGraw-Hill, 1966.
[2] BELLMAN, R. E.: *Dynamic Programming*. Princeton, N.J.: Princeton University Press, 1957.
[3] BRYSON, A. E., AND YU-CHI HO: *Applied Optimal Control*. Waltham, Mass.: Blaisdell, 1969.
[4] CHANG, J. C. D., AND C. C. HOLT: "Orders for Physical Capital Under Uncertainty and Dynamic Costs," paper presented at the North American Meetings of the Econometric Society, New York, December, 1969.
[5] CHILDS, G. L.: *Inventories and Unfilled Orders: A Structural Analysis*. Amsterdam: North-Holland Publishing Co., 1967.
[6] DURBIN, J.: "The Effect of Forecasting Errors in Dynamic Programming with a Quadratic Cost Function," mimeographed, London School of Economics, August, 1959.
[7] FRAZER, R. A., W. J. DUNCAN, AND A. R. COLLAR: *Elementary Matrices*. New York: Macmillan, 1938.
[8] FREEMAN, H.: *Discrete Time Systems*. New York: Wiley, 1965.
[9] HAY, G. A.: "Production, Price and Inventory Theory," *American Economic Review*, 60 (1970), 531–545.
[10] HILDEBRAND, F. B.: *Methods of Applied Mathematics*. Englewood Cliffs, N.J.: Prentice-Hall, 1952.
[11] HOLT, C. C.: "Linear Decision Rules for Economic Stabilization," read at the Twenty-First European Meeting of the Econometric Society, Amsterdam, September, 1959. Mimeographed, Carnegie Institute of Technology.
[12] ———: "Optimal Forecasting for Dynamic Programming with a Quadratic Criterion, I and II," mimeographed, London School of Economics and Carnegie Institute of Technology, February, 1959.
[13] HOLT, C. C., F. MODIGLIANI, J. F. MUTH, AND H. A. SIMON: *Planning Production, Inventories and Work Force*. Englewood Cliffs, N.J.: Prentice-Hall, 1960.
[14] HOLT, C. C., F. MODIGLIANI, AND J. F. MUTH: "Derivation of a Linear Decision Rule for Production and Employment," *Management Science*, 2 (1956), 159–177.
[15] HOLT, C. C., F. MODIGLIANI, AND H. A. SIMON: "A Linear Decision Rule for Production and Employment Scheduling," *Management Science*, 2 (1955), 1–30.
[16] HOLT, C. C., AND H. A. SIMON: "Optimal Decision Rules for Production and Inventory Control," in *Proceedings of the Conference on Production and Inventory Control*. Cleveland, Ohio: Case Institute of Technology, 1954.
[17] JURY, E. I.: *Theory and Application of the z-Transform Method*. New York: John Wiley, 1964.
[18] KALMAN, R. E., L. LAPIDUS, AND E. SHAPIRO: "On the Optimal Control of Chemical and Petroleum Processes," in *Symposium on Instrumentation and Computation in Process Development and Plant Design*. London: Institution of Chemical Engineers, 1959.

[19] LANING, J. H., AND R. H. BATTIN: *Random Processes in Automatic Control*. New York: McGraw-Hill, 1956.

[20] LEE, K. S., and C. C. HOLT: "Employment Decisions of Manufacturing Firms in Response to Demand Fluctuations: Theory and Empirical Results," Paper presented at the North American Meetings of the Econometric Society, New York, December, 1969.

[21] MANSFIELD, E., AND H. WEIN: "Linear Decision Rules and Freight Yard Operations," *Journal of Industrial Engineering*, 9 (1958), 93–98

[22] MODIGLIANI, F.: "A Method for Inverting the Matrix Arising from a Quadratic Decision Criterion," O.N.R. Research Memorandum No. 15, Graduate School of Industrial Administration, Carnegie Institute of Technology, 1954.

[23] PORTER, R.: "Strategies for Discrete-Time Decision Models," Ph.D. dissertation, University of Wisconsin, 1971.

[24] SAMUELSON, P. A.: *Foundations of Economic Analysis*. Cambridge, Mass.: Harvard University Press, 1953.

[25] SCHILD, A.: "On Inventory, Production and Employment Scheduling," *Management Science*, 5 (1959), 157–168.

[26] SIMON, H. A.: "Dynamic Programming Under Uncertainty with a Quadratic Criterion Function," *Econometrica*, 24 (1956), 74–81.

[27] THEIL, H.: "Econometric Models and Welfare Maximization," *Weltwirtschaftliches Archiv*, 72 (1954), 60–83.

[28] ———: *Economic Forecasts and Policy*. Amsterdam: North-Holland Publishing Company, 1958.

[29] ———: "A Note on Certainty Equivalence in Dynamic Planning," *Econometrica*, 25 (1957), 346–349.

[30] VAN DEN BOGAARD, P. J. M., AND A. P. BARTEN: "Optimal Macroeconomic Decision Rules for the Netherlands, 1957–1959," read at the Twenty-First European Meeting of the Econometric Society, Amsterdam, September, 1959.

[31] VAN DEN BOGAARD, P. J. M., AND H. THEIL: "Macrodynamic Policy Making: An Application of Strategy and Certainty Equivalence Concepts to the Economy of the United States, 1933–1936," *Metroeconomica*, 11 (1959), 149–167.

[32] WEINER, N.: *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. New York: Wiley, 1949.