General Disclaimer

One or more of the Following Statements may affect this Document

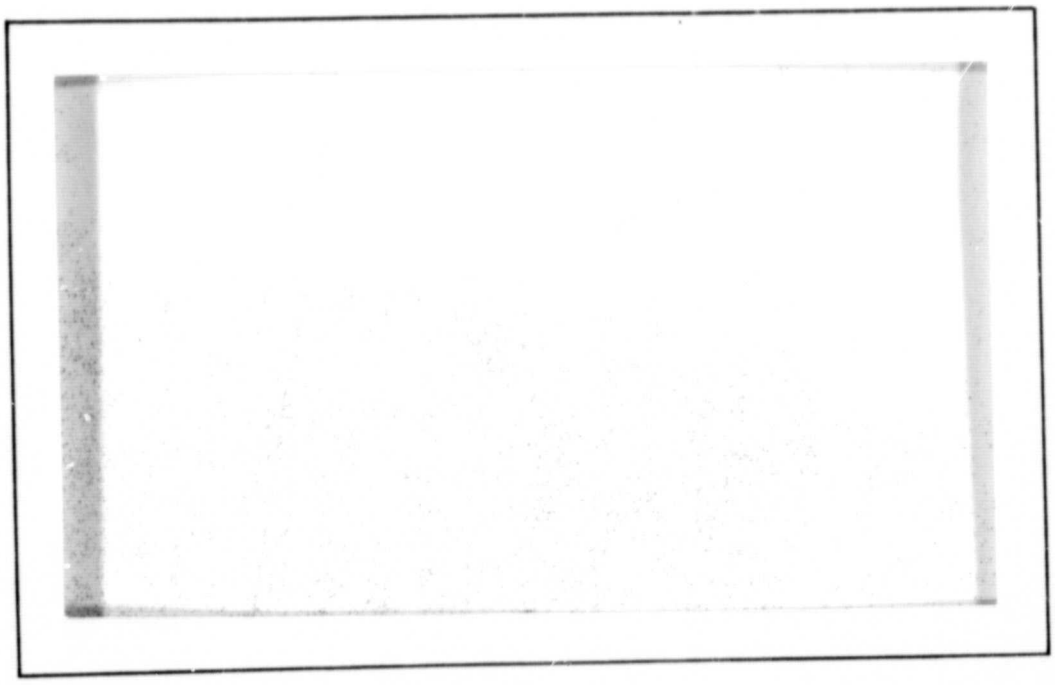
- This document has been reproduced from the best copy furnished by the organizational source. It is being released in the interest of making available as much information as possible.

- This document may contain data, which exceeds the sheet parameters. It was furnished in this condition by the organizational source and is the best copy available.

- This document may contain tone-on-tone or color graphs, charts and/or pictures, which have been reproduced in black and white.

- This document is paginated as submitted by the original source.

- Portions of this document are not fully legible due to the historical nature of some of the material. However, it is the best reproduction available from the original submission.



CENTER FOR
SYSTEM SCIENCE

UNIVERSITY OF ROCHESTER
ROCHESTER, NEW YORK

THE INVERSE PROBLEM
OF THE OPTIMAL REGULATOR

Ryuzo Yokoyama

December, 1969

A Thesis Submitted
in Partial Fulfillment
of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY

Supervised by Professor Edwin Kinnen
Department of Electrical Engineering
The University of Rochester
Rochester, New York

VITA

Ryuzo Yokoyama was born in [redacted], on [redacted]. During the period of 1960-1966, he received undergraduate and graduate training at Tohoku University in Japan and was awarded the Bachelor of Engineering degree and the Master of Engineering degree with a major in Electrical Engineering in 1964 and 1966 respectively.

He entered the University of Rochester in the fall of 1966 to begin a Ph.D. program in Electrical Engineering. He has been assisted in this program by the award of a graduate teaching assistantship in the Department of Electrical Engineering during the period of September 1966 to May 1967, and the award of graduate research assistantships thereafter from the National Aeronautics and Space Administration and the Office of Naval Research, Center for Naval Analyses.

ACKNOWLEDGEMENT

The author expresses his appreciation to Dr. Edwin Kinnen for his guidance and encouragement during the course of this research and to Dr. J.H.B. Kemperman for his helpful advice and assistance.

This work was performed partially with the financial support of the National Aeronautics and Space Administration under grant ~~NsG-574~~ NGR-33-019-014 to the University of Rochester. It was also supported in part by The Center for Naval Analyses of the University of Rochester. This support does not imply endorsement of the content by the Navy.

Finally, I should like to thank Miss Gail Herbst for typing a very difficult thesis.

ABSTRACT

A new phase canonical form is given for a class of multi-input dynamical systems described by time invariant ordinary differential equations. This is based on a modified definition of an equivalent relation for the class of systems. It is shown that a characteristic quantity called a stage distribution defined with respect to the linear part of the system uniquely determines the structure of its canonical form.

The inverse problem of the optimal regulator is considered for this class of systems with integral type performance indices. A convenient analysis of this problem is possible, using the developed phase canonical form. A theorem is stated which asserts necessary and sufficient conditions for optimized performance indices for a specified feedback control law. Further results concern the nonnegativity of loss functions as optimized performance indices under the additional assumptions that the nonlinearities of the system are given as polynomial functions of the state variables and that the feedback control law results in a linear autonomous system. A theorem of necessary conditions for this is given. Sufficient conditions are stated for linear systems. Based on these main theorems, supplimentary theorems and

corollaries are given which reveal other fundamental aspects of optimal feedback control systems.

In comparison with similar studies by other investigators, this work is directed toward more general assumptions on the inverse problem, i.e., generalizations of the system description, the specified feedback control law, and the performance indices. As a consequence, results of other investigators can be described as special cases of those resulting from this work.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Introduction to the Thesis

When a task in the physical world is approached, there naturally occurs the question of the best method to accomplish it. Problems of optimal control are those which attempt to find the best methods through mathematical descriptions of the task. These mathematical descriptions are composed essentially of

(i) a model of the cause-effect dynamics of the task, i.e., the system equation,

(ii) beginning and end parts of the task, i.e., initial and final conditions for the system equations,

(iii) permitted methods to accomplish the task, i.e., admissible control functions, and

(iv) a standard to measure the optimality for each admissible method, i.e., a performance index.

This search for best methods is identified with calculations of a optimal control function to transfer the conditions of the system as desired with the minimum possible value of the performance index.

An elementary problem of optimal control can be formulated from a single initial condition, a single final

condition and some control function of time to be calculated. This is generally called an open loop optimal control problem. Practically these problems are more often recognized under somewhat different circumstances, those of an optimal regulator. Thus the system function is to maintain a specified condition even though it is exposed occasionally to unforeseen disturbances. The recovery to the specified condition after each disturbance is to be in some optimal manner.

Consider, for example, a room in a building with an air temperature of 10°C. It may be desired to change the temperature to a steady 20°C as fast as possible, using a particular heating system. The open loop optimum control problem would be to design the given heating system behavior to minimize the time required for this change, recognizing the characteristics of the room and the heating system. Alternately, the regulation of the room temperature at 20°C could be considered under disturbances due to opening and shutting of doors and to outside weather conditions. To return the temperature to 20°C in some optimal sense, say minimum time, after a change due to these unpredictable causes, is a problem of an optimal regulator.

An optimal regulator problem can be expressed as a family of many individual open loop optimal control problems

with common system equations and assumptions, but with different initial conditions. If optimal control functions are required for many initial conditions, the necessity to calculate them one by one is unreasonable. An optimal control function can be developed alternately in closed form as a function of the system condition. Generally called an optimal feedback control law, this function establishes the optimal control from the present system condition regardless of its preceding career. Consequently, control action for any initial condition can be covered by a single control law.

Practical problems of optimal control may have many controlling and controlled quantities represented by complex algebraic descriptions, e.g., rocket control during space flight, utility power plant and distribution network regulation, and industrial chemical process control. Consider the regulation of a rocket flight path to a fixed trajectory in space. The control variables, representing thrust from individual rockets or combinations of rockets, could be represented in terms of three orthogonal directions of space. The output quantities are the position and velocity of the rocket, each composed of three orthogonal component variables. Consequently, the system has six controlled variables associated with the three controlling variables. The performance index might be a minimum integral

of deviations of the rocket from the fixed trajectory, or alternately, minimum fuel consumption during a specific control action.

The problem of optimal control has been intensively studied as one of the main branches of modern control theory, not only because of the interesting mathematics of the problem, but because of the practical character of the solutions. Many techniques of optimization have been developed in this field, based mostly on the calculus of variations.[1-6] Unsolved problems, however, still exist. Analytically these result from difficulties in finding sufficient conditions for optimality in a general sense and from the rapid increase of required calculations as the size of the system equations becomes realistically large. Complete answers to optimum control problems are restricted at the present time to a few specific classes of problems with relatively simple system equations. Furthermore, the optimal control function solutions to some of these problems may be impossible or inconvenient to reduce to hardware.

This thesis investigates relevant characteristics of optimal feedback control laws for a class of optimal regulators with system equations given by multi-input, time invariant ordinary differential equations. But the direction of the attack is just opposite or inverse to

the usual methods of investigation. The question is asked, what performance indices can be optimized by an assumed feedback control law? The objective is to seek all performance indices shared by a control law. This problem formulation is generally called the inverse problem of the optimal regulator,[6] or briefly the inverse problem.

A study of the inverse problem would (1) disclose practical advantages of using specific classes of control laws in combination with specific performance indices, and (2) distinguish between control laws which are optimal in some sense and those which are not. Consequently, the results would allow future optimal design problems to start with realistic performance indices, and be helpful for understanding observed solutions to control problems that are assumed optimal in some sense.

Before an analysis of the inverse problem is given in the following chapters, however, a canonical form is developed for the class of systems of interest. A canonical form is a compact standard form for describing all systems that are mathematically similar. It is useful both to clearly expose the mathematical composition of the system structure and to allow the analysis and design of the system to conveniently proceed using a compact description. Necessarily the choice of a canonical form for a given

class of problems plays an important role in the succeeding analysis. While various canonical forms have been suggested for linear systems, the new one introduced in this study is apparently necessary for the analysis of the inverse problem considered. It is also more generally useful for demonstrating the mathematical structure of systems. Thus this thesis considers two topics, (1) the inverse problem of the optimal regulator and (2) a canonical form for a broad class of multi-input deterministic systems described by time invariant ordinary differential equations.

1.2 Outline of the Thesis

The material presented in this thesis is divided into eight chapters. Chapter 1 contains an introduction to the topics considered, an outline, and a summary of the results. Chapter 2 is an introduction to the mathematical formalism used in subsequent theoretical developments. Notations, definitions and theorems that are assumed in later chapters are given here.

Chapter 3 summarizes and reviews work that has been published in the area of canonical forms for linear time invariant systems. These canonical forms are grouped in two categories according to descriptive structures, the Jordan standard matrix canonical form and the phase variable

canonical form. The new phase canonical form developed for a class of multi-input systems is given in Chapter 4. First, the development is concerned with linear systems, defining two characteristic quantities, a stage number and a stage distribution, which determine the structure of the systems uniquely, as illustrated by three examples. Second, the results are expanded for more general non-linear systems.

Chapter 5 summarizes and reviews publications on the inverse problem of optimal regulators. Interest is focused on work for problem assumptions similar to those made during this investigation. In Chapter 6, after a precise statement of the inverse problem of the thesis, a general analysis is given. A number of interesting characteristics of optimal feedback control systems revealed by this analysis are then given.

The results of Chapter 6 are applied to linearly synthesized optimal feedback control systems in Chapter 7. Conditions for a nonnegative loss function in optimized performance indices are given, and a principle of necessity of control action is disclosed. Conclusion and suggestions for future studies are in Chapter 8.

1.3 Summary of the Thesis

There are three principal results in this thesis, applicable to a class of deterministic, dynamic systems described by time invariant ordinary differential equations either, linear or nonlinear.

(i) A new phase canonical form is developed for this class of systems. In comparison with other canonical forms, this canonical form has the advantages that (1) it is applicable to a larger class of nonlinear, uncontrollable systems, (2) its structure is uniquely determined for each system by a defined quantity, a stage distribution, and (3) it would appear to be more conveniently used for analyses of optimal control problems than other known forms.

(ii) Necessary and sufficient conditions are given for feedback control laws which are optimal for performance indices given as integral forms and loss functions as sums of penalty functions of the state variables plus positive definite quadratic forms of the control variables. From this result, additional characteristics of optimal feedback control systems are revealed.

(iii) Necessary conditions for nonnegative loss functions in optimized performance indices are given for the inverse problem resulting in linearly synthesized optimal feedback

control systems, assuming the nonlinearities of the systems and the penalty functions are given as polynomials of the state variables. These conditions are also sufficient if the system has no nonlinearity. Specifically it is shown that, for controllable linear systems with linear feedback control laws, nonnegative loss functions in optimized performance indices must be quadratic forms of the state variables and control variables. Additional relevant aspects of linearly synthesized optimal control systems are also given.

Chapter 2

FOUNDATIONS

Material basic to the theoretical developments throughout this thesis is given in this chapter. Following the introduction of mathematical notations for the abstract space descriptions of multi-variable functions, sections are given defining and explaining system modeling, solutions, stability, and controllability. A definition of approachability is introduced.

2.1 Fundamental Mathematical Concepts

2.1.1 Notation for Vectors and Matrices

Vectors and matrices are denoted by underlined capital Roman letters or Greek letters. Their dimensions are apparent either from definitions or are stated explicitly. A null matrix and a null vector are denoted by [0] and $\underline{0}$ respectively. The r x r unit matrix is given by I_r and the inverse matrix by a -1 superscript, e.g., $\underline{A}^{-1}$. The transpose of a matrix or a vector has a T superscript, e.g., $\underline{A}^T$ or $\underline{X}^T$. The scalar product of two vectors $\underline{X}$ and $\underline{Y}$ is

$$\underline{X}^T\underline{Y} \triangleq \Sigma x_i y_i \ . \qquad (2\text{-}1)$$

The quadratic form of $\underline{X}$ associated with a matrix $\underline{A}$ is written as

$$\underline{X}^T \underline{A} \underline{X} \triangleq \sum_{i,j} a_{ij} x_i x_j \ . \qquad (2\text{-}2)$$

The Euclidian norm of a vector $\underline{X}$, denoted by $||\underline{X}||$, is

$$||\underline{X}|| = \sqrt{\underline{X}^T \underline{X}} \ . \qquad (2\text{-}3)$$

All subsequent discussions are assumed to be in finite dimensional Euclidian spaces or the product spaces. R^n designates an n-dimensional Euclidian space.

2.1.2 Multi-Variable Scalar Functions

With vector notation, functions of many variables are conveniently described as

$$f(\underline{X}) = f(x_1, x_2, \ldots, x_n) \ ,$$

$$f(\underline{X}, \underline{U}) = f(x_1, x_2, \ldots, x_n, u_1, u_2, \ldots, u_m) \ ,$$

or $$f(\underline{X}, \underline{U}, t) = f(x_1, x_2, \ldots, x_n, u_1, u_2, \ldots, u_m, t) \qquad (2\text{-}4)$$

defined on R^n, $R^n \times R^m$ and $R^n \times R^m \times R^1$ spaces respectively.

Theorem 2-1:

If $f(\underline{X})$ is continuous on R^n, then $f(\underline{N}\ \underline{X})$ is also continuous on R^n, where $\underline{N}$ is an n x n constant nonsingular matrix.

The proof follows directly from a fundamental theorem of the composition of continuous mappings.[7]

Definition 2-1: Functions of Class C_n.[7]

A scalar function $f(\underline{X})$ defined on R^n is said to be of class C_r in a region $\Gamma \subset R^n$ if it has continuous partial derivatives with respect to all x_i, $(i = 1,2,\ldots n)$, up to order r everywhere in Γ. When Γ is the entire R^n, the phase "in Γ" is omitted.

Provided that $f(\underline{X})$ is of class C_2, the following notations are often used

$$\frac{\partial f}{\partial \underline{X}} = \text{grad } f(\underline{X}) \triangleq \begin{bmatrix} \frac{\partial f(\underline{X})}{\partial x_1} \\ \frac{\partial f(\underline{X})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\underline{X})}{\partial x_n} \end{bmatrix}, \qquad (2\text{-}5)$$

$$\frac{\partial^2 f}{\partial \underline{X} \partial \underline{X}} \triangleq \frac{\partial}{\partial \underline{X}}\{\text{grad } f(\underline{X})\} \triangleq \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \\ \vdots & \vdots & & \\ \frac{\partial f}{\partial x_1 \partial x_n} & \cdots & & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{bmatrix} \tag{2-6}$$

Theorem 2-2:

Let

$$\underline{W}(\underline{X}) = \begin{bmatrix} w_1(\underline{X}) \\ w_2(\underline{X}) \\ \vdots \\ w_n(\underline{X}) \end{bmatrix} \tag{2-7}$$

be a vector valued function of class C_1. Then $\frac{\partial}{\partial \underline{X}}\{\underline{W}(\underline{X})\}$ is symmetric if and only if $\underline{W}(\underline{X})$ is a gradient of a scalar function $\nabla(\underline{X})$, i.e.,

$$\text{grad } \nabla(\underline{X}) = \underline{W}(\underline{X}) \ . \tag{2-8}$$

The proof of this theorem follows directly from the theory of functions of several variables.[7, Chapter 3] Furthermore, $V(\underline{X})$ can be uniquely determined by the line integral

$$V(\underline{X}) = \int_0^{\underline{X}} \underline{W}^T(\underline{X})\, d\underline{X} + C \qquad (2\text{-}9)$$

and is independent of the path of integration. A convenient line integral is

$$V(\underline{X}) = \int_0^{x_1} w_1(\gamma_1, 0\ldots)\, d\gamma_1 + \int_0^{x_2} w_2(x_1, \gamma_2, 0\ldots 0)\, d\gamma_2 + \cdots$$
$$\cdots + \int_0^{x_n} w_n(x_1, x_2, \ldots x_{n4}, \gamma_n)\, d\gamma_n + C\ , \qquad (2\text{-}10)$$

where C is a constant.[8]

Definition 2-2: Sign Definite Functions.[9]

Consider a scalar function $V(\underline{X})$ defined on R^n and an open region $\Gamma \subset R^n$ such that $\underline{0} \in \Gamma$. Then $V(\underline{X})$ is said to be positive (negative) semidefinite in Γ if

$$\text{(i)} \quad V(\underline{0}) = 0 \qquad (2\text{-}11)$$

$$\text{(ii)} \quad V(\underline{X}) \geq 0\ ,\ (V(\underline{X}) \leq 0)\ ,\ \text{for all } \underline{X} \in \Gamma\ . \qquad (2\text{-}12)$$

If $V(\underline{X})$ satisfies

$$\text{(ii)}' \quad V(\underline{X}) > 0 \ , \ (V(\underline{X}) < 0) \ , \text{ for all } \underline{X} \in \Gamma \text{ of } \underline{X} \neq \underline{0} \tag{2-13}$$

instead of (ii), it is said to be positive (negative) definite in Γ. When Γ is the entire R^n, the phase "in Γ" is omitted.

Theorem 2-3:

Assume a finite degree polynomial scalar function $V(\underline{X})$ given by

$$V(\underline{X}) = V^{(2)}(\underline{X}) + V^{(3)}(\underline{X}) + \ldots + V^{(\xi)}(\underline{X}) \ , \tag{2-14}$$

where each $V^{(i)}(\underline{X})$, $(i = 2,3,\ldots,\xi)$, is an i^{th} degree homogeneous function of $\underline{X}$ and $V^{(\xi)}(\underline{X})$, $\xi \geq 2$, is not identically zero. Then, for $V(\underline{X})$ to be positive semi-definite, it is necessary that $V^{(\xi)}(\underline{X})$ be positive semi-definite.

Proof: Consider the contrary, that $V(\underline{X})$ is positive semi-definite but there exists a certain vector $\underline{\chi} \in R^n$ such that

$$V^{(\xi)}(\underline{\chi}) < 0 \ . \tag{2-15}$$

Consider a set of vectors given $y\underline{\chi}$, where y is a scalar

variable. Then

$$f(y) \triangleq V(y\underline{X}) = y^2 V^{(2)}(\underline{X}) + y^3 V^{(3)}(\underline{X}) + \ldots + y^{\xi} V^{(\xi)}(\underline{X}) , \tag{2-16}$$

which is a polynomial function of y. Since $V^{(\xi)}(\underline{X})$ is negative, $f(y)$ becomes negative as $y \to \pm\infty$ and $V(\underline{X})$ becomes negative for $\underline{X} = y\underline{X}$. This contradicts the hypothesis. It also follows that ξ must be even for positive definiteness of $V(\underline{X})$.

2.1.3 Sign Definite Matrices

Assume $\underline{Q}$ to be an n x n real matrix.

Definition 2-3: Sign Definite Matrix.[10]

The matrix $\underline{Q}$ is said to be positive (positive semi) definite if the scalar function $\underline{X}^T\underline{Q}\underline{X}$ is positive (positive semi) definite. If $-\underline{X}^T\underline{Q}\underline{X}$ is positive (positive semi) definite, $\underline{Q}$ is said to be negative (negative semi) definite.

Without loss of generality, $\underline{Q}$ is also assumed to be a symmetric matrix in this section.[10]

Theorem 2-4:[10]

(i) Every $\underline{Q}$ is congruent to

$$\begin{bmatrix} I_{n_1} & & \\ & -I_{n_2} & \\ & & [0] \end{bmatrix}, \text{ all other entries zero,} \qquad (2\text{-}17)$$

where n_1 and n_2 are uniquely determined by $\underline{Q}$.

(ii) $\underline{Q}$ is positive definite if and only if $n_1 = n$, i.e., (2-17) degenerates to the unit matrix, and $\underline{Q}$ is positive semidefinite if and only if $n_2 = 0$.

(iii) An n x n real symmetric matrix $\underline{Q}_1$ is congruent to $\underline{Q}$ if and only if $\underline{Q}_1$ is congruent to (2-17), i.e., sign definiteness of $\underline{Q}$ is invariant for a congruent transformation.

Corollary 2-1:

(i) $\underline{Q}$ is positive definite if and only if all principal minor determinants of $\underline{Q}$ are positive. [10]

(ii) $\underline{Q}$ is positive semidefinite if and only if there exists an n_1 x n matrix $\underline{D}$ of rank $n_1 \leq n$ such that [11]

$$\underline{Q} = \underline{D}^T \underline{D} . \qquad (2\text{-}18)$$

(iii) $\underline{Q}$ is positive definite if and only if $n = n_1$ above.

(iv) $\underline{Q}$ is positive definite if and only if $\underline{Q}^{-1}$ is positive definite.

Proof: The proof of (iii) follows from (ii) of this Corollary and (ii) of the Theorem. For (iv), for $\underline{Q}$ to be positive definite, it must be nonsingular by (i) and $\underline{Q}^{-1}$ exists. Consider a congruent transformation of $\underline{Q}$ by $\underline{Q}^{-1}$. Thus

$$(\underline{Q}^{-1})^T \underline{Q} \underline{Q}^{-1} = (\underline{Q}^{-1})^T = (\underline{Q}^T)^{-1} = \underline{Q}^{-1} , \qquad (2\text{-}19)$$

which follows from the transpose of an inverse matrix[10] and from the symmetry of $\underline{Q}$. From (iii) of Theorem 2-4, $\underline{Q}$ is positive definite if and only if $\underline{Q}^{-1}$ is positive definite.

Theorem 2-5:[13]

Let $\underline{Q}$ be an arbitrary n x n positive definite symmetric matrix. Then there exist two positive numbers, $\lambda_{min}(\underline{Q})$ and $\lambda_{max}(\underline{Q})$, such that

$$\lambda_{min}(\underline{Q}) ||\underline{X}||^2 \leq \underline{X}^T \underline{Q} \underline{X} \leq \lambda_{max}(\underline{Q}) ||\underline{X}||^2 \quad \text{for all } \underline{X} \in R^n . \qquad (2\text{-}20)$$

2.2 Description of Physical Systems[5,12]

Systems which behave according to the Principle of Causality in the physical world can be schematically described as in Figure 1-1. Causes are classified into

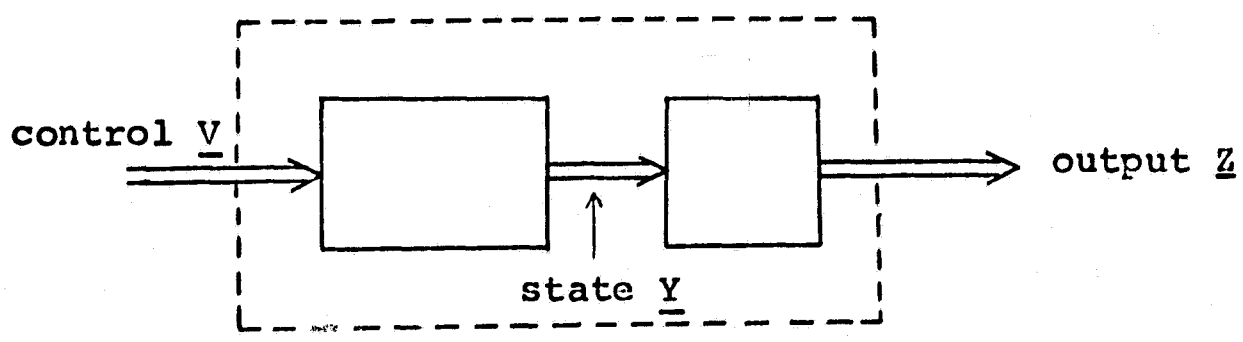


Figure 2-1. A physical system.

controls and disturbances; the former quantities can be specified and manipulated at will but the later cannot. Effects are created by the causes on the physical system. These are outputs or directly observed quantities of effects. To some extent, outputs depend upon the preferences of the observer. To avoid any ambiguity of outputs, a third quantity, states within the system, are conveniently introduced. Always regarded as abstract quantities, the states are defined as the minimal amount of information about the past history of the system which is sufficient to predict the affects upon the future.

Systems considered in this thesis are assumed to belong to a class called deterministic, real, finite dimensional, continuous time, ordinary differential, dynamical systems described by the equations

$$\dot{\underline{Y}} = \hat{\underline{F}}(\underline{Y}, \underline{V}, t) \tag{2-21}$$

$$\underline{Z} = \underline{G}(\underline{Y}, t) \ . \tag{2-22}$$

$\underline{Y}$, $\underline{V}$, and $\underline{Z}$ are called the state, control and output vectors,

$$\underline{Y} = [y_1, y_2, \ldots, y_n]^T \tag{2-23}$$

$$\underline{V} = [v_1, v_2, \ldots, v_m]^T \tag{2-24}$$

$$\underline{Z} = [z_1, z_2, \ldots, z_{n_o}]^T \ . \tag{2-25}$$

By each adjective, the following is meant:

(i) Deterministic: The process described by the system is deterministic.

(ii) Real and finite dimensional: The state and control vectors are defined in real finite dimensional spaces.

(iii) Continuous time: The set of time for these equations is an open interval $(T_o, T_1) \subset R^1$.

(iv) Ordinary differential: The behavior of the state of the system is given by ordinary differential equations as (2-21), where $\cdot = \frac{d}{dt}$.

(v) While a more detailed definition of a dynamical system can be given,[5,12] the following is adequate for this work. Dynamical system: From any $\underline{Y}_o$ and t_o and for any piecewise continuous m dimensional vector valued function $\underline{V}(t)$, each existing in defined regions, there is a unique solution $\hat{\underline{\phi}}_{\underline{V}}(t; \underline{Y}_o, t_o)$ to (2-21), i.e., an n-dimensional, vector valued function, differentiable in t, satisfying

$$(a) \quad \hat{\underline{\phi}}_{\underline{V}}(t_o; \underline{Y}_o, t_o) = \underline{Y}_o \ , \tag{2-26}$$

$$(b) \quad \frac{d}{dt} \hat{\underline{\phi}}_{\underline{V}}(t; \underline{Y}_o, t_o) = \hat{\underline{F}}(\hat{\underline{\phi}}_{\underline{V}}(t; \underline{Y}_o, t_o), \underline{V}(t), t)$$

$$\text{for all } t \in R^1 \ , \tag{2-27}$$

$$(c) \quad \hat{\underline{\phi}}_{\underline{V}}(t; \underline{Y}_o, t_o) = \hat{\underline{\phi}}_{\underline{V}}(t; \hat{\underline{\phi}}_{\underline{V}}(t_1; \underline{Y}_o, t_o), t_1)$$

$$\text{for all } t \geq t_1 \geq t_o \ . \tag{2-28}$$

Generally (2-21), (2-22) and a solution $\hat{\underline{\phi}}_{\underline{V}}(t; \underline{Y}_o, t_o)$ are collectively called the system equation, the output equation, and the solution for $(\underline{Y}_o, t_o)$ and $\underline{V}(t)$. The behavior of the state variables can be identified directly

from the solution, while the output variables are calculated algebraically from (2-22). Consequently substantial analyses of the systems can proceed through system equations only, with the output equations reduced to a secondary role. Thus a dynamical system is usually represented by just the system equations, (2-21), with the output equations, (2-22), implied.

The existence of a unique solution to system equations depends both on the given system equations and a specified control function. A sufficient condition can be stated as follows.

Theorem 2-6:[5]

For the system (2-21), if

(i) $\underline{\hat{F}}(\quad)$ is a continuous function, from $R^n \times R^m \times R^1$ into R^n,

(ii) the partial derivatives $\frac{\partial \hat{f}_i(\quad)}{\partial y_j}$ for $i,j = 1,2,\ldots n$, are continuous functions from $R^n \times R^m \times R^1$ into R^1,

(iii) $\underline{V}(t)$ is a piecewise continuous function from R^1 to R^m, and

(iv) $\underline{Y}_o \in R^n$ and $t_o \in R^1$ are specified,

then there exists a unique solution $\underline{\hat{\Phi}}_{\underline{V}}(t; \underline{Y}_o, t_o)$ on a time interval containing t_o.

Imbedding the specified $\underline{V}(t)$ as $\underline{\hat{F}}(\underline{Y}, t) = \underline{\hat{F}}(\underline{Y}, \underline{V}(t), t)$, another theorem can be given.

Theorem 2-7:[2,13]

Let D be a polyhedron in $R^n \times R^1$ where $(\underline{Y}_o, t_o)$ exists as an interior point. For a system

$$\dot{\underline{Y}} = \underline{\hat{F}}(\underline{Y}, t) \ , \qquad (2\text{-}29)$$

if $\underline{\hat{F}}(\underline{Y}, t)$ is continuous in D and there exists a number $k > 0$ which satisfies

$$||\underline{\hat{F}}(\underline{Y}_1, t) - \underline{\hat{F}}(\underline{Y}_2, t)|| \leq k||\underline{Y}_1 - \underline{Y}_2||$$

$$\text{for all } (\underline{Y}_1, t), (\underline{Y}_2, t) \in D \ , \qquad (2\text{-}30)$$

then there exists a unique solution $\underline{\hat{\Phi}}_{\underline{V}}(t; \underline{Y}_o, t_o)$ in a domain $D^1 \subset D$.

Generally (2-30) is called a Lipschitz condition.

Depending upon characteristics, systems are described by special names.

Definition 2-4: Time Invariant, Free, Autonomous System.[14]

A system is said to be

(i) time invariant (or stationary), if (2-21) is given as

$$\dot{\underline{Y}} = \hat{\underline{F}}(\underline{Y}, \underline{V}) \ , \qquad (2\text{-}31)$$

(ii) free, if (2-21) is given as

$$\dot{\underline{Y}} = \hat{\underline{F}}(\underline{Y}, t) \ , \qquad (2\text{-}32)$$

and

(iii) autonomous, if (2-21) is given as

$$\dot{\underline{Y}} = \hat{\underline{F}}(\underline{Y}) \ , \qquad (2\text{-}33)$$

i.e., time invariant and free.

Definition 2-5: Linear System.[12]

A system is said to be linear if (2-21) is given as

$$\dot{\underline{Y}} = \hat{\underline{A}}(t)\ \underline{Y} + \hat{\underline{B}}(t)\ \underline{V} \qquad (2\text{-}34)$$

where $\hat{\underline{A}}(t)$ and $\hat{\underline{B}}(t)$ are n x n and n x m matrix valued functions.

For a free system, the concept of an equilibrium point is particularly convenient.

Definition 2-6: Equilibrium Point of a System.[9]

A point $\underline{Y}_e$ in R^n is said to be an equilibrium point of a free system (2-32) if

$$\underline{0} = \underline{\hat{F}}(\underline{Y}_e, t) \qquad \text{for all } t \in R^1 . \qquad (2\text{-}35)$$

Practical examples of physical systems described by this symbolism and terminology appear in the literature.[3,5,8]

2.3 Stability and Controllability

For any given system, two descriptive characteristics can be considered, system stability and system controllability. These are useful in the analyses of the system behavior and the syntheses of control functions.

2.3.1 System Stability,[9,14] Stability in the Liapunov Sense

The theory of the behavior of solutions in relation to an equilibrium point of a free system is known as stability theory. It is based largely on concepts originally proposed by Liapunov. Some of the extensive developments of this theory are particularly convenient for application to the class of systems considered during this investigation.

Assume a free system given by

$$\dot{\underline{Y}} = \hat{\underline{F}}(\underline{Y}, t) \qquad (2\text{-}36)$$

defined on $R^n \times R^1$ and, for convenience, also assume $\underline{Y} = \underline{0}$ is an equilibrium point, i.e.,

$$\hat{\underline{F}}(\underline{0}, t) = \underline{0} \qquad \text{for all } t \in R^1 . \qquad (2\text{-}37)$$

Definition 2-7: Stability.[9]

The origin of the system given by (2-36) is said to be stable with respect to t_o if for every $\varepsilon > 0$, there exists $\delta(\varepsilon, t_o) > 0$ such that

$$||\underline{Y}_o|| < \delta \qquad (2\text{-}38)$$

implies

$$||\hat{\underline{\Phi}}_f(t; \underline{Y}_o, t_o)|| < \varepsilon \quad \text{for all } t \geq t_o , \qquad (2\text{-}39)$$

where $\hat{\underline{\Phi}}_f(t; \underline{Y}_o, t_o)$ is a solution of (2-36) from $(\underline{Y}_o, t_o)$.

When the system is a dynamical system, and stability exists for some t_o, then it is stable for any other $t_1 \in R$.[14] Accordingly, the expression "with respect to t_o" can be eliminated from the definition.

Definition 2-8: Asymptotic Stability.[9]

The origin of the system given by (2-36) is said to

be asymptotically stable if

(i) it is stable, and

(ii) for every $\mu > 0$, there exists a $T(\mu, t_o, \delta) > 0$ such that

$$||\underline{\hat{\Phi}}_f(t; \underline{Y}_o, t_o)|| < \mu \quad \text{for all } t \geq t_o + T . \tag{2-40}$$

Heuristically every solution starting in a neighborhood of the equilibrium point at any t_o is required to converge to $\underline{0}$ as $t \to \infty$.

Definition 2-9: Asymptotic Stability in the Large.[14]

The origin of the system of (2-36) is said to be asymptotically stable in the large if

(i) it is stable, and

(ii) every solution converges to $\underline{0}$ as $t \to \infty$.

These three definitions are generally considered to refer to stability in the Liapunov sense. Criteria for satisfying these definitions of stability are established in the so-called theory of the Liapunov direct method.

Theorem 2-8:[6]

Consider an autonomous system

$$\dot{\underline{Y}} = \hat{\underline{F}}(\underline{Y}) \qquad (2\text{-}41)$$

of which $\hat{\underline{F}}(\underline{Y})$ is defined on Γ such that $\underline{0} \in \Gamma$. If there exists scalar function $\hat{\nabla}(\underline{Y})$ of class C_1 in R^n such that

(i) $\hat{\nabla}(\underline{Y})$ is positive definite in Γ,

(ii) $\dot{\hat{\nabla}}(\underline{Y}) = [\frac{\partial \hat{\nabla}}{\partial \underline{Y}}]^T \hat{\underline{F}}(\underline{Y})$ is negative definite in Γ,

then the origin is asymptotically stable. If $\dot{\hat{\nabla}}$ is negative semidefinite in (ii), the origin is stable. If $\hat{\nabla}(\underline{Y})$ also satisfies

(iii) $\lim_{||\underline{Y}|| \to \infty} \hat{\nabla}(\underline{Y}) \to \infty$ in R^n

and Γ can be selected as the entire R^n, the origin is asymptotically stable in the large.

A function $\hat{\nabla}(\underline{Y})$ to identify stability in this sense is generally called a Liapunov function for the system.[14] As the existence of a Liapunov function for a given system is sufficient to guarantee stability without knowledge of specific solutions, the method is particularly valuable for nonlinear systems.

2.3.2 System Controllability

When a system is to be controlled so as to transfer its initial condition to another condition, there is a

question as to the realizability of the requirement, i.e., whether a control function exists for the transition. To summarize this idea, the concept of an admissible control function is introduced first.

Definition 2-10: Admissible Control Function.[5]

A control function $\underline{V}(t)$ defined on $[t_o, t_1] \subset R^1$ is said to be admissible if it is piecewise continuous on the interval.

In effect, an admissible control function provides a unique solution from an arbitrary initial condition to (2-21), as provided by Theorem 2-5.

Definition 2-11: Controllability.[5]

For a system given by

$$\dot{\underline{Y}} = \hat{\underline{F}}(\underline{Y}, \underline{V}, t) \ , \qquad (2\text{-}42)$$

a state $\underline{Y}_o$ is said to be controllable at t_o with respect to a state $\underline{Y}_f$ if there exists an admissible control function $\underline{V}(t)$ defined on $[t_o, t_f]$ such that

$$\hat{\underline{\Phi}}_{\underline{V}}(t_f; \underline{Y}_o, t_o) = \underline{Y}_f \quad \text{for some } t_f \geq t_o \ . \qquad (2\text{-}43)$$

If every $\underline{Y}_o$ is controllable with respect to every $\underline{Y}_f$ at every t_o, then the system is said to be completely controllable. When the system is linear such that

$$\dot{\underline{Y}} = \hat{\underline{A}}\,\underline{Y} + \hat{\underline{B}}\,\underline{V}\ , \qquad (2\text{-}44)$$

where $\hat{\underline{A}}$ and $\hat{\underline{B}}$ are n x n and n x m constant matrices respectively, a criterion of complete controllability is known.

Theorem 2-9:[5]

The linear system given by (2-44) is completely controllable if and only if the n x mn matrix

$$[\hat{\underline{B}},\ \hat{\underline{A}}\,\hat{\underline{B}},\ \hat{\underline{A}}^2\hat{\underline{B}},\ \ldots,\ \hat{\underline{A}}^{n-1}\hat{\underline{B}}] \qquad (2\text{-}45)$$

has rank n.

The following two definitions, extending this idea, are used in later chapters.

Definition 2-12: Reachability.[5]

For a system (2-42), a state $\underline{Y}_1$ is said to be reachable with respect to a state $\underline{Y}_o$ at t_o if there is an admissible control function $\underline{V}(t)$ defined on $[t_o,\ t_f]$ such

that it provides a solution $\hat{\underline{\Phi}}_{\underline{V}}(t; \underline{Y}_o, t_o)$ which satisfies

$$\hat{\underline{\Phi}}_{\underline{V}}(t_1; \underline{Y}_o, t_o) = \underline{Y}_1 \quad \text{for some finite } t_1 \geq t_o . \tag{2-46}$$

For the case of an unlimited t_f, a new definition is introduced.

Definition 2-13: Approachability.

For a system given by (2-42), a state $\underline{Y}_1$ is said to be approachable with respect to a state $\underline{Y}_o$ at t_o if there exist an admissible control function $\underline{V}(t)$ defined on $[t_o, \infty)$ to provide a solution $\hat{\underline{\Phi}}_{\underline{V}}(t; \underline{Y}_o, t_o)$ such that for each $\varepsilon > 0$, there exists a $T > 0$ which satisfies

$$||\hat{\underline{\Phi}}_{\underline{V}}(t; \underline{Y}_o, t_o) - \underline{Y}_1|| < \varepsilon \quad \text{for all } t \geq t_o + T . \tag{2-47}$$

If the $\underline{Y}_1$ is approachable with respect to every state at every t_o, $\underline{Y}_1$ is said to be totally approachable for the system.

Chapter 3

CANONICAL FORMS OF LINEAR SYSTEMS

This chapter reviews the work that has been done to develop canonical forms for linear time-invariant systems. Some fundamental properties of canonical forms are also stated for use in subsequent chapters. A canonical form proposed by Kalman and another based on the Jordan form are described. Others that are described, called phase variable canonical forms, represent original work by Luenberger, Teul, and Asseo.

3.1 Canonical Forms of Systems

It often occurs that two different but mathematically similar systems are sufficiently related that the analyses and solutions for one can be applied to the other with relatively minor modifications. If this can be done among many systems forming a group, it is reasonable to analyze and solve the one system of the group offering the least complexity, and then translate the results to the other systems of the group.

A canonical form is a compact standard form for describing all systems that are mathematically similar. Similarity is associated with an equivalent relation on the set of systems of interest, say S. The equivalent relation used historically to develop canonical forms is as follows.

Definition 3-1: Equivalent System.[5]

Two systems are said to be equivalent if there exists an n x n nonsingular constant matrix $\underline{N}$ such that

$$\underline{N}^{-1}\,\underline{Y} = \underline{X}\,, \tag{3-1}$$

where $\underline{Y}$ and $\underline{X}$ are state vectors of each system.

It can be shown that equivalent systems in S are related by topological relations called reflexive, symmetric and transitive laws, i.e., the nonsingular transformation (3-1) is a topological equivalent relation defined on S. It is known, further, that topological equivalence preserves the stability properties of dynamical systems.[12]

The value of a canonical form depends on the convenience of the specified structure in practical analyses and the extent to which it displays noteworthy characteristics. Historically, the development of canonical forms has been limited to linear systems. These forms can be described by considering the linear system

$$\dot{\underline{Y}} = \hat{\underline{A}}\,\underline{Y} + \hat{\underline{B}}\,\underline{V} \tag{3-2}$$

where $\hat{\underline{A}}$ and $\hat{\underline{B}}$ are assumed n x n and n x m matrices with r the rank of $\hat{\underline{B}}$. By (3-1), an equivalent system is

$$\dot{\underline{X}} = \underline{A}\,\underline{X} + \tilde{\underline{B}}\,\underline{V}\,, \qquad (3\text{-}3)$$

where

$$\underline{A} = \underline{N}^{-1}\hat{\underline{A}}\,\underline{N}$$

and

$$\tilde{\underline{B}} = \underline{N}^{-1}\hat{\underline{B}}\,. \qquad (3\text{-}4)$$

A canonical form is concerned with specifying a particular structure for $\underline{A}$ and $\tilde{\underline{B}}$.

In a general sense, Kalman suggested a canonical form for linear systems so as to conveniently display the controllability property of the system.[12] He stated that $\underline{A}$ and $\tilde{\underline{B}}$ can be specified as

$$\underline{A} = \begin{bmatrix} \underline{A}_{11} & [0] \\ \underline{A}_{21} & \underline{A}_{22} \end{bmatrix} \qquad (3\text{-}5)$$

and

$$\tilde{\underline{B}} = \begin{bmatrix} [0] \\ \tilde{\underline{B}}_2 \end{bmatrix}\,, \qquad (3\text{-}6)$$

where the dimensions of the submatrices are

$$\underline{A}_{11};\ n_1 \times n_1 \qquad \underline{A}_{21};\ n_2 \times n_1$$
$$\underline{A}_{22};\ n_2 \times n_2 \qquad \underline{\tilde{B}}_2;\ n_2 \times m\ , \tag{3-7}$$

such that $n = n_1 + n_2$ and $n_1 \geq 0$, $n_2 \geq 0$. Thus, an equivalent system to the original system is

$$\dot{\underline{X}}_1 = \underline{A}_{11}\,\underline{X}_1$$
$$\dot{\underline{X}}_2 = \underline{A}_{21}\,\underline{X}_1 + \underline{A}_{22}\,\underline{X}_2 + \underline{\tilde{B}}_2\,\underline{V}\ , \tag{3-8}$$

where $\underline{X}_1 = [x_1, x_2, \ldots, x_{n_1}]^T$ and $\underline{X}_2 = [x_{n_1+1}, \ldots, x_n]^T$. Assuming the rank of $\underline{\hat{B}} = r > 0$, Kalman defined the controllability of systems based on this description, instead of Definition 2-11.

Definition 3-2:[12] Controllability (Kalman).

The system (3-2) is completely controllable if it is not equivalent to the system (3-8) with an $n_1 > 0$.

The following theorem gives the equivalency between Definition 2-11 and 3-2.

Theorem 3-1:

The system (3-2) is completely controllable in the

sense of Definition 3-2 if and only if the conditions in Theorem 2-9 are satisfied.

Proof: Assume that (3-2) is equivalent to (3-8) with an $n_1 > 0$. Then from (3-4), (3-5), and (3-6), it follows that

$$[\underline{\hat{B}},\ \underline{\hat{A}}\,\underline{\hat{B}},\ \ldots,\ \underline{\hat{A}}^{n-1}\underline{\hat{B}}] = [\underline{N}\,\underline{\tilde{B}},\ \underline{N}\,\underline{A}\,\underline{\tilde{B}},\ \ldots,\ \underline{N}\,\underline{A}^{n-1}\underline{\tilde{B}}]$$

$$= \underline{N} \left[\begin{bmatrix} [0] \\ \underline{\tilde{B}}_2 \end{bmatrix}, \begin{bmatrix} [0] \\ \underline{A}_{22}\underline{\tilde{B}}_2 \end{bmatrix}, \ldots, \begin{bmatrix} [0] \\ \underline{A}_{22}^{n-1}\underline{\tilde{B}}_2 \end{bmatrix} \right], \tag{3-9}$$

where each [0] is $n_1 \times m$. As the rank of (3-9) is less than n, the system is not controllable by Theorem 2-9. Conversely, assume (3-2) is completely controllable. From Theorem 2-9, the matrix

$$[\underline{\hat{B}},\ \underline{\hat{A}}\,\underline{\hat{B}},\ \ldots,\ \underline{\hat{A}}^{n-1}\underline{\hat{B}}] \tag{3-10}$$

must have rank n. Assume that a nonsingular $\underline{N}$ exists such that $\underline{\hat{A}}$ and $\underline{\hat{B}}$ are transformed by (3-4) to (3-5) and (3-6). Then

$$\underline{N}^{-1}[\underline{\hat{B}},\ \underline{\hat{A}}\,\underline{\hat{B}},\ \ldots,\ \underline{\hat{A}}^{n-1}\underline{\hat{B}}] = [\underline{N}^{-1}\underline{\hat{B}},\ \underline{N}^{-1}\underline{\hat{A}}\,\underline{\hat{B}},\ \ldots,\ \underline{N}^{-1}\underline{\hat{A}}^{n-1}\underline{\hat{B}}] \tag{3-11}$$

must have the rank n. But for each $0 \leq i \leq n-1$,

$$\underline{N}^{-1}\hat{\underline{A}}^{i}\hat{\underline{B}} = \underbrace{(\underline{N}^{-1}\hat{\underline{A}}\,\underline{N})(\underline{N}^{-1}\hat{\underline{A}}\,\underline{N}) \cdots (\underline{N}^{-1}\hat{\underline{A}}\,\underline{N})}_{\text{i stages}}(\underline{N}^{-1}\hat{\underline{B}}) = \underline{A}^{i}\tilde{\underline{B}} \tag{3-12}$$

and (3-11) is reduced to

$$\left[\begin{bmatrix}[0]\\ \tilde{\underline{B}}_2\end{bmatrix}, \begin{bmatrix}[0]\\ \underline{A}_{22}\tilde{\underline{B}}_2\end{bmatrix}, \cdots, \begin{bmatrix}[0]\\ \underline{A}_{22}^{n-1}\tilde{\underline{B}}_2\end{bmatrix}\right], \tag{3-13}$$

where each [0] is n_1 x m. For (3-10) to have the rank n then, n_1 must be zero, that is, the system must not be equivalent to (3-8) with an $n_1 > 0$.

3.2 Jordan Matrix Canonical Form

A canonical form exists if $\underline{A}$ in (3-3) is given by a Jordan canonical form of matrix.[8,15] That is, an $\underline{N}$ can be chosen for a similar transformation of $\hat{\underline{A}}$ such that

$$\underline{A} = \underline{N}^{-1}\hat{\underline{A}}\,\underline{N} = \begin{bmatrix} \underline{J}_{(1)}(\lambda_1) & & & \\ & \underline{J}_{(2)}(\lambda_2) & & \\ & & \ddots & \\ & & & \underline{J}_{(v)}(\lambda_v) \end{bmatrix},$$

all other entries zero, (3-14)

where each $\underline{J}_{(i)}(\lambda_i)$, $(i = 1,2,\ldots,v)$, is an $\ell_i \times \ell_i$ matrix given by

$$\underline{J}_{(i)}\lambda_{(i)} = \begin{bmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix}, \text{ all other entries zero,} \quad (3\text{-}15)$$

$\lambda_1,\lambda_2,\ldots,\lambda_v$ are eigenvalues of $\hat{\underline{A}}$ and v, $\ell_1, \ldots, \ell_v$ are positive integers of the characteristic equation of $\hat{\underline{A}}$,

$$0 = |\hat{\underline{A}}-\lambda I| = (\lambda-\lambda_1)^{\ell_1} \cdot (\lambda-\lambda_2)^{\ell_2} \ldots (\lambda-\lambda_v)^{\ell_v}. \quad (3\text{-}16)$$

Among the numbers $\lambda_1,\lambda_2,\ldots,\lambda_v$, some may be equal, however, one combination of $\underline{N}, v, \ell_1, \ell_2, \ldots, \ell_v$ consistent with (3-15) and (3-16) is guaranteed to exist for the given $\hat{\underline{A}}$.[13] It is convenient to decompose the resulting

$\underline{X}$ and $\underline{\tilde{B}}$ as

$$\underline{X} = [\underline{X}_{(1)}^T, \underline{X}_{(2)}^T, \ldots, \underline{X}_{v}^T]^T \tag{3-17}$$

$$\text{and } \underline{\tilde{B}} = \underline{N}^{-1}\underline{\hat{B}} = [\underline{\tilde{B}}_{(1)}, \underline{\tilde{B}}_{(2)}, \ldots, \underline{\tilde{B}}_{(v)}]^T , \tag{3-18}$$

where each $\underline{\tilde{B}}_{(i)}$, $(i = 1,2,\ldots,v)$, is an ℓ_i x m matrix and

$$\begin{aligned}
\underline{X}_{(1)} &= [x_1, x_2, \ldots, x_{\ell_1}]^T \\
\underline{X}_{(2)} &= [x_{\ell_1+1}, x_{\ell_1+1}, \ldots, x_{\ell_1+\ell_2}]^T \\
&\vdots \\
\underline{X}_{(v)} &= [x_{v-1}, x_{v-1}, \ldots, x_n]^T .
\end{aligned} \tag{3-19}$$

$$(\textstyle\sum_i \ell_i+1) \quad (\textstyle\sum_i \ell_i+2)$$

Then this canonical form can be written as

$$\begin{aligned}
\dot{\underline{X}}_{(1)} &= \underline{J}_{(1)}(\lambda_1)\ \underline{X}_{(1)} + \underline{\tilde{B}}_{(1)}\ \underline{V} && (S_1) \\
\dot{\underline{X}}_{(2)} &= \underline{J}_{(2)}(\lambda_2)\ \underline{X}_{(1)} + \underline{\tilde{B}}_{(2)}\ \underline{V} && (S_2) \\
&\vdots && \vdots \\
\dot{\underline{X}}_{(v)} &= \underline{J}_{(v)}(\lambda_v)\ \underline{X}_{(v)} + \underline{\tilde{B}}_{(v)}\ \underline{V} , && (S_v)
\end{aligned} \tag{3-20}$$

which can be represented by v subsystems each appearing as shown in Figure 3-1.

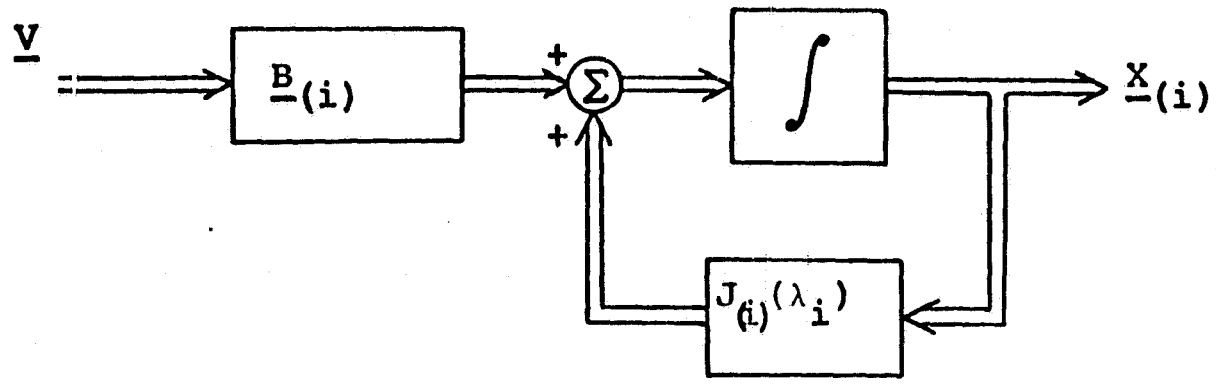


Figure 3-1. Subsystem (S_i) for (3-20).

The principal advantage of this canonical form is the convenience of calculating system solutions. This is evident as the free motion of the system is identified from the eigenvalues of $\underline{\hat{A}}$, and the system is effectively decomposed into independent subsystems with respect to the state variables.[8]

3.3 Phase Variable Canonical Form

In this section, the system (3-2) is assumed

(i) completely controllable, and

(ii) $\underline{\hat{B}}$ is of full rank, r = m.

3.3.1 Phase Variable Canonical Form For Single Input Systems[12,16]

Two canonical forms are known for single input linear systems, i.e., m = 1, meeting the above assumptions. One has the structures of $\underline{A}$ and $\underline{\tilde{B}}$ given as

$$\underline{A} = \underline{N}^{-1}\hat{\underline{A}}\,\underline{N} = \begin{bmatrix} 0 & 1 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ -a_1 & -a_2 & \cdots & \cdots & -a_n \end{bmatrix},$$

all other entries zero, (3-21)

and $$\tilde{\underline{B}} = \underline{N}^{-1}\hat{\underline{B}} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}. \qquad (3\text{-}22)$$

This form can be illustrated as shown in Figure 3-2.

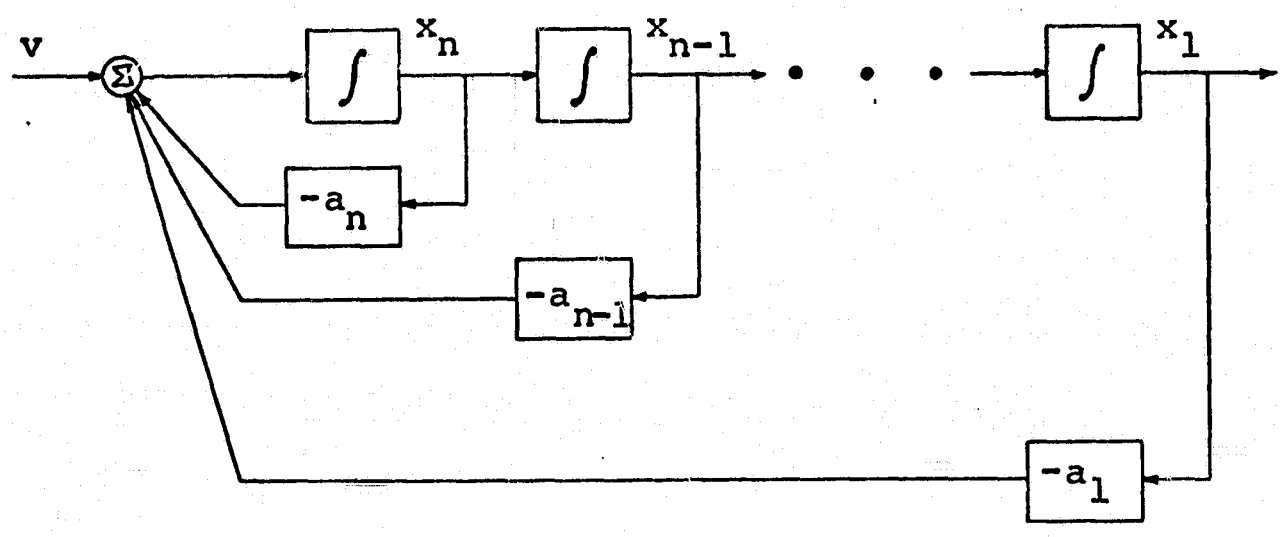


Figure 3-2. Single input phase variable canonical form given by (3-21) and (3-22).

The second canonical form has $\underline{A}$ and $\widetilde{\underline{B}}$ given as

$$\underline{A} = \begin{bmatrix} 0 & & & -a_n \\ 1 & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & 0 & -a_2 \\ & & 1 & -a_1 \end{bmatrix}, \text{ all other entries zero,} \tag{3-23}$$

and

$$\widetilde{\underline{B}} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{3-24}$$

which is illustrated as in Figure 3-3.

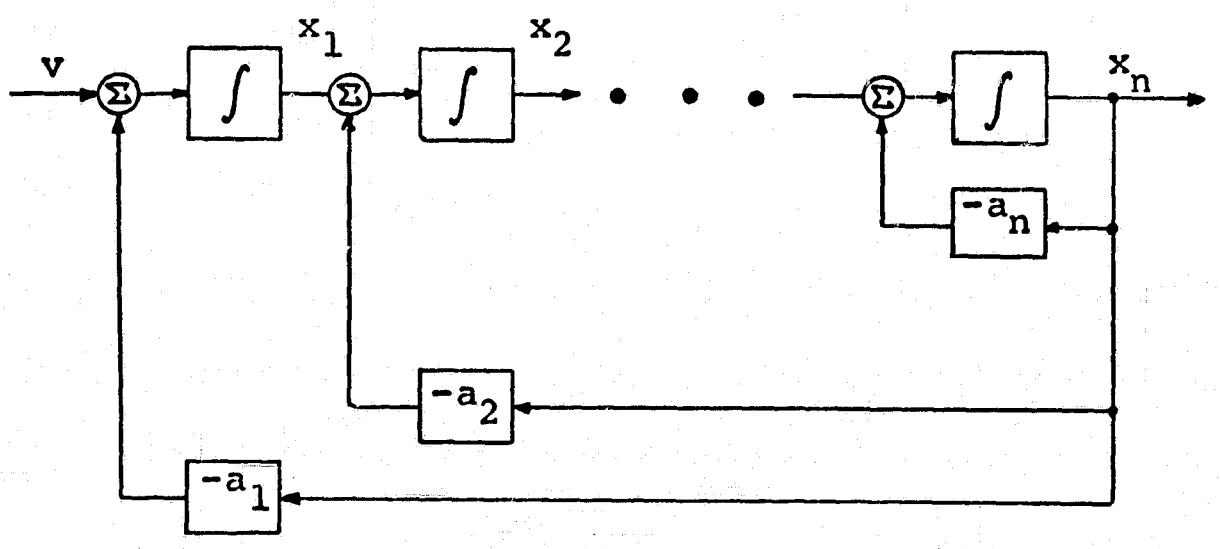


Figure 3-3. Single input phase variable canonical form given by (3-23) and (3-24).

Consequently, the system is viewed as a cascade connection of n-integraters with ordered feedback paths. Since zero and unit elements in $\underline{A}$ and $\underline{\tilde{B}}$ are concentrated regularly, these canonical forms are conveniently used for abstract analyses of system theory. The second canonical form with (3-21) and (3-22), is particularly convenient for optimal control problems. Many papers have been published about reduction techniques for given systems into these canonical forms.[12,17-20]

3.3.2 Phase Variable Canonical Forms for Multi-Input Systems

3.3.2.1 Canonical Form by Tuel[21]

Tuel developed a canonical form, called a control canonical form, in which $\underline{A}$ and $\underline{\tilde{B}}$ are decomposed as

$$\underline{A} = \begin{bmatrix} \underline{A}_{11} & \underline{A}_{12} \\ \underline{A}_{21} & \underline{A}_{22} \end{bmatrix} \tag{3-25}$$

and

$$\underline{\tilde{B}} = \begin{bmatrix} \underline{\tilde{B}}(1) \\ \underline{\tilde{B}}(2) \end{bmatrix} \tag{3-26}$$

such that

(i) there exists a set of r positive integers ℓ_i, (i = 1,2,...,r), (where r is the rank of $\underline{\hat{B}}$), defined

for a given system such that $\sum_{i=1}^{r} \ell_i = n$,

(ii) $\underline{A}_{11}$ and $\underline{A}_{12}$ are (n-r) x (n-r) and (n-r) x r matrices such that

$$\underline{A}_{11} = \begin{bmatrix} \underline{A}_{(1,1)} & & & \\ & \underline{A}_{(2,2)} & & \\ & & \ddots & \\ & & & \underline{A}_{(r,r)} \end{bmatrix}, \text{ all other entries zero,} \qquad (3\text{-}27)$$

and

$$\underline{A}_{12} = \begin{bmatrix} \underline{E}_{(1)} & & & \\ & \underline{E}_{(2)} & & \\ & & \ddots & \\ & & & \underline{E}_{(r)} \end{bmatrix}, \text{ all other entries zero,} \qquad (3\text{-}28)$$

where $\underline{A}_{(i,i)}$ is an $(\ell_i\text{-}1)$ x $(\ell_i\text{-}1)$ matrix

$$\underline{A}_{(i,i)} = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix}, \text{ all other entries zero,} \qquad (3\text{-}29)$$

and $\underline{E}_{(i)}$ is an $(\ell_i\text{-}1)$ column vector

$$\underline{E}_{(i)} = \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 1 \end{bmatrix} , \qquad (3\text{-}30)$$

(iii) $\underline{A}_{21}$ and $\underline{A}_{22}$ are arbitrary r x (n - r) and r x r matrices,

(iv) $\widetilde{\underline{B}}_{(1)} = [0]$ and $\widetilde{\underline{B}}_{(2)}$ is an upper triangular matrix with unit elements on the diagonal.

When m = r = 1, this canonical form is reduced to the canonical form of the single input system, (3-21) and (3-22), i.e.,

$$\underline{A}_{11} = \underline{A}_{(1,1)} = \begin{bmatrix} 0 & 1 & & & \\ & \cdot & \cdot & & \\ & & \cdot & \cdot & \\ & & & \cdot & 1 \\ & & & & 0 \end{bmatrix} , \text{ all other entries zero,} \qquad (3\text{-}31)$$

$$\underline{A}_{12} = \underline{E}_{(1)} = \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 1 \end{bmatrix} , \qquad (3\text{-}32)$$

$$\widetilde{\underline{B}}_{(2)} = [1] , \qquad (3\text{-}33)$$

$\underline{\tilde{B}}_{(1)}$ is the (n-1) x 1 null matrix, and $\underline{A}_{21}$ and $\underline{A}_{22}$ are 1 x (n-1) and 1 x 1 matrices respectively.

This phase canonical form was developed early for solving multi-input optimum control system problems. It compactly describes the original system and illustrates its mathematical structure by an ordered array of zero and unity entries. Another comparable canonical form called an observation canonical form, can also be developed in this manner but is omitted here.[21]

3.3.2.2 Canonical Form by Luenberger[22]

Luenberger suggested a canonical form for linear multi-input systems (3-2), in which $\underline{A}$ and $\underline{\tilde{B}}$ are decomposed into r^2 and r submatrices

$$\underline{A} = \begin{bmatrix} \underline{A}_{(1,1)} & \underline{A}_{(1,2)} & \cdots & \underline{A}_{(1,r)} \\ \vdots & & & \vdots \\ \underline{A}_{(r,1)} & \cdot & \cdots & \underline{A}_{(r,r)} \end{bmatrix} \qquad (3\text{-}34)$$

and

$$\underline{\tilde{B}} = \begin{bmatrix} \underline{\tilde{B}}_{(1)} \\ \underline{\tilde{B}}_{(2)} \\ \vdots \\ \underline{\tilde{B}}_{(r)} \end{bmatrix} \qquad (3\text{-}35)$$

such that

(i) there exists a set of r positive integers ℓ_i, (i = 1,2,...r), (r the rank of $\hat{\underline{B}}$), such that

$$\sum_{i=1}^{r} \ell_i = n ,$$

(ii) each $\underline{A}_{(i,j)}$, (i,j = 1,2,...,r), is an ℓ_i x ℓ_j matrix such that

$$\underline{A}_{(i,i)} = \begin{bmatrix} 0 & 1 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ -a_1 & \cdots & \cdots & \cdots & -a_n \end{bmatrix} , \text{ all other entries zero,} \qquad (3\text{-}36)$$

$$\underline{A}_{(i,j)} = \begin{bmatrix} & & & \\ & & & \\ & & & \\ * & * & \cdots & * \end{bmatrix} , \; i \neq j, \text{ all other entries zero,} \qquad (3\text{-}37)$$

where *'s indicate arbitrary elements,

(iii) each $\underline{\tilde{B}}_{(i)}$ (i = 1,...r), is an ℓ_i x r matrix given by

$$\underline{\tilde{B}}_{(i)} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \leftarrow i^{th} \text{ column}, \quad \text{all other entries zero.} \qquad (3\text{-}38)$$

The form is illustrated in Figure 3-4. With this canonical form, a multi-input system can be viewed as interactions of r individual single-input systems, each with the canonical form given as (3-21) and (3-22). Consequently, the conveniences of the phase variable canonical form for single-input systems can be appreciated for the multi-input systems. Again a modification of this canonical form can be developed corresponding to the use of (3-23) and (3-24) instead of (3-21) and (3-22); this is omitted.[22]

3.3.2.3 Canonical Form by Asseo[23]

Asseo described a canonical form in which $\underline{A}$ and $\underline{\tilde{B}}$ are decomposed into (3-25) and (3-26) such that the following are satisfied.

(i) $\underline{A}_{11}$ is the (n-r) x r null matrix and $\underline{A}_{12}$ is the (n-r) x (n-r) unit matrix,

(ii) $\underline{A}_{21}$ and $\underline{A}_{22}$ are r x r and r x (n-r) arbitrary matrices, and

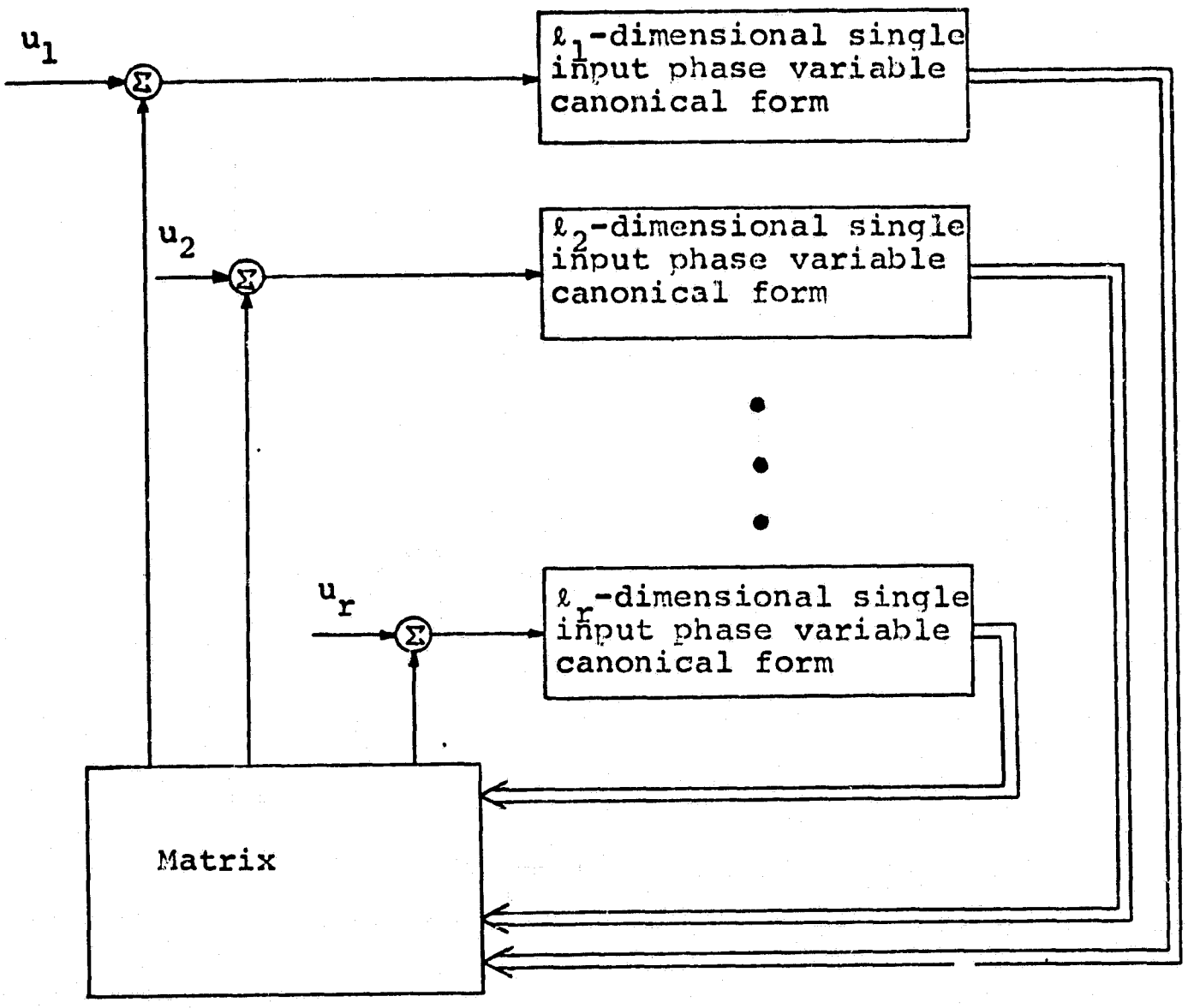


Figure 3-4. Luenberger canonical form.

(iii) $\underset{\sim}{\widetilde{B}}_{(1)}$ is a (n-r) x r null matrix and $\underset{\sim}{\widetilde{B}}_{(2)}$ is the r x r unit matrix.

The structure of this form appears simple and convenient for analysis in comparison with other suggested forms. Contrary to his assertion, however, the canonical form cannot be used for all systems which satisfy the two assumptions of Section 3.3.

As a counter example to illustrate this, consider a system (3-2) with n = 4, r = 2, i.e.,

$$\underset{\sim}{\hat{A}} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \tag{3-39}$$

and

$$\underset{\sim}{\hat{B}} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{3-40}$$

It can be shown that this system is completely controllable by Theorem 2-9. From (3-4) let $\underset{\sim}{N}^{-1}$ for the equivalent reduction to $\underset{\sim}{A}$ and $\underset{\sim}{\widetilde{B}}$ be

$$\underline{N}^{-1} = \begin{bmatrix} n_{11} & n_{12} & n_{13} & n_{14} \\ n_{21} & & & \cdot \\ \cdot & & & \cdot \\ n_{41} & \cdot & \cdot & n_{44} \end{bmatrix} . \qquad (3\text{-}41)$$

To obtain $\underline{\tilde{B}}$ according to statement (iii) above, it follows that $n_{13} = n_{14} = n_{23} = n_{24} = n_{34} = n_{43} = 0$ and $n_{33} = n_{44} = 1$. But in addition, from $\underline{A} = \underline{N}^{-1}\ \underline{\hat{A}}\ \underline{N}$, (3-41) and (i) above,

$$\underline{A}\ \underline{N}^{-1} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ * & * & * & * \\ * & * & * & * \end{bmatrix} \begin{bmatrix} n_{11} & n_{12} & 0 & 0 \\ n_{21} & n_{22} & 0 & 0 \\ n_{31} & n_{32} & 1 & 0 \\ n_{41} & n_{42} & 0 & 1 \end{bmatrix} = \begin{bmatrix} n_{31} & n_{32} & 1 & 0 \\ n_{41} & n_{42} & 0 & 1 \\ * & * & * & * \\ * & * & * & * \end{bmatrix}$$

(3-42)

and from (3-39)

$$\underline{N}^{-1}\ \underline{\hat{A}} = \begin{bmatrix} n_{11} & n_{12} & 0 & 0 \\ n_{21} & n_{22} & 0 & 0 \\ n_{31} & n_{32} & 1 & 0 \\ n_{41} & n_{42} & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & n_{11} & 0 & n_{12} \\ 0 & n_{22} & 0 & n_{22} \\ 0 & n_{31} & 0 & n_{32} \\ 0 & n_{41} & 0 & n_{42} \end{bmatrix}$$

(3-43)

where * indicates arbitrary elements.

The identity of (3-42) and (3-43) fails for the (1,3) element.

3.4 Comments

The canonical forms described in this chapter display the internal or elemental mathematical structure of systems from different viewpoints. The usefulness of the phase canonical forms for optimal control problems, however, could be improved if (i) a greater number of elements in $\underline{A}$ and $\underline{\tilde{B}}$ were reduced to zero or unity, and (ii) these zero and unit elements were arranged in both a simple and unique order.

It is shown in the next chapter that the phase canonical form for single input systems is uniquely determined. It would appear, therefore, that this canonical form cannot be improved for this application. There are possibilities for improvement, however, when multi-input systems are considered. Both the Luenberger and Tuel canonical forms are valid for multi-input systems. While they have the same number of zero and unit elements in $\underline{A}$ and $\underline{\tilde{B}}$, they both have some ambiguity or arbitrariness about the dimensions of the decomposed submatrices; these dimensions are not unique but depend upon the chosen matrix $\underline{N}$. While the Asseo canonical form has a particularly simplified structure and is free from submatrix dimensional

ambiguity, the application is for a more limited subclass of systems.

Applications of known phase canonical forms are also restricted to systems which are completely controllable and with full rank of $\hat{\underline{B}}$. The possibility exists, therefore, to remove this limitation. The extension of canonical forms for use with nonlinear systems has also been avoided in past work.

The new canonical form given in the next chapter is developed to have a unique and regular distribution of zero and unit elements in $\underline{A}$ and $\tilde{\underline{B}}$ for linear systems, without the restrictive assumptions of controllability or on the rank of $\hat{\underline{B}}$. The canonical form is also proposed for use with a class of nonlinear systems by application to the linear part of these systems.

Chapter 4

DEVELOPMENT OF A CANONICAL FORM

In this chapter, a new phase variable canonical form is developed for a class of multi-input systems. This particular canonical form is shown to be superior to those reviewed in the previous chapter. It also provides the form for the analysis of the inverse problem of the optimal regulator in subsequent chapters. First a new definition of an equivalent system is given. Section 4.2 introduces two theorems on matrix transformations. Based on these ideas, the new canonical form is then presented in Section 4.3. The uniqueness of the structure of the canonical form for linear systems is discussed, and finally the canonical form is extended for applications to a class of nonlinear systems.

Throughout this chapter only, matrices are indicated by capital Roman or Greek letters without the underline and are constant unless otherwise noted. Vectors are underlined.

4.1 Equivalent System

Systems which are considered in this chapter are given by

$$\dot{\underline{Y}} = \hat{\underline{F}}(\underline{Y}) + \hat{B}\,\underline{V} \qquad (4\text{-}1)$$

where $\hat{\underline{F}}(\underline{Y}) = [\hat{f}_1(\underline{Y}), \hat{f}_2(\underline{Y}), \ldots, \hat{f}_n(\underline{Y})]^T$, $\hat{B}$ is an n x m matrix with $0 < m \leq n$ and the rank $\hat{B}$ is r, $0 < r \leq m$. One definition of an equivalent system is given by Definition 3-1. A more rigorous statement is possible, however, for the particular class of systems given by (4-1).

Definition 4-1: Equivalent System.

For two arbitrary systems

$$\dot{\underline{Y}} = \hat{\underline{F}}(\underline{Y}) + \hat{B}\,\underline{V}\,, \quad S_1$$
$$\dot{\underline{X}} = \underline{F}(\underline{X}) + B\,\underline{U}\,, \quad S_2\,, \tag{4-2}$$

S_1 is said to be equivalent to S_2 if there exist non-singular matrices N and M which are n x n and m x m respectively, satisfying

$$\underline{X} = N^{-1}\,\underline{Y}$$
$$\underline{U} = M^{-1}\,\underline{V}\,, \tag{4-3}$$

or equivalently

$$\underline{F}(\underline{X}) = N^{-1}\,\hat{\underline{F}}(N\,\underline{X})$$
$$B = N^{-1}\,\hat{B}\,M\,. \tag{4-4}$$

The nonsingular transformation (4-3) can be identified as a topological equivalent relation[24] defined for this class of systems. The reflective law is satisfied directly by choosing M and N to be unit matrices. The symmetric law is satisfied by substituting

$$\begin{aligned} \underline{Y} &= N\,\underline{X} \\ \underline{V} &= M\,\underline{U} \end{aligned} \qquad (4\text{-}5)$$

into (4-4). To show that the transitive law is also satisfied, consider a system

$$\dot{\underline{Z}} = \tilde{\underline{F}}(\underline{Z}) + \tilde{B}\,\underline{W}\ , \quad S_3\ , \qquad (4\text{-}6)$$

to which S_2 is equivalent. Thus there exist nonsingular matrices $\tilde{N}$ and $\tilde{M}$ such that

$$\begin{aligned} \underline{Z} &= \tilde{N}^{-1}\,\underline{X} \\ \underline{W} &= \tilde{M}^{-1}\,\underline{U}\ . \end{aligned} \qquad (4\text{-}7)$$

S_1 is made an equivalent system of S_3 by defining

$$\begin{aligned} \underline{Z} &= (N\,\tilde{N})^{-1}\,\underline{Y} \\ \underline{W} &= (M\,\tilde{M})^{-1}\,\underline{V}\ , \end{aligned} \qquad (4\text{-}8)$$

where both $(N\ \tilde{N})$ and $(M\ \tilde{M})$ are nonsingular.

If M in (4-3) is fixed as the unit matrix, Definition 4-1 is essentially reduced to Definition 3-1. Thus, Definition 4-1 is a generalization of Definition 3-1, but with application restricted to the class of systems given by (4-1). The flexibility of the additional matrix M, however, allows a more compact canonical form to be given for the class of systems (4-1).

4.2 Fundamental Theorems

Theorem 4-1:

Let A be an n x n matrix. For each positive integer $s \leq n$, there exists an n x n nonsingular matrix N which transforms A into

$$A \triangleq N^{-1}AN \triangleq \begin{bmatrix} A_{(1,1)} & A_{(1,2)} & \cdots & A_{(1,v)} \\ \vdots & & & \vdots \\ A_{(v,1)} & \cdot & \cdots & A_{(v,v)} \end{bmatrix}$$

$$= \begin{bmatrix} A_{(1,1)} & A_{(1,2)} & [0] & \cdots & [0] \\ [0] & [0] & A_{(2,3)} & \cdots & \vdots \\ \vdots & & \ddots & \ddots & [0] \\ [0] & \cdots & [0] & A_{(v-1,v)} & \\ A_{(v,1)} & \cdots & \cdots & A_{(v,v)} & \end{bmatrix}, \quad (4\text{-}9)$$

where each submatrix $A_{(i,j)}$, $(i,j = 1,2,\ldots,v)$, is $\ell_i \times \ell_j$ such that

(i) $\ell_v = s$; (4-10)

(ii) $v, \ell_1, \ell_2, \ldots, \ell_{v-1}$ are positive integers dependent on A and s such that

$$\sum_{i=1}^{v} \ell_i = n \ ; \quad (4\text{-}11)$$

(iii) if $v \geq 2$, then either

(a) $A_{(1,1)}$ is a Jordan canonical form and

$$A_{(1,2)} = [0] \text{ , or} \tag{4-12}$$

(b)
$$A_{(1,1)} = [0] \tag{4-13}$$

$$A_{(1,2)} = \begin{bmatrix} [0] & I_{\ell_1} \end{bmatrix} , \tag{4-14}$$

where, if (b), then $\ell_1 \leq \ell_2$;

(iv) in addition, if $v \geq 3$, then

(a)
$$\ell_i \leq \ell_j \quad \text{for } 2 \leq i \leq j \leq v , \tag{4-15}$$

(b)
$$A_{(i,i+1)} = \begin{bmatrix} [0] & I_{\ell_i} \end{bmatrix} \quad \text{for } 2 \leq i \leq v-1 , \tag{4-16}$$

(c) $A_{(v,i)}$, for $i = 1,2,\ldots,v$, are unspecified.

The proof of the theorem follows three lemmas.

Lemma 4-1:

(a) If H is a $\theta_1 \times \theta_2$ matrix with rank θ_3, there exist nonsingular matrices H_1 and H_2 which are $\theta_1 \times \theta_1$ and $\theta_2 \times \theta_2$ respectively, satisfying

$$H_1 \, H \, H_2 = \begin{bmatrix} [0] & [0] \\ [0] & I_{\theta_3} \end{bmatrix} . \qquad (4\text{-}17)$$

(b) If H is full rank, i.e., $\theta_3 = \text{Min}(\theta_1, \theta_2)$, and $\theta_1 < \theta_2$ ($\theta_1 > \theta_2$) is satisfied, H_1(H_2) can be the unit matrix.

(c) If $\theta_1 = \theta_2$, then either H_1 or H_2 can be the unit matrix.

Proof: The reduction of H to (4-17) follows directly from the equivalence of matrices.[10] To show part (b) of the lemma, assume $\theta_1 < \theta_2$. Since $\theta_3 = \theta_1$, H has θ_1 independent columns. Let $H_2^{(1)}$ be a $\theta_2 \times \theta_2$ nonsingular matrix resulting from an interchange of columns of H such that

$$H \, H_2^{(1)} = \begin{bmatrix} H_3 & H_4 \end{bmatrix} , \qquad (4\text{-}18)$$

where H_4 is a $\theta_1 \times \theta_1$ nonsingular matrix composed of θ_1 independent columns of H. Define a $\theta_2 \times \theta_2$ nonsingular matrix such that

$$H_2^{(2)} \triangleq \begin{bmatrix} I_{\theta_2 - \theta_1} & [0] \\ -H_4^{-1} H_3 & H_4^{-1} \end{bmatrix} \qquad (4\text{-}19)$$

and

$$H_2 \triangleq H_2^{(1)} \; H_2^{(2)} \; .$$

Then it follows that

$$H \; H_2 = \begin{bmatrix} [0] & I_{\theta_1} \end{bmatrix} \qquad (4\text{-}20)$$

and H_1 is the unit matrix. If $\theta_2 < \theta_1$, a similar method can be applied. If $\theta_1 = \theta_2 = \theta_3$, then either H_1 or H_2 can be H^{-1}.

Lemma 4-2:

Let H_1 and H_2 be $\theta_2 \times \theta_1$ and $\theta_2 \times \theta_3$ matrices respectively with

$$0 \leq \theta_2 \leq \theta_3 \; , \qquad (4\text{-}21)$$

and let the rank of H_2 be θ_2. Then for any $\theta_1 > 0$, there exists a $\theta_3 \times \theta_1$ matrix K satisfying

$$H_1 = H_2 K \ . \tag{4-22}$$

Proof: The proof is complete if a construction of K can be demonstrated. Since H_2 is of full rank, from (4-21) and Lemma 4-1 there exists a $\theta_3 \times \theta_3$ nonsingular matrix K_1 such that

$$H_2 K_1 = \begin{bmatrix} [0] & I_{\theta_2} \end{bmatrix} \ . \tag{4-23}$$

Define K_2 to be a $\theta_3 \times \theta_1$ matrix such that

$$K_2 \triangleq \begin{bmatrix} [0] \\ \\ H_1 \end{bmatrix} \tag{4-24}$$

and define

$$K \triangleq K_1 K_2 \ . \tag{4-25}$$

Then K is $\theta_2 \times \theta_1$ and satisfies

$$H_2 K = H_1 \ . \tag{4-26}$$

Lemma 4-3:

Let A be an n x n matrix. For each positive integer $s \leq n$, there exists an n x n nonsingular matrix $\hat{N}$ to transform A into

$$\tilde{A} \triangleq \hat{N}^{-1}A\hat{N} \triangleq \begin{bmatrix} \tilde{A}_{(1,1)} & \tilde{A}_{(1,2)} & \cdots & \tilde{A}_{(1,v)} \\ \tilde{A}_{(2,1)} & & & \vdots \\ \vdots & & & \vdots \\ \tilde{A}_{(v,1)} & \cdot & \cdots & \tilde{A}_{(v,v)} \end{bmatrix}$$

$$= \begin{bmatrix} \tilde{A}_{(1,1)} & \tilde{A}_{(1,2)} & [0] & \cdots & [0] \\ \vdots & & \ddots & & \vdots \\ & & & \ddots & [0] \\ & & & & \tilde{A}_{(v-1,v)} \\ \tilde{A}_{(v,1)} & \cdot & \cdot & \cdot & \tilde{A}_{(v,v)} \end{bmatrix}, \qquad (4\text{-}27)$$

where each submatrix $\tilde{A}_{(i,j)}$, $(i,j = 1,2,\ldots,v)$, is ℓ_i x ℓ_j such that

(i) $\ell_v = s$; (4-28)

(ii) $v, \ell_1, \ell_2, \ldots \ell_{v-1}$ are positive integers dependent on A and s such that $\sum_{i=1}^{v} \ell_i = n$;

(iii) if $v \geq 2$, then either

(a) $$\tilde{A}_{(1,2)} = [0] \text{ , or} \tag{4-29}$$

(b) $$\tilde{A}_{(1,2)} = \left[[0] \quad I_{\theta_1} \right] , \tag{4-30}$$

where, for (b) it follows that $\ell_1 \leq \ell_2$; (4-31)

(iv) in addition, if $v \geq 3$,

(a) $$\ell_i \leq \ell_j \text{ , for } 2 \leq i \leq j \leq v \text{ ,} \tag{4-32}$$

(b) $$\tilde{A}_{(i,i+1)} = \left[[0] \quad \hat{N}_{ii}^{(i)^{-1}} \right] \text{ for } 2 \leq i \leq v-1 \text{ ,} \tag{4-33}$$

where $\hat{N}_{ii}^{(i)^{-1}}$ is a $\ell_i \times \ell_i$ nonsingular matrix,

(c) all other submatrices are unspecified.

Proof: (1) First, if s = n, define

$$\begin{aligned} v &= 1 \\ \ell_1 &= s \text{ .} \end{aligned} \tag{4-34}$$

The statement is satisfied by $\hat{N} = I$.

(2) If $s < n$, A can be decomposed as

$$A \triangleq \begin{bmatrix} \overbrace{A_{22}}^{n-s} & \overbrace{A_{21}}^{s} \\ A_{12} & A_{11} \end{bmatrix} \begin{matrix} \} \, n-s \\ \} \, s \end{matrix}, \qquad (4-35)$$

where each A_{ij}, $(i,j = 1,2)$, has the dimensions indicated. According to Lemma 4-1 $\hat{N}_{11}^{(1)}$ and $\hat{N}_{22}^{(1)}$ are s x s and (n-s) x (n-s) nonsingular matrices satisfying

$$\hat{N}_{22}^{(1)} A_{21} \hat{N}_{11}^{(1)} = \begin{bmatrix} [0] & [0] \\ [0] & I_{n_1} \end{bmatrix}, \qquad (4-36)$$

where n_1 is the rank of A_{21}. If $n_1 = s \leq n-s$, $\hat{N}_{11}^{(1)}$ can be the unit matrix, according to Lemma 4-1. Define an n x n nonsingular matrix such that

$$\hat{N}^{(1)} \triangleq \begin{bmatrix} \hat{N}_{22}^{(1)^{-1}} & [0] \\ [0] & \hat{N}_{11}^{(1)} \end{bmatrix}. \qquad (4-37)$$

Then

$$A^{(1)} \triangleq \hat{N}^{(1)^{-1}} A \hat{N}^{(1)} = \begin{bmatrix} \hat{N}_{22}^{(1)} A_{22} \hat{N}_{22}^{(1)^{-1}} & \hat{N}_{22}^{(1)} A_{21} \hat{N}_{11}^{(1)} \\ \hat{N}_{11}^{(1)^{-1}} A_{12} \hat{N}_{22}^{(1)^{-1}} & \hat{N}_{11}^{(1)^{-1}} A_{11} \hat{N}_{11}^{(1)} \end{bmatrix} . \tag{4-38}$$

(3) As a subclass of (2), if $n_1 = 0$ or $n_1 = n - s \leq s$ in (4-36),

$$\hat{N}_{22}^{(1)} A_{21} \hat{N}_{11}^{(1)} = \begin{cases} [0] \ ; \quad \text{if } n_1 = 0 & (4-39) \\ \begin{bmatrix} [0] & I_{n_1} \end{bmatrix} \ ; \text{ if } n_i = n - s \leq s \ . & (4-40) \end{cases}$$

Then define

$$\begin{aligned} v &= 2 \\ \ell_1 &= n - s \\ \ell_2 &= s \\ \hat{N} &= \hat{N}^{(1)} \ . \end{aligned} \tag{4-41}$$

It follows that

$$\tilde{A} = \begin{bmatrix} \tilde{A}_{(1,1)} & \tilde{A}_{(1,2)} \\ \tilde{A}_{(2,1)} & \tilde{A}_{(2,2)} \end{bmatrix} , \qquad (4\text{-}42)$$

where $\tilde{A}_{(1,2)}$ is either (4-39) or (4-40). For (4-40), it follows that

$$\ell_1 = \text{rank of } \tilde{A}_{(1,2)} = \text{Min}(\ell_1, \ell_2) \leq \ell_2 . \qquad (4\text{-}43)$$

Therefore the assertion is proved for the case of $s < n$ and $v = 2$.

(4) The remaining subclass of (2) is $0 < n_1 < n-s$. Again $A^{(1)}$ can be decomposed, using (4-36), as

$$A^{(1)} = \begin{bmatrix} A_{33}^{(1)} & A_{32}^{(1)} & A_{31}^{(1)} \\ A_{23}^{(1)} & A_{22}^{(1)} & A_{21}^{(1)} \\ A_{13}^{(1)} & A_{12}^{(1)} & A_{11}^{(1)} \end{bmatrix} = \begin{bmatrix} A_{33}^{(1)} & A_{32}^{(1)} & [0] \\ A_{23}^{(1)} & A_{22}^{(1)} & A_{21}^{(1)} \\ A_{13}^{(1)} & A_{12}^{(1)} & A_{11}^{(1)} \end{bmatrix} \begin{matrix} \} \, n-s-n_1 \\ \} \, n_1 \\ \} \, s \end{matrix} ,$$

$$\underbrace{\phantom{A_{33}^{(1)}}}_{(n-s-n_1)} \quad \underbrace{\phantom{A_{32}^{(1)}}}_{n_1} \quad \underbrace{\phantom{A_{21}^{(1)}}}_{s}$$

(4-44)

where each $A_{ij}^{(1)}$ has the dimension indicated. It also

follows from (4-36) that

$$A_{21}^{(1)} = \begin{bmatrix} [0] & I_{n_1} \end{bmatrix} . \qquad (4\text{-}45)$$

Using the previous process, let $\hat{N}_{22}^{(2)}$ and $\hat{N}_{33}^{(2)}$ be $n_1 \times n_1$ and $(n-s-n_1) \times (n-s-n_1)$ nonsingular matrices satisfying

$$\hat{N}_{33}^{(2)} A_{32}^{(1)} \hat{N}_{22}^{(2)} = \begin{bmatrix} [0] & [0] \\ [0] & I_{n_2} \end{bmatrix} , \qquad (4\text{-}46)$$

where n_2 is the rank of $A_{32}^{(1)}$. If $n_2 = n_1 \leq n - s - n_1$, $\hat{N}_{22}^{(2)}$ can be the unit matrix according to Lemma 4-1.

Define an n x n nonsingular matrix

$$N^{(2)} \triangleq \begin{bmatrix} \hat{N}_{33}^{(2)-1} & [0] & [0] \\ [0] & \hat{N}_{22}^{(2)} & [0] \\ [0] & [0] & I_s \end{bmatrix} \qquad (4\text{-}47)$$

and calculate

$$A^{(2)} \triangleq \hat{N}^{(2)^{-1}} A^{(1)} \hat{N}^{(2)}$$

$$= \begin{bmatrix} \hat{N}_{33}^{(2)} A_{33}^{(1)} \hat{N}_{33}^{(2)^{-1}} & \hat{N}_{33}^{(2)} A_{32}^{(1)} \hat{N}_{22}^{(2)} & [0] \\ \hat{N}_{22}^{(2)^{-1}} A_{23}^{(1)} \hat{N}_{33}^{(2)^{-1}} & \hat{N}_{22}^{(2)^{-1}} A_{22}^{(1)} \hat{N}_{22}^{(2)} & \begin{bmatrix} [0] & \hat{N}_{22}^{(2)^{-1}} \end{bmatrix} \\ A_{13}^{(1)} \hat{N}_{33}^{(2)^{-1}} & A_{12}^{(1)} \hat{N}_{22}^{(2)} & A_{11}^{(1)} \end{bmatrix}, \tag{4-48}$$

where (4-45) is used.

(5) A subclass of (4) is $n_2 = 0$ or $n_2 = n - s - n_1 \leq n_1$ in (4-48). From Lemma 4-1,

$$\hat{N}_{33}^{(2)} A_{32}^{(1)} \hat{N}_{22}^{(2)} = \begin{cases} [0] \; ; \quad \text{if} \quad n_2 = 0 & (4\text{-}49) \\ \begin{bmatrix} [0] & I_{n_2} \end{bmatrix} \; ; \; \text{if } n_2 = n - s - n_1 \leq n_1 \, . & (4\text{-}50) \end{cases}$$

Then define

$$v = 3$$

$$\ell_1 = n_2 = n - n_1 - s$$

$$\ell_2 = n_1 \qquad (4\text{-}51)$$

$$\ell_3 = s$$

$$\hat{N} = \hat{N}^{(1)} \; \hat{N}^{(2)}$$

and

$$\tilde{A} \triangleq A^{(2)} \triangleq \begin{bmatrix} \tilde{A}_{(1,1)} & \tilde{A}_{(1,2)} & [0] \\ \tilde{A}_{(2,1)} & \tilde{A}_{(2,2)} & \tilde{A}_{(2,3)} \\ \tilde{A}_{(3,1)} & \tilde{A}_{(3,2)} & \tilde{A}_{(3,3)} \end{bmatrix}, \qquad (4\text{-}52)$$

where $\tilde{A}_{(1,2)}$ is either (4-49) or (4-50) and

$$\tilde{A}_{(2,3)} = \begin{bmatrix} [0] & \hat{N}_{22}^{(2)^{-1}} \end{bmatrix} \qquad (4\text{-}53)$$

with $\hat{N}_{22}^{(2)^{-1}}$ nonsingular. With respect to the rank of $\tilde{A}_{(2,3)}$

$$\ell_2 = \text{rank of } \tilde{A}_{(2,3)} = \text{Min}(\ell_2, \ell_3) \leq \ell_3 \; . \qquad (4\text{-}54)$$

If, in addition, (4-50) is satisfied, it is necessary from

the rank of $\widetilde{A}_{(1,2)}$ that

$$\ell_1 \leq \ell_2 \,. \tag{4-55}$$

Therefore the assertion is proved for the case of $s < n$ and $v = 3$.

(6) This process can be continued for the remaining subclass of (4). Consider, instead, a general description of $A^{(k)}$ as a transformed matrix from $A^{(k-1)}$ by a nonsingular matrix $\hat{N}^{(k)}$ such that

$$A^{(k)} = \hat{N}^{(k)^{-1}} A^{(k-1)} \hat{N}^{(k)} \tag{4-56}$$

with

$$A^{(k-1)} = \begin{bmatrix} A^{(k-1)}_{k+1\ k+1} & \cdots & A^{(k-1)}_{k-1\ 1} \\ \vdots & & \vdots \\ A^{(k-1)}_{1\ k+1} & \cdots & A^{(k-1)}_{11} \end{bmatrix} \tag{4-57}$$

and

$$\hat{N}^{(k)} = \begin{bmatrix} \hat{N}^{(k)^{-1}}_{k+1\ k+1} & & \\ & \hat{N}^{(k)}_{kk} & \\ & & I_{n-h_1^{(k)}-h_2^{(k)}} \end{bmatrix},$$

all other entries zero, (4-58)

where

(a) each $A^{(k-1)}_{ij}$, $(i,j = 1,2,\ldots,k+1)$, is an $h_i^{(k-1)} \times h_j^{(k-1)}$,

$h_1^{(k-1)} = s$, $h_2^{(k-1)} = n_1$, ... , $h_h^{(k-1)} = n_{k-1}$, and

$h_{h+1}^{(k-1)} = n - s - \sum_{i=1}^{k-1} n_i$ such that

$$\text{rank } A^{(k-1)}_{i\ i-1} = n_{i-1} , \quad \text{for } 2 \le i \le k . \qquad (4\text{-}59)$$

(b) $\hat{N}^{(k)}_{kk}$ and $\hat{N}^{(k)}_{k+1\ k+1}$ are nonsingular matrices which are $n_{k-1} \times n_{k-1}$ and $(n - s - \sum_{i=1}^{k-1} n_i) \times (n - s - \sum_{i=1}^{k-1} n_i)$ such that

$$\hat{N}^{(k)}_{k+1\ k+1}\, A^{(k-1)}_{k+1\ k}\, \hat{N}^{(k)}_{kk} = \begin{bmatrix} [0] & [0] \\ [0] & I_{n_k} \end{bmatrix}, \qquad (4\text{-}60)$$

where the rank of $A^{(k-1)}_{k+1\ k}$ is n_k. If $n_k = n_{k-1} \leq n - s - \sum_{i=1}^{k-1} n_i$, $\hat{N}^{(k)}_{kk}$ can be the unit matrix, from Lemma 4-1. According to the preceding transformations and the decomposition of $A^{(k)}$, it must follow that

$$A^{(k-1)}_{ij} = [0] \qquad \text{for } 1 \leq j \leq i + 2 \leq k + 2$$

$$A^{(k-1)}_{k\ k-1} = \begin{bmatrix} [0] & I_{n_{k-1}} \end{bmatrix} \qquad (4\text{-}61)$$

$$A^{(k-1)}_{i+1\ i} = \begin{bmatrix} [0] & \hat{N}^{(i)\,-1}_{ii} \end{bmatrix} \quad \text{for } i = 2,3,\ldots,k-2\ ,$$

that is, (4-57) becomes

$$A^{(k-1)} = \begin{bmatrix} A^{(k-1)}_{k+1\ k+1} & A^{(k-1)}_{k+1\ k} & & & & & \\ \cdot & \cdot & \begin{bmatrix}[0] & I_{n_k}\end{bmatrix} & & & & \\ \cdot & & & \cdot & & & \\ \cdot & & & & \begin{bmatrix}[0] & \hat{N}^{(k-1)^{-1}}_{k+1\ k-1}\end{bmatrix} & & \\ & & & & \cdot & \cdot & \\ & & & & & \cdot & \begin{bmatrix}[0] & \hat{N}^{(2)^{-1}}_{22}\end{bmatrix} \\ & & & & & & \cdot \\ A^{(k-1)}_{1\ k+1} & \cdot & \cdot & \cdot & \cdot & \cdot & A^{(k-1)}_{11} \end{bmatrix} , \tag{4-62}$$

all other entries zero,

and

$$A^{(k)}$$

$$= \begin{bmatrix}
\hat{N}^{(k)}_{k+1\ k+1} A^{(k-1)}_{k+1\ k+1} \hat{N}^{(k)^{-1}}_{k+1\ k+1} & \hat{N}^{(k)}_{k+1\ k+1} A^{(k-1)}_{k+1\ k} \hat{N}^{(k)}_{kk} & & \\
\hat{N}^{(k)^{-1}}_{kk} A^{(k-1)}_{k\ k+1} \hat{N}^{(k)^{-1}}_{k+1\ k+1} & \hat{N}^{(k)^{-1}}_{kk} A^{(k-1)}_{kk} \hat{N}^{(k)}_{kk} & \left[0\right]\ \hat{N}^{(k)^{-1}}_{kk} & \\
A^{(k-1)}_{k-1\ k+1} \hat{N}^{(k)^{-1}}_{k+1\ k+1} & A^{(k-1)}_{k-1\ k} \hat{N}^{(k)}_{kk} & A^{(k-1)}_{k-1\ k-1} & \left[0\right]\ \hat{N}^{(k-1)^{-1}}_{k-1\ k-1} \\
\vdots & & & \\
A^{(k-1)}_{1\ k+1} \hat{N}^{(k)^{-1}}_{k+1\ k+1} & A^{(k-1)}_{1\ k} \hat{N}^{(k)}_{kk} & A^{(k-1)}_{1\ k-1} & \cdots \quad A^{k-1}_{1\ 1}
\end{bmatrix}$$

$$\begin{bmatrix} & & & \\ & & & \\ & & & \\ & & & \left[0\right]\ \hat{N}^{(2)^{-1}}_{22} \end{bmatrix}$$

all other entries zero.

(4-63)

Repeating the transformation η times, either the rank of $A^{(\eta-1)}_{\eta+1\ \eta}$ becomes

(a) zero, when it follows that

$$\hat{N}^{(\eta)}_{\eta+1\ \eta+1} A^{(\eta-1)}_{\eta+1} \hat{N}^{(\eta)}_{\eta\eta} = [0] , \qquad (4\text{-}64)$$

or (b) full, i.e., the rank $A^{(\eta-1)}_{\eta+1\ \eta} = n - s - \sum_{i=1}^{\eta-1} n_i$,

when it follows that

$$\hat{N}^{(\eta)}_{\eta+1\ \eta+1} A^{(\eta-1)}_{\eta+1\ \eta} \hat{N}^{(\eta)}_{\eta\ \eta} = \left[[0] \quad I_{n_\eta} \right] . \qquad (4\text{-}65)$$

Define

$$\begin{aligned} v &= \eta + 1 \\ \ell_1 &= n - s - \sum_{i=1}^{\eta-1} n_i \\ \ell_2 &= n_{v-2} \\ &\vdots \\ \ell_{v-1} &= n_1 \\ \ell_v &= s \end{aligned} \qquad (4\text{-}66)$$

$$\hat{N} = \hat{N}^{(1)} \hat{N}^{(2)} \cdots\cdots \hat{N}^{(v-1)} ,$$

where $\ell_1, \ell_2, \ldots \ell_v$ are positive integers and $\hat{N}$ is non-singular. With k = n in (4-63), and

$$A^{(n)} = A^{(v-1)} = \tilde{A} , \tag{4-67}$$

$\tilde{A}$ given by (4-27), it follows that

$$\text{(a)} \quad \tilde{A}_{(1,2)} = \begin{cases} [0] , & \text{from (4-64), or} \\ \begin{bmatrix} [0] & I_{\ell_1} \end{bmatrix} , & \text{from (4-65) ,} \end{cases} \tag{4-68}$$

where, for the second case, from the rank of $\tilde{A}_{(1,2)}$,

$$\ell_1 \leq \ell_2 ;$$

$$\text{(b)} \quad \tilde{A}_{(i,i+1)} = \begin{bmatrix} [0] & \hat{N}^{(v+1-i)^{-1}}_{v+1-i\ v+1-i} \end{bmatrix} , \text{ for } i = 2,3,\ldots,v-1, \tag{4-69}$$

where each $\hat{N}^{(v+1-i)^{-1}}_{v+1-i\ v+1-i}$ is nonsingular;

(c) according to the rank of (4-69),

$$\ell_i \leq \ell_{i+1} , \quad \text{for } i = 2,3,\ldots,v-1 ; \tag{4-70}$$

$$\text{(d)} \quad \tilde{A}_{(i,j)} = [0] , \quad \text{for } 2 \leq i \leq j + 2 \leq v + 2 . \tag{4-71}$$

Thus the lemma is proved.

From the structure of each $\hat{N}^{(i)}$ in (4-58), $\hat{N}$ in (4-66) is

$$\hat{N} = \hat{N}^{(1)}\ \hat{N}^{(2)} \cdots \hat{N}^{(v-1)} = \begin{bmatrix} \hat{N}_{22}^{(1)^{-1}} & \\ & \hat{N}_{1\,1}^{(1)} \end{bmatrix} \begin{bmatrix} \hat{N}_{33}^{(2)^{-1}} & \hat{N}_{22}^{(2)} \\ & I_{\ell_v} \end{bmatrix} \begin{bmatrix} \hat{N}_{44}^{(3)^{-1}} & \hat{N}_{33}^{(3)} \\ & I_{\ell_v+\ell_{v-1}} \end{bmatrix} \cdots$$

$$\cdots \begin{bmatrix} \hat{N}_{vv}^{(v-1)} & \\ & \hat{N}_{v-1\,v-1}^{(v-1)} \\ & & I_{n-\ell_1-\ell_2} \end{bmatrix} \triangleq \begin{bmatrix} \hat{N}_{11} & [0] \\ \hat{N}_{21} & \hat{N}_{22} \end{bmatrix}, \qquad (4\text{-}72)$$

where $\hat{N}_{11}$ and $\hat{N}_{22}$ are (n-s) x (n-s) and s x s respectively.

Proof of Theorem 4-1:

It is sufficient to show the existence of an n x n nonsingular matrix $\tilde{N}$ to similarly transform $\tilde{A}$ as given in Lemma 4-3 such that

$$A = \tilde{N}^{-1} \tilde{A} \tilde{N}, \tag{4-73}$$

where A is given by the statement of the theorem. Then N of the theorem is given by

$$N = \hat{N}\tilde{N} \quad . \tag{4-74}$$

(1) If v = 1, let $\tilde{N} = I_n$.

(2) Assume $v \geq 2$. For a null $\tilde{A}_{(1,2)}$, (4-29), define

$$\tilde{N}^{(1)} = \begin{bmatrix} \tilde{N}_{1\,1}^{(1)} & [0] \\ [0] & I_{n-\ell_1} \end{bmatrix} , \tag{4-75}$$

where $\tilde{N}_{11}^{(1)}$ is ℓ_1 x ℓ_1 nonsingular to provide a Jordan canonical form for $\tilde{N}_{1\,1}^{(1)^{-1}} \tilde{A}_{(1,1)} \tilde{N}_{1\,1}^{(1)}$. For $\tilde{A}_{(1,2)}$

given by (4-30), define

$$\widetilde{N}^{(1)} = \begin{bmatrix} I_{\ell_1} & [0] & \\ \widetilde{N}^{(1)}_{21} & \widetilde{N}^{(1)}_{22} & [0] \\ [0] & & I_{n-\ell_1-\ell_2} \end{bmatrix} , \qquad (4\text{-}76)$$

where $\widetilde{N}^{(1)}_{22} = I_{\ell_2}$ and $\widetilde{N}^{(1)}_{21}$ is an $\ell_2 \times \ell_1$ matrix satisfying

$$\widetilde{A}_{(1,1)} + \widetilde{A}_{(1,2)} \widetilde{N}^{(1)}_{21} = [0] . \qquad (4\text{-}77)$$

This $\widetilde{N}^{(1)}_{21}$ exists from Lemma 4-2, where $\widetilde{A}_{(1,2)}$ is of full rank and $\ell_1 \leq \ell_2$. Define

$$\widetilde{A}^{(1)} = \widetilde{N}^{(1)^{-1}} \widetilde{A} \widetilde{N}^{(1)} \triangleq \begin{bmatrix} \widetilde{A}^{(1)}_{(1,1)} & \widetilde{A}^{(1)}_{(1,2)} & & \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \widetilde{A}^{(1)}_{(v-1,v)} \\ \widetilde{A}^{(1)}_{(v,1)} & \cdot & \cdot & \widetilde{A}^{(1)}_{(v,v)} \end{bmatrix} ,$$

all other entries zero, (4-78)

where (i) each $\widetilde{A}^{(1)}_{(i,j)}$ is an $\ell_i \times \ell_j$ matrix;

(ii) either

(a) for (4-29),

$$\widetilde{A}^{(1)}_{(1,1)} = \widetilde{N}^{(1)^{-1}}_{1\,1} \widetilde{A}_{(1,1)} \widetilde{N}^{(1)}_{1\,1}, \text{ a Jordan canonical form}$$

$$\widetilde{A}^{(1)}_{(1,2)} = [0] \qquad (4\text{-}79)$$

or

(b) for (4-30)

$$\widetilde{A}^{(1)}_{(1,1)} = \widetilde{A}_{(1,1)} + \widetilde{A}_{(1,2)} \widetilde{\widetilde{N}}^{(1)}_{2\,1} = [0] \ ,$$

$$\widetilde{A}^{(1)}_{(1,2)} = \begin{bmatrix} [0] & I_{\ell_1} \end{bmatrix} \ ; \qquad (4\text{-}80)$$

(iii) $$\widetilde{A}^{(1)}_{(i,i+1)} = \widetilde{A}_{(i,i+1)} = \begin{bmatrix} [0] & \hat{N}^{(v+1-i)^{-1}}_{v+1-i\ v+1-i} \end{bmatrix} \ ,$$

$$\text{for} \quad i = 2,3,\ldots,v-1 \ , \qquad (4\text{-}81)$$

as given by (4-27) and (4-69). Therefore the theorem is proved for v = 2 by defining $\widetilde{N} = \widetilde{N}^{(1)}$.

(3) If $v \geq 3$, consider

$$\tilde{N}^{(2)} \triangleq \begin{bmatrix} I_{\ell_1} & & & & \\ & I_{\ell_2} & & & \\ \tilde{N}^{(2)}_{31} & \tilde{N}^{(2)}_{32} & \tilde{N}^{(2)}_{33} & & \\ & & & I_{n-\ell_1-\ell_2-\ell_3} & \end{bmatrix},$$

all other entries zero, (4-82)

where $\tilde{N}^{(2)}_{3i}$ are chosen to satisfy

$$\tilde{A}^{(1)}_{(2,1)} + \tilde{A}^{(1)}_{(2,3)} \tilde{N}^{(2)}_{31} = [0] ,$$

$$\tilde{A}^{(1)}_{(2,2)} + \tilde{A}^{(1)}_{(2,3)} \tilde{N}^{(2)}_{32} = [0] , \quad (4-83)$$

$$\tilde{N}^{(2)}_{33} = \begin{bmatrix} I_{\ell_3-\ell_2} & [0] \\ [0] & \hat{N}^{(v-1)}_{v-1\ v-1} \end{bmatrix}$$

and $\hat{N}^{(v-1)}_{v-1\ v-1}$ is given by (4-81). $\tilde{N}^{(2)}_{21}$ and $\tilde{N}^{(2)}_{32}$ exist from Lemma 4-2. Since $\tilde{N}^{(2)}_{33}$ is nonsingular, $\tilde{N}^{(2)}$ is nonsingular and its inverse is

$$\widetilde{N}^{(2)^{-1}} = \begin{bmatrix} & I_{\ell_1} & & \\ & & I_{\ell_2} & \\ -\widetilde{N}_{33}^{(2)^{-1}}\widetilde{N}_{31}^{(2)} & -\widetilde{N}_{33}^{(2)^{-1}}\widetilde{N}_{32}^{(2)} & \widetilde{N}_{33}^{(2)^{-1}} & \\ & & & I_{n-\ell_1-\ell_2-\ell_3} \end{bmatrix},$$

all other entries zero. (4-84)

Define

$$\widetilde{N}^{(2)^{-1}}\,\widetilde{A}^{(1)}\,\widetilde{N}^{(2)} \triangleq \widetilde{A}^{(2)} \triangleq \begin{bmatrix} \widetilde{A}_{(1,1)}^{(2)} & \widetilde{A}_{(1,2)}^{(2)} & & & \\ \widetilde{A}_{(2,1)}^{(2)} & \widetilde{A}_{(2,2)}^{(2)} & \widetilde{A}_{(2,3)}^{(2)} & & \\ \vdots & \vdots & \ddots & \ddots & \\ \vdots & \vdots & & \ddots & \widetilde{A}_{(v-1,v)}^{(2)} \\ \widetilde{A}_{(v,1)}^{(2)} & \cdots & \cdots & \cdots & \widetilde{A}_{(v,v)}^{(2)} \end{bmatrix},$$

all other entries zero, (4-85)

where

(i) each $\widetilde{A}_{(i,j)}^{(2)}$ is an $\ell_i \times \ell_j$ matrix,

(ii)

$$\widetilde{A}^{(2)}_{(1,1)} = \widetilde{A}^{(1)}_{(1,1)}$$

$$\widetilde{A}^{(2)}_{(1,2)} = \widetilde{A}^{(1)}_{(1,2)} \qquad (4\text{-}86)$$

$$\widetilde{A}^{(2)}_{(i,i+1)} = \widetilde{A}^{(1)}_{(i,i+1)} = \begin{bmatrix} [0] & \widehat{N}^{(v+1-i)^{-1}}_{v+1-i\ v+1-i} \end{bmatrix}$$

for $i = 4,5,\ldots,v-1$, if $v \geq 5$

by (4-79~81),

(iii)

$$\widetilde{A}^{(2)}_{(2,1)} = \widetilde{A}^{(1)}_{(2,1)} + \widetilde{A}^{(1)}_{(2,3)}\ \widetilde{N}^{(2)}_{31} = [0]$$

$$\widetilde{A}^{(2)}_{(2,2)} = \widetilde{A}^{(1)}_{(2,2)} + \widetilde{A}^{(1)}_{(2,3)}\ \widetilde{N}^{(2)}_{32} = [0] \qquad (4\text{-}87)$$

$$\widetilde{A}^{(2)}_{(2,3)} = \widetilde{A}^{(1)}_{(2,3)}\ \widetilde{N}^{(2)}_{33} = \begin{bmatrix} [0] & I_{\ell_2} \end{bmatrix}$$

from (4-83),

(iv) if $v \geq 4$

$$\widetilde{A}^{(2)}_{(3,4)} = \widetilde{N}^{(2)^{-1}}_{33}\ \widetilde{A}^{(1)}_{(3,4)} = \begin{bmatrix} [0] & \widetilde{N}^{(2)^{-1}}_{33}\ \widehat{N}^{(v-2)^{-1}}_{v-2\ v-2} \end{bmatrix}, \qquad (4\text{-}88)$$

where $\widehat{N}^{(v-2)}_{v-2\ v-2}$ is given by (4-69). The theorem is proved for the case of $v = 3$ by defining $\widetilde{N} = \widetilde{N}^{(1)}\ \widetilde{N}^{(2)}$.

(4) For $v \geq 4$, this process can be repeated $(v-1)$ times. Generally, defining

$$\tilde{N}^{(i+1)^{-1}} \tilde{A}^{(i)} \tilde{N}^{(i+1)} \triangleq \tilde{A}^{(i+1)} \triangleq \begin{bmatrix} \tilde{A}^{(i+1)}_{(1,1)} & \cdots & \tilde{A}^{(i+1)}_{(1,i+1)} \\ \vdots & & \vdots \\ \tilde{A}^{(i+1}_{(i+1,1)} & \cdots & \tilde{A}^{(i+1)}_{(i+1,i+1)} \end{bmatrix}, \tag{4-89}$$

where

(i) $\tilde{A}^{(i)}$ has a structure after (i-1) reductions, such that

$$\tilde{A}^{(i)} = \begin{bmatrix} \tilde{A}^{(1)}_{(1,1)} & \tilde{A}^{(1)}_{(1,2)} & [0] & \cdots & [0] \\ [0] & [0] & \begin{bmatrix} [0] & I_{\ell_2} \end{bmatrix} & & \vdots \\ \vdots & & & \ddots & \\ [0] & \cdots & [0] & \begin{bmatrix} [0] & I_{\ell_i} \end{bmatrix} & \\ \tilde{A}^{(i)}_{(i+1,1)} & \cdots & & \tilde{A}^{(i)}_{(i+1,i+2)} & \cdots \\ \vdots & & & & [0] \\ & & & & \tilde{A}^{(i)}_{(v-1,v)} \\ \tilde{A}^{(i)}_{(v,1)} & \cdots & & \cdots & \tilde{A}^{(i)}_{(v,v)} \end{bmatrix} \tag{4-90}$$

where

$$\widetilde{A}^{(i)}_{(i+1,i+2)} = \left[[0] \quad \widetilde{N}^{(i)\,-1}_{i+1\ i+1} \widehat{N}^{(v-i)\,-1}_{v-i\ v-i} \right],$$

$$(\widetilde{N}^{(i)\,-1}_{i+1\ i+1} \widehat{N}^{(v-i)\,-1}_{v-i\ v-i} \text{ is } \ell_{i+1} \times \ell_{i+1} \text{ nonsingular}),$$

(4-91)

$$\widetilde{A}^{(i)}_{(i,i+1)} = \left[[0] \quad \widehat{N}^{(v-j)\,-1}_{v-j\ v-j} \right] \text{ for } j = i+1, i+2, \ldots, v-1 ,$$

(ii) $\widetilde{N}^{(i+1)}$ is such that

$$\widetilde{N}^{(i+1)} = \begin{bmatrix} I_{\ell_1} & & & & & & \\ & I_{\ell_2} & & & & & \\ & & \ddots & & & & \\ & & & I_{\ell_i} & & & \\ \widetilde{N}^{(i+1)}_{i+2\ 1} & & \cdots & & \widetilde{N}^{(i+1)}_{i+2\ i+2} & & \\ & & & & & I_{\ell_{i+2}} & \\ & & & & & & I_{\ell_v} \end{bmatrix},$$

all other entries zero, (4-92)

where each $\widetilde{N}^{(i+1)}_{i+2\ j}$ satisfies

$$\text{(a)}\quad \widetilde{A}^{(i)}_{(i+1,j)} + \widetilde{A}^{(i)}_{(i+1,i+2)}\widetilde{N}^{(i+1)}_{i+2\ j} = [0] \quad \text{for } j = 1,2,\ldots,i+1, \tag{4-93}$$

where existence is by Lemmas 4-2, and by using (4-91),

(b)

$$\widetilde{N}^{(i+1)}_{i+2\ i+2} = \begin{bmatrix} I_{\ell_{i+2}-\ell_{i+1}} & [0] \\ [0] & \widehat{N}^{(v-i)}_{v-i\ v-i}\widetilde{N}^{(i)}_{i+1\ i+1} \end{bmatrix}, \tag{4-94}$$

Since (4-94) is nonsingular, $\widetilde{N}^{(i+1)}$ is nonsingular. For the last transformation of the sequence, i = v - 1 and defining

$$\widetilde{N} \triangleq \widetilde{N}^{(1)}\ \widetilde{N}^{(2)} \ldots .\ \widetilde{N}^{(v-1)}, \tag{4-95}$$

which is nonsingular, Theorem 4-1 is proved.

From the structure of each $\widetilde{N}^{(i)}$, (i = 1,2,...,v-1), (4-92), it follows that

$$\tilde{N} = \begin{bmatrix} \tilde{N}_{11} & [0] \\ \tilde{N}_{21} & \tilde{N}_{22} \end{bmatrix}, \qquad (4\text{-}96)$$

where $\tilde{N}_{11}$ and $\tilde{N}_{22}$ are (n-s) x (n-s) and s x s respectively. From (4-42) and (4-89),

$$N = \hat{N}\,\tilde{N} = \begin{bmatrix} N_{11} & [0] \\ N_{21} & N_{22} \end{bmatrix}, \qquad (4\text{-}97)$$

where N_{11} and N_{22} are (n-s) x (n-s) and s x s respectively and N_{22} is nonsingular because both $\hat{N}_{22}$ and $\tilde{N}_{22}$ are nonsingular.

Theorem 4-2:

Let $\hat{A}$ and $\hat{B}$ be n x n and n x m matrices with rank $\hat{B}$ = r and $0 < r \leq m \leq n$. There exist nonsingular matrices N and M which are n x n and m x m respectively such that

$$N^{-1}\,\hat{A}\,N = A \qquad (4\text{-}98)$$

and

$$N^{-1}\hat{B}M = \begin{bmatrix} [0] & [0] \\ [0] & I_r \end{bmatrix} \triangleq B \ , \qquad (4\text{-}99)$$

where A is given by Theorem 4-1, (4-9), with s = r.

Proof: First consider nonsingular matrices $N^{(1)}$ and $M^{(1)}$, n x n and m x m, such that

$$N^{(1)^{-1}}\hat{B}M^{(1)} = \begin{bmatrix} [0] & [0] \\ [0] & I_r \end{bmatrix} \ , \qquad (4\text{-}100)$$

by Lemma 4-1. Define

$$A \triangleq N^{(1)^{-1}}\hat{A}N^{(1)} \ . \qquad (4\text{-}101)$$

Then from Theorem 4-1, there exists a nonsingular matrix N such that

$$A = N^{-1}AN \ , \qquad (4\text{-}102)$$

where the structure of N is given in (4-97). Also define

$$M^{(2)} \triangleq \begin{bmatrix} I_{m-r} & [0] \\ [0] & N_{22} \end{bmatrix} , \qquad (4\text{-}103)$$

where N_{22} is defined in (4-88). If

$$\begin{aligned} N &= N^{(1)} N \\ M &= M^{(1)} M^{(2)} , \end{aligned} \qquad (4\text{-}104)$$

then (4-98) and (4-99) follow.

4.3 Development of a New Canonical Form for Linear Systems

4.3.1 Canonical Form for Linear Systems

For the class of linear systems given by

$$\dot{\underline{Y}} = \hat{A}\, \underline{Y} + \hat{B}\, \underline{V} , \qquad (4\text{-}105)$$

such that

(a) $\hat{A}$ and $\hat{B}$ are n x n and n x m respectively with $0 < m \leq n$,

(b) rank $\hat{B}$ = r with $0 < r \leq m$, a new canonical form is suggested by Theorems 4-1 and 4-2.

Precisely, the new canonical form of (4-105) is given by

$$\dot{\underline{X}} = A\,\underline{X} + B\,\underline{U}$$

$$= \begin{bmatrix} A_{(1,1)} & A_{(1,2)} & & & & \\ & & A_{(2,3)} & & & \\ & & & \ddots & & \\ & & & & \ddots & \\ & & & & & A_{(v-1,v)} \\ A_{(1,v)} & \cdot & \cdot & \cdot & \cdot & A_{(v,v)} \end{bmatrix} \underline{X} + \begin{bmatrix} & & \\ & & \\ & & \\ & & \\ & & \\ & I_r & \end{bmatrix} \underline{U} \,,$$

all other entries zero, (4-106)

where

$$\underline{Y} = N\,\underline{X} \qquad A = N^{-1}\,\hat{A}\,N$$

$$\underline{V} = M\,\underline{U}\,, \quad \text{or equivalently} \qquad B = N^{-1}\,\hat{B}\,M\,. \tag{4-107}$$

The existence of N and M are guaranteed by Theorem 4-2.

4.3.2 Controllability of the System Determined from the Canonical Form

From Theorem 2-9, the controllability of (4-105) can be determined by the rank of

$$\hat{G} = [\hat{B},\ \hat{A}\,\hat{B},\ \ldots,\ \hat{A}^{n-1}\,\hat{B}]\,. \tag{4-108}$$

Using the nonsingular M in (4-107), define an mn x mn matrix such that

$$\hat{M} = \begin{bmatrix} M & & & \\ & M & & \\ & & \ddots & \\ & & & M \end{bmatrix}, \text{ all other entries zero,} \qquad (4\text{-}109)$$

n stages

which is also nonsingular. Define

$$G \triangleq N^{-1} \hat{G} \hat{M} = [N^{-1} \hat{B} M, N^{-1} \hat{A} \hat{B} M, \ldots, N^{-1} \hat{A}^{n-1} \hat{B} M]$$
$$= [B, A B, \ldots, A^{n-1} B] \qquad (4\text{-}110)$$

from (4-107), by applying the technique used in (3-12). The rank of $\hat{G}$ is equal to that of G because N^{-1} and M are nonsingular. Thus the following theorem is proved.

Theorem 4-3:

The given system (4-105) is completely controllable if and only if its canonical form (4-106) is completely controllable.

The alternate definition of controllability can also be considered, using Theorem 3-1. If v = 1, for (4-106), $B = I_r$ and the system is completely controllable by this

theorem. If $v \geq 2$, the controllability of the system depends upon the structures of the $A_{(1,1)}$ and $A_{(1,2)}$ submatrices in the canonical form given in Theorem 4-3.

Corollary 4-1:

Provided that the canonical form has $v \geq 2$, the canonical form is controllable if and only if

$$A_{(1,1)} = [0]$$
$$A_{(1,2)} = \begin{bmatrix} [0] & I_{\ell_1} \end{bmatrix} . \qquad (4\text{-}111)$$

Equivalently the canonical form is uncontrollable if and only if

$$A_{(1,2)} = [0] . \qquad (4\text{-}112)$$

The proof follows directly from Theorems 3-1 and 4-3 and the structure of the canonical form.

4.3.3 Heuristic Explanation of the Canonical Form

Conveniently decompose $\underline{U}$ of (4-106) to

$$\underline{U}_d = [u_1, u_2, \ldots, u_{m-r}]^T$$
$$\underline{U}_e = [u_{m-r+1}, u_{m-r+2}, \ldots, u_m]^T \qquad (4\text{-}113)$$

and define an n x r matrix

$$B_e \triangleq \begin{bmatrix} [0] \\ \\ I_r \end{bmatrix} ; \qquad (4\text{-}114)$$

then the canonical form (4-106) is reduced to

$$\dot{\underline{X}} = A\,\underline{X} + B_e\,\underline{U}_e \,. \qquad (4\text{-}115)$$

Thus only r control variables out of m are effective in multi-input systems, where r is the rank of $\hat{B}$.

Conveniently call each ℓ_i of Theorem 4-1 the i^{th} stage number of the canonical form and decompose the state vector

$$\underline{X} = \begin{bmatrix} \underline{X}_{(1)} \\ \underline{X}_{(2)} \\ \vdots \\ \underline{X}_{(v)} \end{bmatrix} , \qquad (4\text{-}116)$$

where each $\underline{X}_i = [x_{p_i+1}, x_{p_i+2}, \ldots, x_{p_i+\ell_i}]^T$ with $p_i = \sum_{k=1}^{i-1} \ell_k$.

Then the canonical form (4-106) becomes

(i) if the original system is controllable,

$$\begin{aligned}
\dot{\underline{X}}_{(1)} &= A_{(1,2)}\ \underline{X}_{(2)} \\
\dot{\underline{X}}_{(2)} &= A_{(2,3)}\ \underline{X}_{(3)} \\
&\vdots \\
\dot{\underline{X}}_{(v-1)} &= A_{(v-1,v)}\ \underline{X}_{(v)} \\
\dot{\underline{X}}_{(v)} &= A_{(v,1)}\ \underline{X}_{(1)} + \cdots + A_{(v,v)}\ \underline{X}_{(v)} + \underline{U}_e \text{ , or}
\end{aligned} \tag{4-117}$$

(ii) if it is uncontrollable,

$$\begin{aligned}
\dot{\underline{X}}_{(1)} &= A_{(1,1)}\ \underline{X}_{(1)} \\
\dot{\underline{X}}_{(2)} &= A_{(2,3)}\ \underline{X}_{(3)} \\
&\vdots \\
\dot{\underline{X}}_{(v-1)} &= A_{(v-1,v)}\ \underline{X}_{(v)} \\
\dot{\underline{X}}_{(v)} &= A_{(v,1)}\ \underline{X}_{(1)} + \cdots + A_{(v,v)}\ \underline{X}_{(v)} + \underline{U}_e \text{ .}
\end{aligned} \tag{4-118}$$

Provided $v \geq 2$, these equations appear schematically as shown in Figures 4-1 and 4-2.

In (4-116), the stable variables in $\underline{X}_{(1)}$ are uncontrollable because they behave as $e^{A_{(1,1)}(t-t_o)}\ \underline{X}_{(1)0}$ from an intial condition $(\underline{X}_{(1)0},\ t_o)$. However, the other state variables, $\underline{X}_{(2)}, \underline{X}_{(3)}, \ldots, \underline{X}_{(v)}$, can be controlled everywhere by $\underline{U}_e$. To show this, consider an initial states $\underline{X}_o$, and define a control function

$$\underline{U}_e(t) = \underline{u}(t) - A_{(v,1)}\ e^{A_{(1,1)}(t-t_o)}\ \underline{X}_{(1)0} \tag{4-119}$$

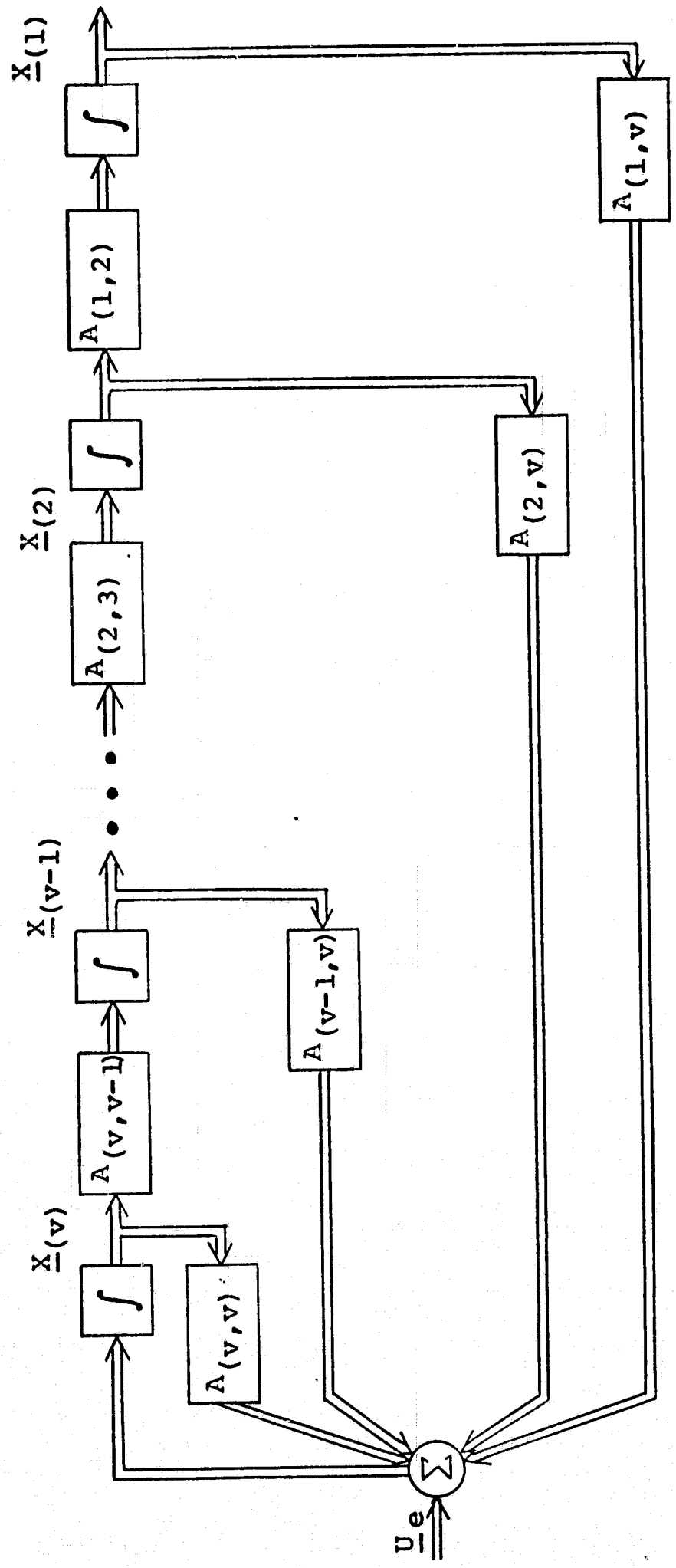


Figure 4-1. Canonical form of a completely controllable linear system (4-117).

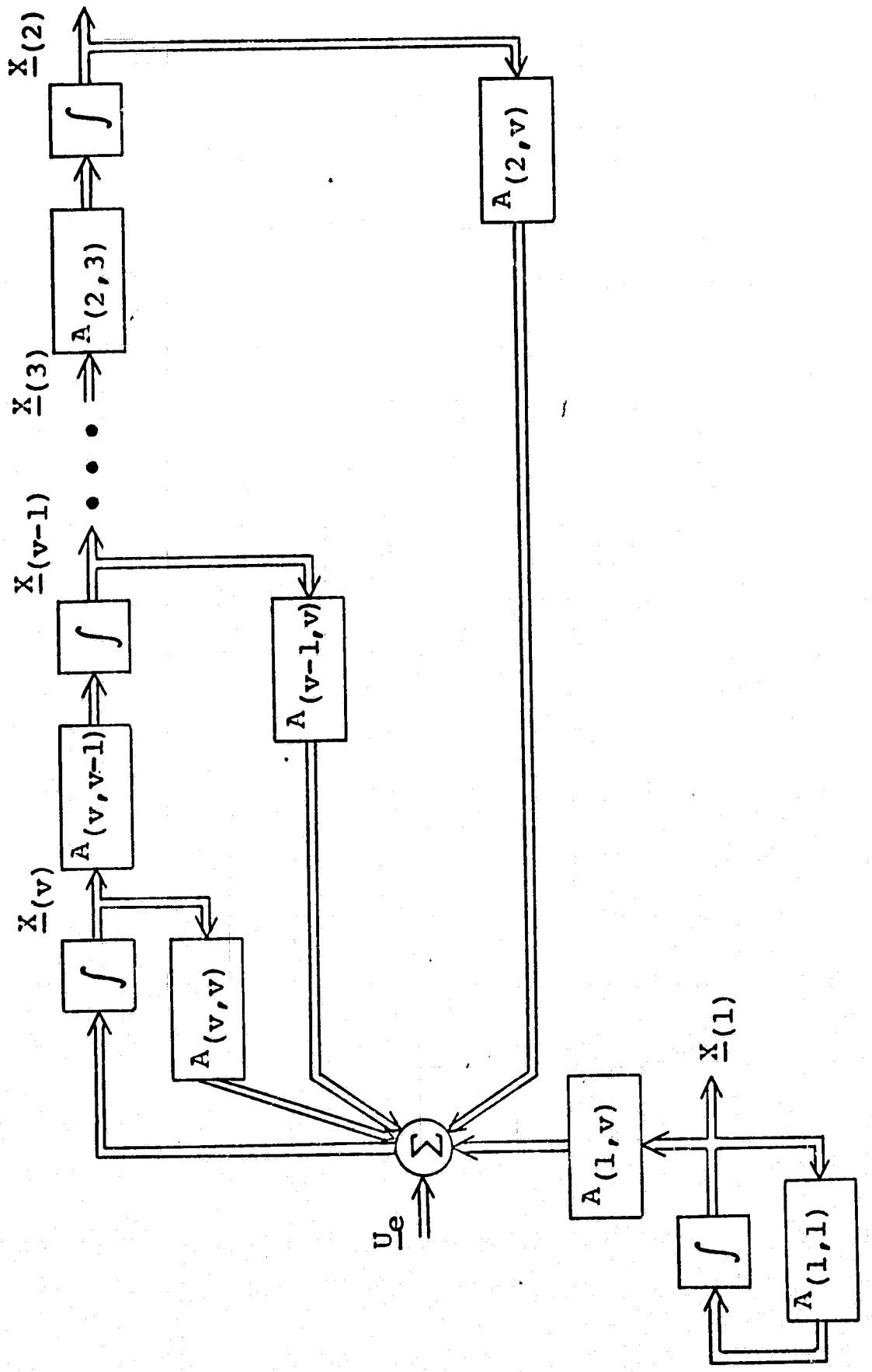


Figure 4-2. Canonical form of an uncontrollable linear system (4-118).

Then, from (4-116), it follows that

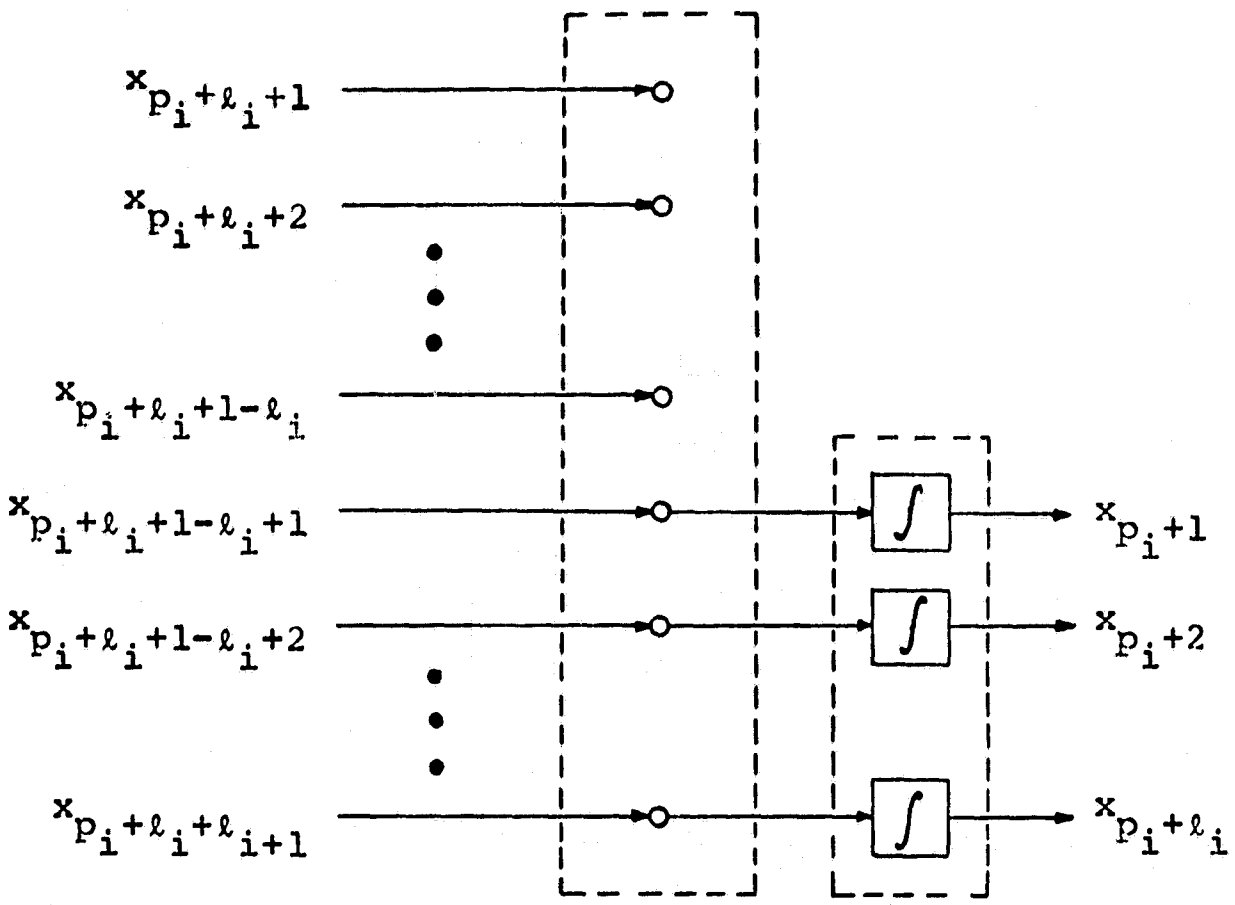
$$\begin{aligned}
\dot{\underline{X}}_{(2)} &= A_{(2,3)}\ \underline{X}_{(3)} \\
\dot{\underline{X}}_{(3)} &= A_{(3,4)}\ \underline{X}_{(4)} \\
&\vdots \\
\dot{\underline{X}}_{(v-1)} &= A_{(v-1,v)}\ \underline{X}_{(v)} \\
\dot{\underline{X}}_{(v)} &= A_{(v,2)}\ \underline{X}_{(2)} + A_{(v,3)}\ \underline{X}_{(3)} + \cdots + A_{(v,v)}\ \underline{X}_{(v)} + \underline{u}(t).
\end{aligned} \tag{4-120}$$

The state variables in $\underline{X}_{(2)}, \underline{X}_{(3)}, \ldots, \underline{X}_{(v)}$ are completely controllable, by (4-117). Conveniently call the state variables in $\underline{X}_{(1)}$ of (4-118) the uncontrollable state variables.

For subsystems of (4-117) or (4-118) given by

$$\dot{\underline{X}}_{(i)} = A_{(i,i+1)}\ \underline{X}_{(i+1)}\ , \tag{4-121}$$

with $A_{(i,i+1)} = \begin{bmatrix} [0] & I_{\ell_i} \end{bmatrix}$, a more detailed representation can be made than is shown in Figure 4-4 and Figure 4-2, i.e., Figure 4-3.



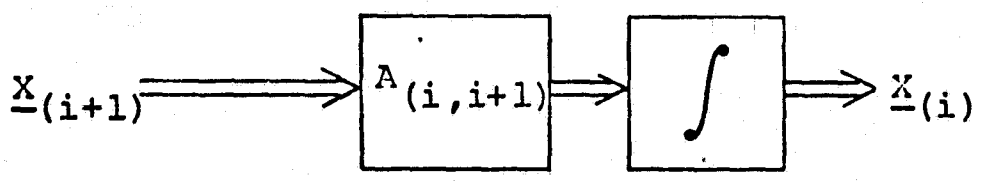


Figure 4-3. Subsystem of the canonical form in (4-117) and (4-118).

That is, the last ℓ_i state variables of $\underline{X}_{(i+1)}$ are integrated to become $\underline{X}_{(i)}$. Let the ordered set $\{\ell_1,\ell_2,\ldots,\ell_v\}$ be the state distribution of the canonical form. Physically, this refers to the numbers of integrators at each stage of the partitioned $\underline{X}_{(i)}$, in sequence.

4.4 Uniqueness of the Structure of the Canonical Form

A given system does not have a unique canonical form, (4-106), but one that is dependent on the choice of different combinations of the M and N matrices in (4-107). The stage distributions for these different forms, however, are unique.

Theorem 4-4:

The stage distribution $\{\ell_1,\ell_2,\ldots,\ell_v\}$ for all canonical forms of a given system is uniquely determined.

Proof: If $\ell_1 = n = r$, $v = 1$ is determined from the construction and the stage distribution is unique. Consider $v \geq 2$, i.e., $0 < r < n$ and two transformations

$$\underline{X} = N^{-1}\,\underline{Y}$$
$$\underline{U} = M^{-1}\,\underline{V} \qquad (4\text{-}122)$$

and

$$\underline{\chi} = \eta^{-1} \underline{Y}$$
$$\underline{u} = m^{-1} \underline{V} \qquad (4\text{-}123)$$

such that they provide the canonical form

$$\dot{\underline{X}} = A \underline{X} + B \underline{U} \qquad (4\text{-}124)$$

with a stage distribution

$$\{\ell_1, \ell_2, \ldots, \ell_v\} \qquad (4\text{-}125)$$

and

$$\dot{\underline{\chi}} = A \underline{\chi} + B \underline{u} \qquad (4\text{-}126)$$

with a stage distribution

$$\{\ell_1', \ell_2', \ldots \ell_\mu'\} , \qquad (4\text{-}127)$$

where $\ell_v = \ell_\mu' = r$. Define

$$\Theta = \eta^{-1} N$$
$$\pi = m^{-1} M \ . \qquad (4\text{-}128)$$

Then from the transitive law of the equivalent relation for the canonical form, (4-126) must be transformed into (4-124) by

$$\underline{\chi} = \Theta \underline{x}$$
$$\underline{u} = \pi \underline{v} \ , \qquad (4\text{-}129)$$

which is equivalent to

$$A = \Theta^{-1} A \Theta$$
$$B = \Theta^{-1} B \pi \quad . \qquad (4\text{-}130)$$

Assume the contrary of the theorem statement, that is, in (4-125) and (4-127), there exists a positive integer $\rho \leq \text{Min}(\nu, \mu)$ such that

$$\ell_{\nu-i} = \ell'_{\mu-i} \quad , \quad \text{for } i = 0,1,\ldots,\rho \ ,$$
$$\ell_{\nu-\rho-1} \neq \ell'_{\mu-\rho-1} \ . \qquad (4\text{-}131)$$

Identify

$$A = \begin{bmatrix} A_{(0,0)} & A_{(0,\alpha)} & & & & \\ & & A_{(\alpha,\alpha+1)} & & & \\ & & & \cdot & & \\ & & & & \cdot & \\ & & & & & A_{(\mu-1,\mu)} \\ A_{(\mu,0)} & \cdot & \cdot & \cdot & & A_{(\mu,\mu)} \end{bmatrix},$$

all other entries zero, (4-132)

and

$$A = \begin{bmatrix} A_{(0,0)} & A_{(0,\alpha)} & & & & \\ & & A_{(\beta,\beta+1)} & & & \\ & & & \cdot & & \\ & & & & \cdot & \\ & & & & & A_{(v-1,v)} \\ A_{(v,0)} & \cdot & \cdot & \cdot & & A_{(v,v)} \end{bmatrix},$$

all other entries zero, (4-133)

where

$$\alpha = \mu - \rho + 1 \tag{4-134}$$

$$\beta = \nu - \rho + 1 \ , \tag{4-135}$$

each $A_{(i,j)}$, $(i,j = 0,\alpha,\alpha+1,\ldots,\mu)$, is $\gamma_i \times \gamma_j$ such that

$$\gamma_o = n - \sum_{i=\alpha}^{\mu} \ell_i' = n - \sum_{i=\beta}^{\mu} \ell_i$$

$$\gamma_i = \ell_i' = \ell_{\nu-\mu+i} \ , \quad \mu-\rho+1 \le i \le \mu \ , \tag{4-136}$$

and A is decomposed in the same partitioning as A.

From (4-121), it must follow that

$$A_{(\mu-i,\mu-i+1)} = A_{(\nu-i,\nu-i+1)} = \begin{bmatrix} [0] & I_{\ell_{\nu-i}} \end{bmatrix}$$

$$\text{for } i = 1,2,\ldots,\rho-1, \tag{4-137}$$

but the rank of $A_{(0,\alpha)}$ differs from the rank of $A_{(0,\alpha)}$.

Let

$$\textcircled{H} \triangleq \begin{bmatrix} \textcircled{H}_{(0,0)} & \textcircled{H}_{(0,\alpha)} & \cdots & \textcircled{H}_{(0,\mu)} \\ \vdots & & & \vdots \\ \textcircled{H}_{(\mu,0)} & \textcircled{H}_{(\mu,\alpha)} & \cdots & \textcircled{H}_{(\mu,\mu)} \end{bmatrix} \tag{4-138}$$

and

$$\Theta^{-1} \triangleq \begin{bmatrix} \tilde{\Theta}_{(0,0)} & \tilde{\Theta}_{(0,\alpha)} & \cdot & \cdot & \cdot & \tilde{\Theta}_{(0,\mu)} \\ \cdot & & & & & \cdot \\ \cdot & & & & & \cdot \\ \cdot & & & & & \cdot \\ \tilde{\Theta}_{(\mu,0)} & \tilde{\Theta}_{(\mu,\alpha)} & \cdot & \cdot & \cdot & \tilde{\Theta}_{(\mu,\mu)} \end{bmatrix}, \tag{4-139}$$

where the decomposition corresponds to that of A (4-132). From (4-130),

$$B = \begin{bmatrix} & \\ & I_r \end{bmatrix} = \Theta^{-1} B \Pi = \begin{bmatrix} \left[[0] \quad \tilde{\Theta}_{(0,\mu)} \right] \Pi \\ \left[[0] \quad \tilde{\Theta}_{(\alpha,\mu)} \right] \Pi \\ \vdots \\ \left[[0] \quad \tilde{\Theta}_{(\mu,\mu)} \right] \Pi \end{bmatrix} . \tag{4-140}$$

Comparing the submatrices in (4-140), it follows that

$$[0] = \left[[0] \quad \tilde{\Theta}_{(i,\mu)} \right] \Pi , \quad \text{for } i = 0,\alpha,\alpha+1,\ldots,\mu-1,$$

$$\left[[0] \quad I_r \right] = \left[[0] \quad \tilde{\Theta}_{(\mu,\mu)} \right] \Pi . \tag{4-141}$$

As Π is nonsingular,

$$[0] = \widetilde{\Theta}_{(i,\mu)} \text{ , for } i = 0,\alpha,\alpha+1,\ldots,\mu-1 \text{ ,} \qquad (4\text{-}142)$$

and $\widetilde{\Theta}_{(\mu,\mu)}$ must be nonsingular. Subsequently, from the inverse relation between (4-138) and (4-139),

$$\Theta_{(i,\mu)} = [0] \text{ , for } i = 0,\alpha,\alpha+1,\ldots,\mu-1,$$
$$\Theta^{-1}_{(\mu,\mu)} = \widetilde{\Theta}_{(\mu,\mu)} \text{ .} \qquad (4\text{-}143)$$

From the first equation of (4-130), it follows that

$$A = \Theta^{-1} A \Theta = \begin{bmatrix} \tilde{\Theta}_{(0,0)} & \cdots & \tilde{\Theta}_{(0,\mu-1)} & [0] \\ \vdots & & \vdots & \vdots \\ \tilde{\Theta}_{(\mu,0)} & \cdots & \tilde{\Theta}_{(\mu,\mu-1)} & \tilde{\Theta}_{(\mu,\mu)} \end{bmatrix}$$

$$\begin{bmatrix} A_{(0,0)} & A_{(0,\mu)} & [0] & \cdots & [0] \\ [0] & [0] & A_{(\mu,\mu+1)} & \ddots & \vdots \\ \vdots & & \ddots & \ddots & [0] \\ [0] & \cdots & [0] & A_{(\mu-1,\mu)} & \\ A_{(\mu,0)} & \cdots & \cdots & \cdots & A_{(\mu,\mu)} \end{bmatrix}$$

$$\begin{bmatrix} \Theta_{(0,0)} & \cdots & \Theta_{(0,\mu-1)} & [0] \\ \vdots & & \vdots & \vdots \\ \Theta_{(\mu,0)} & \cdots & \Theta_{(\mu,\mu-1)} & \Theta_{(\mu,\mu)} \end{bmatrix} . \qquad (4\text{-}144)$$

Comparing the submatrices in the second last columns of (4-144) and (4-134),

$$[0] = \tilde{\Theta}_{(i,\mu-1)} A_{(\mu-1,\mu)} \Theta_{(\mu,\mu)} = \left[[0] \quad \tilde{\Theta}_{(i,\mu-1)} \right] \Theta_{(\mu,\mu)},$$

$$\text{for } i = 0,\alpha,\ldots,\nu-2 \tag{4-145}$$

$$A_{(\nu-1,\nu)} = \left[[0] \quad I_{\ell_{\nu-1}} \right] = \tilde{\Theta}_{(\mu-1,\mu-1)} A_{(\mu-1,\mu)} \Theta_{(\mu,\mu)}$$

$$= \left[[0] \quad \tilde{\Theta}_{(\mu-1,\mu-1)} \right] \Theta_{(\mu,\mu)} .$$

As $\Theta_{(\mu,\mu)}$ is nonsingular, it must be concluded that

$$[0] = \tilde{\Theta}_{(i,\mu-1)}, \quad \text{for } i = 0,\alpha,\alpha+1,\ldots,\nu-2 , \tag{4-146}$$

and from the inverse relation between (4-138) and (4-139),

$$\Theta_{(i,\mu-1)} = [0] , \quad \text{for } i = 0,\alpha,\alpha+1,\ldots,\mu-1 ,$$

$$\Theta_{(\mu-1,\mu-1)} = \tilde{\Theta}^{-1}_{(\mu-1,\mu-1)} . \tag{4-147}$$

Repeating the decomposition of Θ ρ times, it follows that for (4-126) to be satisfied,

$$\Theta_{(i,j)} = \tilde{\Theta}_{(i,j)} = [0] , \quad \text{for } i < j,$$

$$\Theta^{-1}_{(i,i)} = \tilde{\Theta}_{(i,i)} , \quad \text{for } i = 0,\alpha,\alpha+1,\ldots\mu , \tag{4-148}$$

where $\Theta_{(i,i)}$ is nonsingular. From (4-134) and (4-147)

$$A_{(0,\alpha)} = \tilde{\Theta}_{(0,0)} \; \mathcal{A}_{(0,\alpha)} \; \Theta_{(\alpha,\alpha)} \tag{4-149}$$

and the rank of $A_{(0,\alpha)}$ must be equal to that of $\mathcal{A}_{(0,\alpha)}$, i.e.,

$$\ell'_{\mu-\rho-1} = \ell_{\nu-\rho-1} \, , \tag{4-150}$$

as both $\tilde{\Theta}_{(0,0)}$ and $\tilde{\Theta}_{(\alpha,\alpha)}$ are nonsingular. But this contradicts the hypothesis (4-131).

Consequently, the canonical forms (4-124) and (4-126) differ only for elements in submatrices of A and $\mathcal{A}$ with indices (v,1), (v,2), ..., (v,v), and also (1,1) if the system is uncontrollable, i.e.,

$$\mathcal{A} = \begin{bmatrix} \mathcal{A}_{(1,1)} & A_{(1,2)} & & & & \\ & & A_{(2,3)} & & & \\ & & & \ddots & & \\ & & & & A_{(\nu-1,\nu)} & \\ \mathcal{A}_{(\nu,1)} & \cdots & \cdots & \cdots & \cdots & \mathcal{A}_{(\nu,\nu)} \end{bmatrix} ,$$

all other entries zero. (4-151)

The uniqueness of the structure of canonical forms quaranteed for each given system is due to the uniqueness of the stage distribution. The ambiguity in sizes of the submatrices of A and B is thereby avoided in contrast to other suggested canonical forms.

From Theorem 4-4 and (4-146), a corollary follows directly from (4-148).

Corollary 4-2:

The matrix Θ in (4-138) must be such that

$$\Theta = \begin{bmatrix} \Theta_{(1,1)} & & & & \\ & \Theta_{(2,2)} & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \Theta_{(v,v)} \end{bmatrix},$$

all other entries zero. (4-152)

where each $\Theta_{(i,j)}$, $(i,j = 1,2,\ldots,v)$, is $\ell_i \times \ell_j$ and $\Theta_{(i,i)}$ is nonsingular.

4.5 Uniqueness of the Canonical Form

Furthermore, consider a set of canonical forms for a given system with a fixed M (4-107), and observe the variety of matrices A corresponding to various N. This can be done effectively by observing the differences in A and $\mathcal{A}$ of (4-124) and (4-126) under the restriction $M = \mathcal{M}$ in (4-122) and (4-123), or equivalently $\mathcal{T} = I_m$.

Theorem 4-5:

Consider a completely controllable system (4-105) with the stage distribution

$$\{r, r, \ldots, r\} \quad , \tag{4-153}$$

where r is the rank of $\hat{B}$ and assume that a transformation (4-107) is made. Then N is unique.

Proof: Assume two transformations, (4-124) and (4-126), with $\mathcal{M} = M$. Then the proof is based on (4-130), showing that Θ in (4-138) is the unit matrix.

For the case of $v = 1$, since $n = 4$, N^{-1} must be $(\hat{B}\ M)^{-1}$ to satisfy (4-107) with $B = I_r = I_n$, and the uniqueness follows. For $v \geq 2$, consider A and $\mathcal{A}$ as given by (4-124) and (4-126). Then from the controllability

of the system and Corollary 4-1, (4-153) and (4-151), it follows that

$$A = \begin{bmatrix} [0] & I_r & & \\ & \ddots & \ddots & \\ & & [0] & I_r \\ A_{(v,1)} & \cdots & \cdots & A_{(v,v)} \end{bmatrix}, \text{ other entries zero,} \qquad (4\text{-}154)$$

and

$$A = \begin{bmatrix} [0] & I_r & & \\ & \ddots & \ddots & \\ & & [0] & I_r \\ A_{(v,1)} & \cdots & \cdots & A_{(v,v)} \end{bmatrix}, \text{ other entries zero,} \qquad (4\text{-}155)$$

where each submatrix is r x r. Then from (4-130) and $\pi = I_m$,

$$A \Theta = \Theta A \qquad (4\text{-}156)$$

and

$$B = \Theta B , \qquad (4\text{-}157)$$

where Θ is given as (4-152). Then from (4-157)

$$\Theta_{(v,v)} = I_r \,. \tag{4-158}$$

Alternately, from (4-154∿156),

$$A\Theta = \begin{bmatrix} \Theta_{(2,1)} & \Theta_{(2,2)} & [0] & \cdots & [0] \\ \Theta_{(3,1)} & & \Theta_{(3,3)} & \cdots & \\ \vdots & & & \ddots & [0] \\ \Theta_{(v,1)} & \cdots & \cdots & \cdots & \Theta_{(v,v)} \\ \Delta & \Delta & \cdots & \cdots & \Delta \end{bmatrix}, \tag{4-159}$$

$$\Theta A = \begin{bmatrix} [0] & \Theta_{(1,1)} & [0] & \cdots & [0] \\ \vdots & \Theta_{(2,1)} & \Theta_{(2,2)} & \cdots & \\ & \vdots & & \ddots & [0] \\ [0] & \Theta_{(v-1,1)} & \cdots & \cdots & \Theta_{(v-1,v-1)} \\ \Delta & \Delta & \Delta & \cdots & \Delta \end{bmatrix}, \tag{4-160}$$

where Δ refers to entries of no importance to the following.

Comparing submatrices in the first v - 1 rows,

$$\Theta_{(i,j)} = [0] \ , \quad \text{if } i > j \ , \tag{4-161}$$

and $$\Theta_{(i,i)} = \Theta_{(i+1,i+1)} \ , \text{ for } i = 1,2,\ldots,(v-1) \ . \tag{4-162}$$

But from (4-158), (4-162) reduces to

$$\Theta_{(i,i)} = I_r \ , \text{ for } i = 1,2,\ldots,v \ , \tag{4-163}$$

and Θ must be the unit matrix.

Theorem 4-5 establishes the uniqueness of the canonical form for a given completely controllable system, with a stage distribution given by (4-153). If it is possible to select M as a unit matrix, then this canonical form is reduced to the canonical forms discussed in Chapter 3. Sufficient conditions which allow this choice of M are given in the next theorem.

Theorem 4-6:

Consider the system (4-105) with a stage distribution

$$\{\ell_1, r, r, \ldots, r\} \ , \quad 0 < \ell_1 \leq r \ , \tag{4-164}$$

$$r = \text{rank } \hat{B} = m \ , \tag{4-165}$$

where it is assumed that $\ell_1 = r$ if $v = 1$. Then the canonical transformation (4-107) is possible with $M = I_r = I_m$.

Proof: The resulting structure of A is

$$A = \begin{bmatrix} [0] & [[0] \quad I_{\ell_1}] & & & \\ & & I_r & & \\ & & & \ddots & \\ & & & & I_r \\ A_{(v,1)} & A_{(v,2)} & \cdot & \cdot & \cdot \quad A_{(v,v)} \end{bmatrix}, \text{ all other entries zero.} \tag{4-166}$$

If $v = 1$, then $n = r$, $\hat{B}$ is an $n \times n$ nonsingular matrix by assumption, and the proof follows by specifying $N = \hat{B}^{-1}$. Consider the case of $v \geq 2$. Since $\ell_v = r$, $M^{(1)}$ in (4-100) can be chosen as I_r according to Lemma 4-1. For the transformation of A (4-101) into $\tilde{A}$ (4-27), $\hat{N}$ of (4-72) can be chosen as

$$\hat{N} = \begin{bmatrix} \hat{N}_{11} & [0] \\ \hat{N}_{21} & I_r \end{bmatrix}. \tag{4-167}$$

This follows as each $\hat{N}_{ii}^{(i)}$ of (4-58), $i = 1,2,\ldots,v$, can

be chosen as I_r according to Lemma 4-1 due to the rull rank of each $A^{(i-1)}_{(i+1,i)}$, i = 2,3,...v, by assumption (4-164). Then from (4-63) and (4-67),

$$\tilde{A} = A^{(v-1)} = \begin{bmatrix} \tilde{A}_{(1,1)} & \tilde{A}_{(1,2)} & & & \\ \tilde{A}_{(2,1)} & \cdot & I_r & & \\ \vdots & & \cdot & \cdot & \\ \cdot & & & \cdot & I_r \\ \tilde{A}_{(v,1)} & \cdot & \cdot & \cdot & \tilde{A}_{(v,v)} \end{bmatrix},$$

all other entries zero. (4-168)

Then from (4-92) and (4-05), $\tilde{N}$ can be chosen as

$$\tilde{N} = \begin{bmatrix} \tilde{N}_{11} & [0] \\ \tilde{N}_{21} & I_r \end{bmatrix} \tag{4-169}$$

because each $\tilde{N}^{(i)}_{i+1\ i+1}$, i = 1,2,...,v-1, can be chosen as I_r, according to (4-94) with all $\hat{N}^{(i)}_{ii} = I_r$. Subsequently, it follows that

$$N = \hat{N}\tilde{N} = \begin{bmatrix} N_{11} & [0] \\ N_{21} & I_r \end{bmatrix} \tag{4-170}$$

and $M^{(2)}$ of (4-103) is also the unit matrix. Therefore

$$M = M^{(1)} M^{(2)} = I_r . \qquad (4\text{-}171)$$

Combining Theorems 4-5 and 4-6, the following is evident.

Corollary 4-3:

Consider a completely controllable system (4-105) with the stage distribution given by (4-153) and rank $\hat{B} = r = m$. Then there exists a unique canonical form for the system with $M = I_r$.

If the stage distribution of the system is given by (4-164), the canonical form corresponds to Asseo's canonical form, A given (4-166). However, in the reduction to the Asseo's canonical form in the sense of Definition 4-1, it is necessary that:

(i) the stage distribution be given as (4-164), and

(ii) the canonical transformation (4-107) is possible with the unit matrix M.

Consequently, the application of Asseo's unique compact decomposition to his canonical form is limited to sub-class of systems (4-105) for which this decomposition is

applicable and equally unique and compact.

The transformation in Theorem 4-6 corresponds to the canonical transformations used in Definition 3-1. Thus the stage distribution of a single input completely controllable system must be $\{1,1,\ldots,1\}$, and it follows form Theorem 4-6 that the canonical form of the system uniquely exists in the sense of Definition 3-1, i.e.,

$$\dot{\underline{X}} = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdot & \cdot & 0 \\ 0 & 0 & 1 & 0 & \cdot & \cdot & 0 \\ \cdot & & & \cdot & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & \cdot & 0 & & 1 \\ a_1 & \cdot & \cdot & \cdot & a_{n-1} & & a_n \end{bmatrix} \underline{X} + \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 1 \end{bmatrix} u \ . \qquad (4\text{-}172)$$

Consequently, the statement of the uniqueness of this canonical form in Section 3.4 is verified.

4.6 Examples

Example 4-1. Consider the system

$$\begin{aligned}
\dot{y}_1 &= y_1 + 3y_3 - y_4 + v_2 \\
\dot{y}_2 &= 2y_1 + y_2 - y_3 + 4y_4 + v_1 \\
\dot{y}_3 &= -y_1 + y_2 - y_3 + 2y_4 + v_1 \\
\dot{y}_4 &= y_1 + 2y_3 - y_4 + v_2 ,
\end{aligned} \tag{4-173}$$

and from (4-100)

$$\hat{A} = \begin{bmatrix} 1 & 0 & 3 & -1 \\ -2 & 1 & -1 & 4 \\ -1 & 1 & -1 & 2 \\ 1 & 0 & 2 & -1 \end{bmatrix} \quad \text{and} \quad \hat{B} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} . \tag{4-174}$$

For

$$N(1)^{-1} \hat{B} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} , \tag{4-175}$$

it is sufficient to assume

$$N(1) = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} . \tag{4-176}$$

Then

$$A = \tilde{A} = {}_N(1)^{-1}\, \hat{A}\, {}_N(1) = \begin{bmatrix} 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \\ -1 & 1 & 0 & 1 \\ 1 & 0 & 2 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} \tilde{A}_{(1,1)} & \tilde{A}_{(1,2)} \\ \tilde{A}_{(2,1)} & \tilde{A}_{(2,2)} \end{bmatrix}. \qquad (4\text{-}177)$$

Comparing this to (4-78) and (4-80), $\tilde{A}_{(1,1)}$ can be made [0] if

$$N^{(1)} = N = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}. \qquad (4\text{-}178)$$

Then

$$N^{-1}{}_N(1)^{-1}\, A\, {}_N(1)N = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}, \text{ and} \qquad (4\text{-}179)$$

$$N^{-1}\, {}_{N(1)}{}^{-1}\, \hat{B} = B \ , \tag{4-180}$$

which is the desired canonical form. Defining

$$N \triangleq {}_{N(1)}N = \begin{bmatrix} 2 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \tag{4-181}$$

and

$$M = I_2 \ , \tag{4-182}$$

then, by the transformation,

$$\begin{aligned} \underline{Y} &= N\, \underline{X} \\ \underline{V} &= M\, U \ , \end{aligned} \tag{4-183}$$

the canonical form of the system becomes

$$\begin{aligned} \dot{x}_1 &= x_3 \\ \dot{x}_2 &= x_4 \\ \dot{x}_3 &= x_2 + x_4 + u_1 \\ \dot{x}_4 &= x_1 + x_3 + u_2 \ . \end{aligned} \tag{4-184}$$

The system is controllable by Theorem 4-3. As $M = I_2$, (4-184) is the unique canonical form of (4-173), from Corollary 4-3.

Example 4-2. Consider

$$\begin{aligned}
\dot{y}_1 &= y_1 + 3y_2 - 3y_3 - y_4 + v_2 \\
\dot{y}_2 &= -2y_1 + 4y_4 + v_1 \\
\dot{y}_3 &= -y_1 + 2y_4 + v_1 \\
\dot{y}_4 &= y_1 + 2y_2 - 2y_3 - y_4 + v_2 \,.
\end{aligned} \tag{4-185}$$

With the same transformation of Example 4-1,

$$\begin{aligned}
\dot{x}_1 &= x_2 \\
\dot{x}_2 &= x_4 \\
\dot{x}_3 &= x_4 + u_1 \\
\dot{x}_4 &= x_1 + x_2 + u_2 \,,
\end{aligned} \tag{4-186}$$

which is controllable. This is not a unique canonical form, however. Selecting an alternate N for (4-178) such that

$$N = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} , \tag{4-187}$$

(4-185) becomes

$$\begin{aligned}
\dot{x}_1 &= x_2 \\
\dot{x}_2 &= x_4 \\
\dot{x}_3 &= -x_2 + x_4 + u_1 \\
\dot{x}_4 &= x_1 + x_2 + u_2 \ ,
\end{aligned} \tag{4-188}$$

where the first two equations of (4-188) are identical to those of (4-186) as determined by the structure of the canonical form.

Example 4-3. Consider

$$\begin{aligned}
\dot{y}_1 &= -y_1 + y_2 + y_4 + v_2 \\
\dot{y}_2 &= -y_1 - y_2 + y_3 + 2y_4 + v_1 \\
\dot{y}_3 &= -y_2 + y_3 + v_1 \\
\dot{y}_4 &= y_2 + v_2 \ .
\end{aligned} \tag{4-189}$$

With the same transformation as Example 4-1, i.e., (4-181) and (4-182),

$$\begin{aligned} \dot{x}_1 &= -x_1 \\ \dot{x}_2 &= x_4 \\ \dot{x}_3 &= -x_2 + u_1 \\ \dot{x}_4 &= x_1 + x_2 + x_3 + u_2 \end{aligned} \qquad (4\text{-}190)$$

which is uncontrollable as x_1 is isolated.

The diagrams of these example canonical systems are shown in Figures 4-4∿6.

4.7 Application of the Canonical Form to General Systems

Consider a class of systems given by (4-1) which can be expressed by

$$\hat{\underline{F}}(\underline{Y}) \triangleq \hat{A}\,\underline{Y} + \hat{\underline{F}}(\underline{Y}) \ , \qquad (4\text{-}191)$$

where $\hat{A}$ is an n x n matrix such that $\hat{A}\ \underline{Y}$ describes the first degree homogeneous function of $\underline{Y}$ in $\hat{\underline{F}}(\underline{Y})$, with $\hat{\underline{F}}(\underline{Y})$ the remainder. A canonical form for this class of systems, using the development in this chapter, is

$$\begin{aligned} \dot{\underline{X}} &= A\,\underline{X} + \underline{F}(\underline{X}) + B\,\underline{U} \\ &\triangleq \underline{F}(\underline{X}) + B\,\underline{U} \ , \end{aligned} \qquad (4\text{-}192)$$

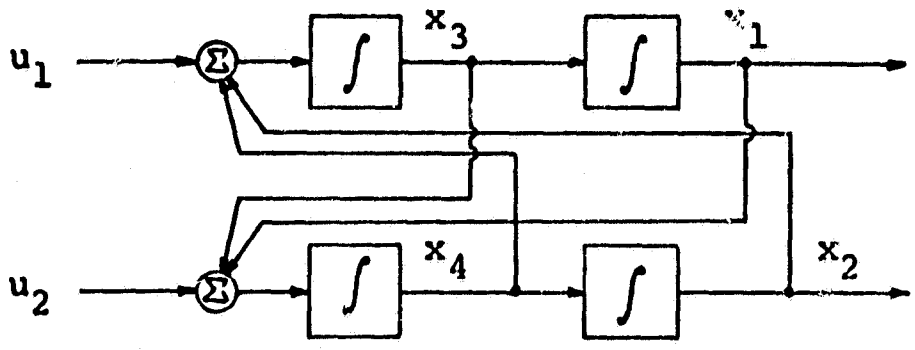


Figure 4-4. Canonical form of Example 4-1, (4-184).

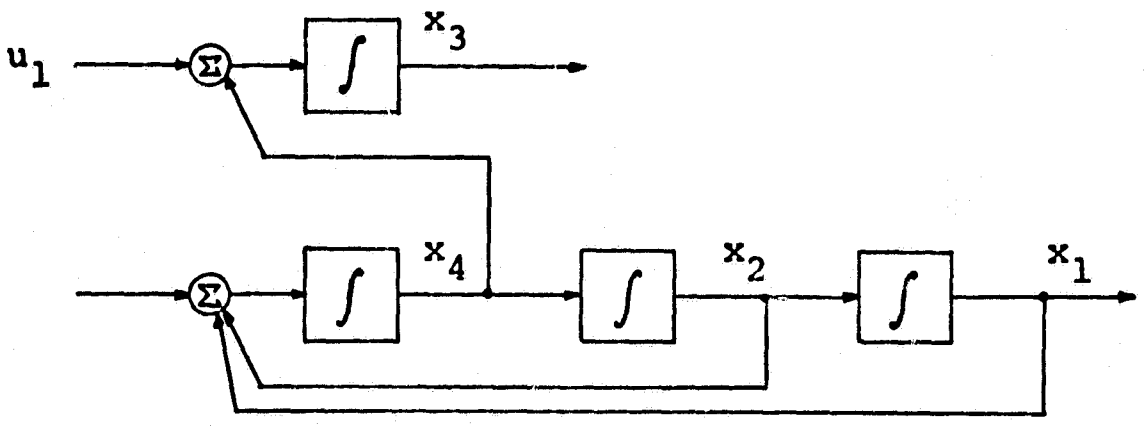


Figure 4-5. Canonical form of Example 4-2, (4-186).

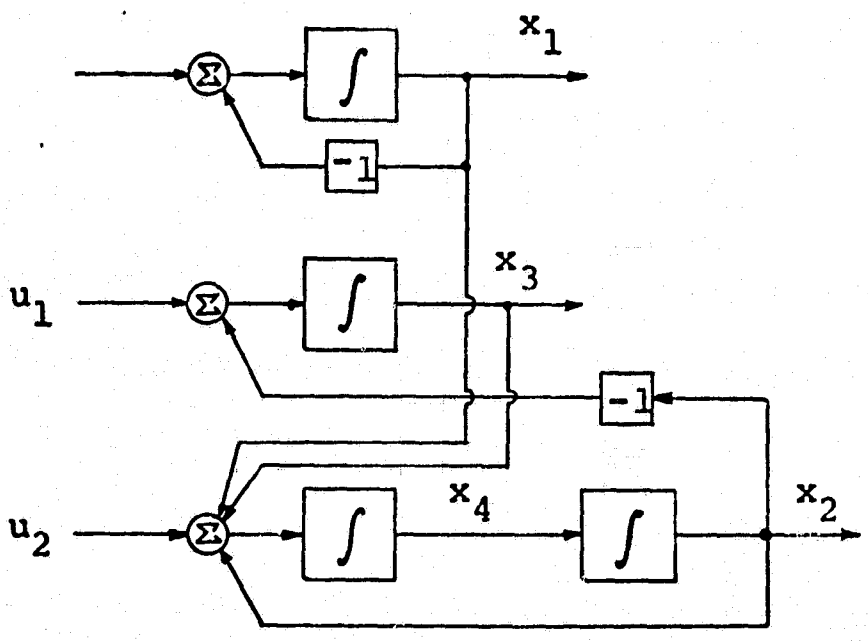


Figure 4-6. Canonical form of Example 4-3, (4-190).

where A and B are given by (4-106). For the canonical transformation (4-107),

$$A = N^{-1} \hat{A} N$$

$$B = N^{-1} \hat{B} M \qquad (4\text{-}193)$$

$$\underline{F}(\underline{X}) = N^{-1} \hat{\underline{F}}(N\underline{X}) .$$

The characteristics of the canonical form discussed in Sections 4.4 and 4.5 then exist for the linear part of (4-192). Furthermore, if $\hat{\underline{F}}(\underline{Y})$ is of class C_2, then $\underline{F}(\underline{X})$ in (4-193) is also of class C_2 (according to Theorem 2-1), and the uniqueness of solutions for the canonical form (4-192) is preserved according to Theorem 2-6.

Finally, a solution for the canonical form can be characterized as follows.

Lemma 4-4:

For the system (4-1), arbitrarily assume an initial condition $(\underline{Y}_o, t_o)$, an admissible control function $\underline{V}(t)$, and a solution $\hat{\underline{\Phi}}_{\underline{V}}(t; \underline{Y}_o, t_o)$. Then the solution of the canonical form (4-185) with the initial condition $(N^{-1}\underline{Y}_o, t_o)$ and the control function $\underline{U}(t) \triangleq M^{-1} \underline{V}(t)$ is

$$\underline{\Phi}_{\underline{U}}(t;\ N^{-1}\underline{Y}_o,\ t_o) = N^{-1}\hat{\underline{\Phi}}_{\underline{V}}(t;\ \underline{Y}_o,\ t_o)\ . \qquad (4\text{-}194)$$

Proof: By the hypothesis, $\hat{\underline{\Phi}}_{\underline{V}}(t;\ \underline{Y}_o,\ t_o)$ satisfies the characteristics of the solution given in Chapter 2.

From (2-26),

$$\underline{\Phi}_{\underline{U}}(t_o;\ N^{-1}\underline{Y}_o,\ t_o) = N^{-1}\ \hat{\underline{\Phi}}_{\underline{V}}(t;\ \underline{Y}_o,\ t_o) = N^{-1}\underline{Y}_o = \underline{X}_o\ . \qquad (4\text{-}195)$$

From (2-27),

$$\frac{d}{dt}\{NN^{-1}\ \hat{\underline{\Phi}}_{\underline{V}}(t;\ \underline{Y}_o,\ t_o)\} = \underline{F}(NN^{-1}\ \hat{\underline{\Phi}}_{\underline{V}}(t;\ \underline{Y}_o,\ t_o) + B\ MM^{-1}\ \underline{V}(t)\ , \qquad (4\text{-}196)$$

which reduces to

$$\frac{d}{dt}\{N^{-1}\ \hat{\underline{\Phi}}_{\underline{V}}(t;\ \underline{Y}_o,\ t_o)\} = N^{-1}\ \underline{F}(NN^{-1}\ \hat{\underline{\Phi}}_{\underline{V}}(t;\ \underline{Y}_o,\ t_o) + N^{-1}\ B\ MM^{-1}\ \underline{V}(t) = N^{-1}\underline{F}(N\hat{\underline{\Phi}}_{\underline{V}}(t;\ \underline{Y}_o,\ t_o) + B\ \underline{U}(t)\ . \qquad (4\text{-}197)$$

Therefore,

$$\frac{d}{dt}\{\underline{\Phi}_{\underline{U}}(t;\ N^{-1}\underline{Y}_o,\ t_o)\} = \underline{F}(\underline{\Phi}_{\underline{U}}(t;\ N^{-1}\underline{Y}_o\ t_o) + B\ \underline{U}(t)\ . \tag{4-199}$$

From (2-28),

$$\underline{\Phi}_{\underline{U}}(t;\ N^{-1}\underline{Y}_o,\ t_o) = N^{-1}\hat{\underline{\Phi}}_{\underline{V}}(t;\ \underline{Y}_o,\ t_o)$$

$$= N^{-1}\hat{\underline{\Phi}}_{\underline{V}}(t,\ N^{-1}N\hat{\underline{\Phi}}_{\underline{V}}(t_1;\ \underline{Y}_o,\ t_o),\ t_1)$$

$$= N^{-1}\hat{\underline{\Phi}}_{\underline{V}}(t,\ N^{-1}\underline{\Phi}_{\underline{U}}(t_1;\ \underline{X}_o,\ t_o),\ t_1)$$

$$= \underline{\Phi}_{\underline{U}}(t;\ \underline{\Phi}_{\underline{U}}(t_1;\ \underline{X}_o,\ t_o),\ t_1), \text{ for all } t \geq t_1 \geq t_o\ . \tag{4-200}$$

From (4-195), (4-199), and (4-200), the function $\underline{\Phi}_{\underline{U}}(t;N^{-1}\underline{Y}_o,\ t_o)$ satisfies the characteristics of the solution of (4-192); therefore the assertion is justified.

Chapter 5

THE OPTIMAL FEEDBACK CONTROL LAW AND THE INVERSE PROBLEM OF THE OPTIMAL REGULATOR

The purpose of this chapter is to review work that has been done on the inverse problem of the optimal regulator. Initially, however, the problem of optimal control as explained in Chapter 1 is mathematically restated. A theoretical background for the problem of the optimal regulator is given in Section 5.2.2 based on the principle of optimality and Caratheodory's lemma. In Section 5.3 studies of the inverse problem by Kalman, Suga and Thau are reviewed. Finally comments about these studies are given.

5.1 Formulation of the Optimal Control Problem and the Inverse Problem

Initial and final conditions for a system of objects are generally defined as manifolds in $R^n \times R^1$. It is convenient to call them starting manifolds M_s and terminating manifolds M_t, and the space $R^n \times R^1$ a motion space. The magnitudes of the control variables and the state variables may be restricted during control action to subdomains of $R^m \times R^1$ and $R^n \times R^1$ for practical reasons, e.g., structual design limitations. Call these admissible domains

the available control region, A.C.R., and the available state region, A.S.R.

Definition 5-1: Suitable Control.

An admissible control function $\underline{V}(t)$ defined on $[t_o, t_1] \in R^1$ is said to be suitable for the problem specifications, or simply a suitable control, if it remains in the A.C.R. and provides a solution to (2-21) such that

$$(\hat{\underline{\Phi}}_{\underline{V}}(t_1; \underline{Y}_o, t_o), t_1) \in M_t \qquad (5\text{-}1)$$

which remains in A.S.R. When the time interval is given by $[t_o, \infty)$, (5-1) can be restated: for each $\varepsilon > 0$, there exists a $T > 0$ satisfying

$$\inf_{\underline{Y}_1 \in M_t} \{||\hat{\underline{\Phi}}_{\underline{V}}(\tau; \underline{Y}_o, t_o) - \underline{Y}_1||\} < \varepsilon \ , \quad \text{for all } \tau \geq t_o + T \ . \qquad (5\text{-}2)$$

A performance index for control action is usually given as

$$\tilde{J}[\underline{Y}_o, t_o, \underline{V}(t)] = \tilde{K}[\underline{Y}_o, t_o, \underline{Y}_1, t_1] + \int_{t_o}^{t_1} \tilde{L}(\underline{\Phi}_{\underline{V}}(\tau; \underline{Y}_o, t_o), \underline{V}(\tau), \tau) d\tau \ , \qquad (5\text{-}3)$$

where t_1 is defined as

$$t_1 \triangleq \inf_t [\{t \in R^1 / t \geq t_o \ , \ \inf_{\underline{Y}_1 \in M_t} ||\underline{\Phi}_{\underline{V}}(t; \underline{Y}_o, t_o) - \underline{Y}_1|| = 0\}] \ . \tag{5-4}$$

The function $\tilde{K}[\ ]$, called a terminal cost, is a penalty for the choice of the starting and terminating points on M_s and M_t and is usually assumed to be nonnegative valued on $R^n \times R^1 \times R^n \times R^1$. If $\tilde{K}$ is constant on M_s for each fixed $(\underline{Y}_1, t_1) \in M_t$ or on M_t for each fixed $(\underline{Y}_o, t_o) \in M_s$, or is constant on both M_s and M_t, the penalty function is constant and generally omitted from (5-3). The loss function, $\tilde{L}(\underline{Y}, \underline{V}, t)$, can be considered as a penalty for each point in motion space and is generally assumed to be nonnegative valued on $R^n \times R^m \times R^1$. The problem of optimal control for an open loop control function is stated as follows. For a given set of problem specifications, i.e., a system equation, M_s, M_t, A.C.R., A.S.R. and a performance index, find a $(\underline{Y}_o, t_o)_o \in M_s$ and a suitable control function $\underline{V}^o(t)$ to minimize the value of the performance index. The function $\underline{V}^o(t)$ is called the open loop optimal control function.

For the problem of the optimal regulator, there exists a collection of M_s, say $\mathcal{M}_s$, but a unique M_t and optimal control functions are required for each element of $\mathcal{M}_s$. For this problem, an optimal feedback

control law, $\underline{V}^{o}(\underline{Y}, t)$, generally provides the optimal control function. Thus for each $(\underline{Y}_o, t_o) \in M_s$,

$$\underline{V}^{o}(t; \underline{Y}_o, t_o) = \underline{V}^{o}(\hat{\underline{\Phi}}_f(t; \underline{Y}_o, t_o), t) , \qquad (5\text{-}5)$$

where $\hat{\underline{\Phi}}_f(t, \underline{Y}_o, t_o)$ is a free solution of

$$\dot{\underline{Y}} = \hat{\underline{F}}(\underline{Y}, \underline{V}^{o}(\underline{Y}, t), t) . \qquad (5\text{-}6)$$

Assuming suitable feedback control law instead of a performance index (5-3), it is possible to attempt to find all performance indices for which the assumed control law is optimal. This is called the inverse problem of the optimal regulator. Specifically, consider $\mathcal{L}$ and $\mathcal{V}$ to be spaces of all performance indices and all suitable feedback control laws for the given optimal regulator problem. The usual or forward problem of the optimal regulator can be given as a mapping in an optimum sense from $\mathcal{L}$ to $\mathcal{V}$, while the inverse problem is from $\mathcal{V}$ to $\mathcal{L}$. As the space $\mathcal{L}$ is too large for analytical treatment, some additional assumptions usually restrict the objects to a subset of $\mathcal{L}$ and $\mathcal{V}$, e.g., the restriction of $\mathcal{L}$ to a sum of quadratic forms in $\underline{Y}$ and $\underline{V}$. [25]

5.2 Optimal Feedback Control Law

5.2.1 Statement of the Problem

The fundamental characteristics of an optimal feedback control law are considered under the assumptions that

(i) a system is given by

$$\dot{\underline{Y}} = \hat{\underline{F}}(\underline{Y}, \underline{V}, t) \ , \qquad (5\text{-}7)$$

where $\hat{\underline{F}}(\underline{Y}, \underline{V}, t)$ is defined on $R^n \times R^m \times R^1$ and is of class C_2 with respect to all arguments;

(ii) M_t, the final condition of the system, is a smooth manifold in $R^n \times R^1$ and M_s , the set of initial conditions composed of all reasonable points in $R^n \times R^1$;

(iii) A.S.R. and A.C.R. are the entire $R^n \times R^1$ and $R^m \times R^1$ spaces respectively;

(iv) a performance index is given by

$$\tilde{J}[\underline{Y}_o, t_o, \underline{V}(t)] = \tilde{K}[\underline{Y}_1, t_1] + \int_{t_o}^{t_1} \tilde{L}(\underline{Y}, \underline{V}, t)\,dt \ , \qquad (5\text{-}8)$$

where

(iva) the terminal cost $\tilde{K}[\underline{Y}, t]$, considered only for final conditions, is of class C_2 with respect to all arguments, and

(ivb) the loss function $\tilde{L}(\underline{Y}, \underline{V}, t)$ is of class C_1 with respect to all arguments.

Since M_s is a set of starting points, there is no terminal cost with respect to initial condition. It is possible to imbed the terminal cost of (5-8) into the integral. It has been shown [27] that an optimal control function from each $(\underline{Y}_o, t_o) \in M_s$ for the performance index (5-8) must be equal to that for

$$\hat{J}[\underline{Y}_o, t_o, \underline{V}(t)] = \int_{t_o}^{t_1} \hat{L}(\underline{Y}, \underline{V}, t)dt , \qquad (5-9)$$

where

$$\hat{L}(\underline{Y},\underline{V},t) = \tilde{L}(\underline{Y},\underline{V},t) + \{\text{grad } \tilde{K}[\underline{Y},t]\}^T \hat{\underline{F}}(\underline{Y},\underline{V},t) + \frac{\partial \tilde{K}[\underline{Y},t]}{\partial t} . \qquad (5-10)$$

$\hat{L}(\underline{Y}, \underline{V}, t)$ is of class C_1 with respect to all arguments by the assumptions (iva) and (ivb). For convenience, therefore, the following analysis proceeds with (5-9) instead of (5-8).

5.2.2 Fundamental Lemma

The principle of optimality states that any portion of an optimal solution is also an optimal soltuion. [2] Mathematically, let $\underline{V}^o(t)$, $t_o \le t \le t_f$, be an optimal

control function from $(\underline{Y}_o, t_o)$ to $(\underline{Y}_f, t_f)$. Then, if $t_o \leq t_1 \leq t_2 \leq t_f$, the control $\underline{V}^o(t)$ considered on the interval $t_1 \leq t \leq t_2$ is an optimal control function from $(\hat{\underline{\phi}}_{\underline{V}^o}(t_1; \underline{Y}_o, t_o), t_1)$ to $(\hat{\underline{\phi}}_{\underline{V}^o}(t_2; \underline{Y}_o, t_o), t_2)$ with $\hat{\underline{\phi}}_{\underline{V}^o}(t; \underline{Y}_o, t_o)$, $t_1 \leq t \leq t_2$, the corresponding trajectory.

Caratheodory's lemma is a sufficiency statement of optimality.

Lemma 5-1:[5,26] Caratheodory's Lemma.

If there exists a suitable feedback control law $\underline{V}^*(\underline{Y}, t)$ for this problem of the optimal regulator such that for all $(\underline{Y}, t) \in R^n \times R^1$

$$\hat{L}(\underline{Y}, \underline{V}^*(\underline{Y},t), t) = 0$$
$$\hat{L}(\underline{Y}, \underline{V}, t) \geq 0 \quad \text{if} \quad \underline{V} \neq \underline{V}^*(\underline{Y}, t) , \tag{5-11}$$

then the function $\underline{V}^*(\underline{Y}, t)$ is an optimal feedback control law.

Necessarily the corresponding optimal performance index from every initial condition is identically zero.

5.2.3 Heuristic Approach to the Optimal Feedback Control Law[4,5,26,27]

Define a function $\hat{V}^o(\underline{Y}, t)$, called an optimal performance index function, such that, for each $(\underline{Y}_o, t_o) \in M_s$,

$$\hat{V}^o(\underline{Y}_o, t_o) = \inf_{\underline{V}(t;\underline{Y}_o,t_o)} \{\hat{J}[\underline{Y}_o, t_o, \underline{V}(t; \underline{Y}_o, t_o)]\}$$

$$= \hat{J}[\underline{Y}_o, t_o, \underline{V}^o(t; \underline{Y}_o, t_o)] , \qquad (5\text{-}12)$$

where $\underline{V}(t; \underline{Y}_o, t_o)$ is any suitable open loop control function from $(\underline{Y}_o, t_o)$. In the following, it is assumed that $\hat{V}^o(\underline{Y}, t)$ is of class C_2 with respect to all arguments. From the definition of t_1 in (5-4), if $(\underline{Y}_o, t_o) \in M_t$, then

$$\hat{V}^o(\underline{Y}_o, t_o) = 0 . \qquad (5\text{-}13)$$

Consider an arbitrary $(\underline{Y}_o, t_o) \in M_s$ and assume a corresponding optimal control function $\underline{V}^o(t; \underline{Y}_o, t_o)$. Consider also a perturbed control function $\underline{V}_d(t; \underline{Y}_o, t_o)$ from $(\underline{Y}_o, t_o)$ such that for an incremental Δt_o ,

$$\underline{V}_d(\tau; \underline{Y}_o, t_o) \triangleq \underline{V}_\alpha \in R^m \quad \text{for } \tau \in [t_o; t_o + \Delta t_o] \qquad (5\text{-}14)$$

transforms the system condition to $(\underline{Y}_o + \Delta\underline{Y}_o,\ t_o + \Delta t_o)$

and

$$\underline{V}_d(\tau;\ \underline{Y}_o,\ t_o) = \underline{V}^o(\tau;\ \underline{Y}_o + \Delta\underline{Y}_o,\ t_o + \Delta t_o) \quad \text{for } \tau > t_o + \Delta t_o \ . \tag{5-15}$$

The performance index becomes

$$\hat{J}[\underline{Y}_o, t_o, \underline{V}_d(t;\underline{Y}_o,t_o)] = \int_{t_o}^{t_o+\Delta t_o} \hat{L}(\hat{\underline{\phi}}_{\underline{V}_d}(\tau;\underline{Y}_o,t_o), \underline{V}_\alpha, \tau)\, d\tau + \hat{V}^o(\underline{Y}_o+\Delta\underline{Y}_o, t_o+\Delta t_o) \ . \tag{5-16}$$

Since Δt_o is small and $\hat{L}(\underline{Y},\ \underline{V},\ t)$ and $\hat{\underline{F}}(\underline{Y},\ \underline{V},\ t)$ are of class C_1, this integral can be approximated as

$$\hat{L}(\underline{Y}_o,\ \underline{V}_\alpha,\ t_o)\Delta t_o + 0(\Delta t_o) \ , \tag{5-17}$$

and $\hat{V}^o(\underline{Y}_o + \Delta\underline{Y}_o,\ t_o + \Delta t_o)$ as

$$\hat{V}^o(\underline{Y}_o,t_o) + \{[\text{grad}\ \hat{V}^o(\underline{Y}_o,t_o)]^T\ \hat{\underline{F}}(\underline{Y}_o,\underline{V}_\alpha,t_o) + \hat{V}_t{}^o(\underline{Y}_o,t_o)\}\Delta t_o + \sigma(\Delta t_o) \ , \tag{5-18}$$

where

$$\Delta\underline{Y}_o = \hat{\underline{F}}(\underline{Y}_o,\ \underline{V}_d,\ t_o)\Delta t_o \ , \tag{5-19}$$

$0(\Delta t_o)$ and $\sigma(\Delta t_o)$ are higher orders of Δt_o, i.e.,

$$\lim_{\Delta t_o \to 0} \frac{0(\Delta t_o)}{\Delta t_o} = 0$$

$$\lim_{\Delta t_o \to 0} \frac{\sigma(\Delta t_o)}{\Delta t_o} = 0 \ , \qquad (5\text{-}20)$$

and

$$\hat{V}_t{}^o(\underline{Y}, t) \triangleq \frac{\partial \hat{V}^o(\underline{Y}, t)}{\partial t} \ . \qquad (5\text{-}21)$$

From the assumption that $\underline{V}^o(t; \underline{Y}_o, t_o)$ is an optimum control function

$$\hat{V}^o(\underline{Y}_o, t_o) = \hat{J}[\underline{Y}_o, t_o, \underline{V}^o(t; \underline{Y}_o, t_o)] \leq \hat{J}[\underline{Y}_o, t_o, \underline{V}_d(t; \underline{Y}_o, t_o)]$$

$$\text{for } \underline{V}_d \in R^m \ . \qquad (5\text{-}22)$$

But from (5-18∿20),

$$\{\hat{L}(\underline{Y}_o, \underline{V}_\alpha, t_o) + [\text{grad } \hat{V}^o(\underline{Y}_o, t_o)]^T \hat{\underline{F}}(\underline{Y}_o, \underline{V}_\alpha, t_o) + \hat{V}_t{}^o(\underline{Y}_o, t_o)\}\Delta t_o$$

$$+ 0(\Delta t_o) + \sigma(\Delta t_o) \geq 0 \ . \qquad (5\text{-}23)$$

Dividing by Δt_o and taking $\lim_{\Delta t_o \to 0}$,

$$\hat{L}(\underline{Y}_o,\underline{V}_\alpha,t_o) + [\text{grad } \hat{V}^o(\underline{Y}_o,t_o)]^T \hat{\underline{F}}(\underline{Y}_o,\underline{V}_\alpha,t_o) + \hat{V}_t{}^o(\underline{Y}_o,t_o) \geq 0 , \tag{5-24}$$

where the equality is satisfied if

$$\underline{V}_\alpha = \underline{V}^o(t_o;\ \underline{Y}_o,\ t_o) . \tag{5-25}$$

As any point of any motion can be regarded as an initial condition for an optimal control problem, (5-24) is valid for all points in $R^n \times R^1$, according to the principle of optimality. The important point is that an optimal feedback control law $\underline{V}^o(\underline{Y},\ t)$ is a suitable control law which satisfies the equality of (5-24) at every point in $R^n \times R^1$.

The function $\hat{V}^o(\underline{Y},\ t)$ describes the value of the performance index for the optimal control function for each $(\underline{Y},\ t)$. Its time derivative, governed by the system equation, is

$$\dot{\hat{V}}^o(\underline{Y},t) = [\text{grad } \hat{V}^o(\underline{Y},t)]^T \hat{\underline{F}}(\underline{Y},\underline{V},t) + \hat{V}_t{}^o(\underline{Y},t), \tag{5-26}$$

which depends upon $\underline{V}$ at each $(\underline{Y},\ t)$. Thus (5-24) for any $(\underline{Y}_o,\ t_o)$, is generally written as

$$\hat{L}(\underline{Y},\underline{V},t) + \dot{\hat{V}}^o(\underline{Y},t) \geq 0 \quad \text{for } (\underline{Y},t) \in R^n \times R^1 . \tag{5-27}$$

The equality is satisfied when an optimal control is used. The optimal feedback control law $\underline{V}^o(\underline{Y}, t)$ can be described as the suitable feedback control law which satisfies the equality in (5-27) at every $(\underline{Y}, t) \in R^n \times R^1$.

For a calculation of the function $\hat{V}^o(\underline{Y}, t)$, it is convenient to introduce a function called the Hamiltonian of the problem such that

$$\hat{H}(\underline{Y},\underline{\psi},\underline{V},t) \triangleq \underline{\psi}^T \hat{\underline{F}}(\underline{Y},\underline{V},t) + \hat{L}(\underline{Y},\underline{V},t) , \qquad (5\text{-}28)$$

where $\underline{\psi}$ is an arbitrary n-dimensional vector of variables ψ_i. Then from (5-13) and (5-24), $\hat{V}^o(\underline{Y}, t)$ is a solution of a specific partial differential equation such that

$$\min_{\underline{V} \in R^m} \{\hat{H}(\underline{Y}, \frac{\partial \hat{V}(\underline{Y},t)}{\partial \underline{Y}}, \underline{V},t)\} + \hat{V}_t(\underline{Y},t) = 0$$

$$\text{for all } (\underline{Y},t) \in R^n \times R^1 \qquad (5\text{-}29)$$

with the boundary condition

$$\hat{V}(\underline{Y}, t) = 0 \quad \text{for all } (\underline{Y},t) \in M_t. \qquad (5\text{-}30)$$

Identifying $\hat{H}(\underline{Y},\psi,\underline{V},t)$ as a function of $\underline{V}$ at each $(\underline{Y},\psi,t) \in R^n \times R^n \times R^1$, denote $\tilde{\underline{V}}(\underline{Y},\psi,t)$ as a function to provide the absolute minimum value to $\hat{H}(\underline{Y},\underline{\psi},\underline{V},t)$ everywhere.

Then (5-29) can be written

$$\hat{H}(\underline{Y}, \frac{\partial \hat{V}(\underline{Y},t)}{\partial \underline{Y}}, \underline{\tilde{V}}(\underline{Y}, \frac{\partial \hat{V}(\underline{Y},t)}{\partial \underline{Y}}, t)) + \hat{V}_t(\underline{Y},t) = 0 , \quad (5\text{-}31)$$

which is generally called the Hamilton-Jacobi equation of the problem. As the fundamental condition for optimality in dynamic programing, Bellman called (5-29) Bellman's equation and its solution a Bellman function.[1]

Alternately, it is possible to recognize grad$[\hat{V}(\underline{Y},t)]$ as an independent variable in the Hamilton-Jacobi equation. Pontryagin[2] developed a different technique of calculating open loop optimal control functions, in effect, by doing so. This is called his maximum principle (or sometimes the minimum principle). Systematically calculating open loop optimal control functions from various initial conditions and observing their common characteristics, a synthesis of an optimal feedback control law is possible.[2] However, at this point two major difficulties exist for these calculations. From a practical aspect, no general method for solving the Hamilton-Jacobi equation is known and solutions can only be calculated for a few classes of problems with restrictive assumptions. Secondly, from a theoretical aspect, an optimal performance index function $\hat{V}^o(\underline{Y}, t)$ must be a solution to the Hamilton-Jacobi equation but this is not a sufficient condition. Thus a technique

to identify the $\hat{V}^o(\underline{Y}, t)$ among solutions must be developed when more than one solution exists. If the number of solutions to the Hamilton-Jacobi equation is finite and small, $\hat{V}^o(\underline{Y}, t)$ may be identified by comparing each of the solutions. These difficulties are not avoided if the maximum principle is used.

This second difficulty can be avoided, however, if a unique solution can be shown to exist for the Hamilton-Jacobi equation. But this depends upon the specified $\hat{L}(\underline{Y}, \underline{V}, t)$ and $\hat{\underline{F}}(\underline{Y}, \underline{V}, t)$ in the problem.

Definition 5-2: A Normal Hamiltonian.[5]

If the Hamiltonian of a problem is minimized by a unique value of $\underline{V} \in R^m$ at each $(\underline{Y}, \underline{\psi}, t) \in R^n \times R^n \times R^1$, then the Hamiltonian is said to be normal. In this case, the function $\underline{V}^o(\underline{Y}, \underline{\psi}, t)$ which provides the absolute minimum to the Hamiltonian is called the H-minimal control law.

In order to be normal, the Hamiltonian must be a strict convex function of $\underline{V}$, i.e., the matrix

$$\frac{\partial^2 \hat{H}}{\partial \underline{V} \partial \underline{V}} = \begin{bmatrix} \frac{\partial^2 \hat{H}}{\partial v_1 \partial v_1} & \cdots & \frac{\partial^2 \hat{H}}{\partial v_m \partial v_1} \\ \vdots & & \vdots \\ \frac{\partial^2 \hat{H}}{\partial v_1 \partial v_m} & \cdots & \frac{\partial^2 \hat{H}}{\partial v_m \partial v_m} \end{bmatrix} \qquad (5\text{-}32)$$

must be positive definite for every argument.[5] Then the H-minimal control law is calculated from

$$\underline{0} = \left. \frac{\partial \hat{H}(\underline{Y}, \underline{\psi}, \underline{V}, t)}{\partial \underline{V}} \right|_{\underline{V} = \underline{V}^{o}(\underline{Y}, \underline{\psi}, t)}$$

$$= \left\{ \underline{\psi}^{T} \left[\frac{\partial \hat{\underline{F}}(\underline{Y}, \underline{V}, t)}{\partial \underline{V}} \right] + \frac{\partial \hat{L}(\underline{Y}, \underline{V}, t)}{\partial \underline{V}} \right\}_{\underline{V} = \underline{V}_{o}(\underline{Y}, \underline{\psi}, t)} . \qquad (5\text{-}33)$$

Theorem 5-1:[5]

If the Hamiltonian of a problem is normal, then a solution of the Hamilton-Jacobi equation is the optimal performance index function $\hat{V}^{o}(\underline{Y}, t)$.

This theorem is proved by the uniqueness of the solution to the Hamilton-Jacobi equation and Lemma 5-1. The

corresponding optimal feedback control law is given by

$$\underline{V}^{o}(\underline{Y}, \frac{\partial \hat{V}^{o}(\underline{Y},t)}{\partial \underline{Y}}, t) \quad . \tag{5-34}$$

Examples of this theory are given in the literature.[5,6,26]

5.2.4 Miscellaneous Comments

The preceding considerations are given under restrictive assumptions for simplicity of discussion. Studies have been made for fewer restrictions. Based on measure theory, Bridgland[28] generalized the theory with relaxed assumptions on $\hat{\underline{F}}(\underline{Y}, \underline{V}, t)$ and $\hat{L}(\underline{Y}, \underline{V}, t)$ and generalized the integral interval of the performance index to $[t_o, \infty)$. Boltyanskii[29] extended the theory for a continuous $\hat{V}^{o}(\underline{Y}, t)$, (not necessarily of class C_2), but under other conditions. Thus the existence of continuous $\hat{V}^{o}(\underline{Y}, t)$ (not necessarily of C_2) for the time optimal control problem,[2] as mentioned by Pontryagin, is justified.

5.3 Review of Studies on the Inverse Problem

Results reported for the inverse problem are reviewed but with interest directed to those with similar problem assumptions as those made Chapters 6 and 7.

5.3.1 Study of Kalman[25]

Kalman considered an inverse problem with the assumptions that

(i) system is completely controllable and given by

$$\dot{\underline{Y}} = \hat{\underline{A}}\,\underline{Y} + \hat{\underline{B}}\upsilon, \qquad (5\text{-}35)$$

where $\hat{\underline{A}}$ and $\hat{\underline{B}}$ is constant and the control vector υ is one dimensional,

(ii) the control law is time-invariant such that

$$\upsilon(\underline{Y}) = -\hat{\underline{K}}^T\,\underline{Y}\ , \qquad (5\text{-}36)$$

where $\hat{\underline{K}}$ is an n x 1 constant matrix and all real parts of the eigenvalues of $(\hat{\underline{A}} - \hat{\underline{B}}\,\hat{\underline{K}}^T)$ are negative, thus predicting the final desired condition of system to be $\underline{Y} = \underline{0}$,

(iii) the form of the performance index is restricted to

$$\int_0^\infty (\underline{Y}^T\,\hat{\underline{H}}^T\,\hat{\underline{H}}\,\underline{Y} + \upsilon^2)\,dt\ , \qquad (5\text{-}37)$$

where $\hat{\underline{H}}$ is a n_1 x n constant matrix with the rank $n_1 \leq n$ and

$$\text{rank } [\hat{\underline{H}},\ \hat{\underline{A}}\,\hat{\underline{H}},\ \ldots,\ \hat{\underline{A}}^{n-1}\,\hat{\underline{H}}] = n\ , \qquad (5\text{-}38)$$

$\hat{\underline{H}}^T\,\hat{\underline{H}}$ is positive semidefinite by (ii) of Corollary 2-1. In another study,[26] Kalman proved that the optimal performance index function for this problem is given by

$$\hat{V}^o(\underline{Y}) = \underline{Y}^T\,\hat{\underline{P}}\,\underline{Y}\ , \qquad (5\text{-}39)$$

where $\hat{\underline{P}}$ is a symmetric positive semidefinite matrix. Using Theorem 5-1, he showed the corresponding Hamilton-Jacobi equation to be

$$\underline{Y}^T\,\{\hat{\underline{P}}\,\hat{\underline{A}} + \hat{\underline{A}}^T\,\hat{\underline{P}} + \hat{\underline{H}}^T\,\hat{\underline{H}}\}\underline{Y} + 2\,\mathcal{U}\,\hat{\underline{B}}^T\,\hat{\underline{P}} + \mathcal{U}^2 = 0 \qquad (5\text{-}40)$$

and the optimal feedback control law

$$\mathcal{U}^o(\underline{Y}) = -\,\hat{\underline{B}}^T\,\hat{\underline{P}}\,\underline{Y}\ . \qquad (5\text{-}41)$$

Theorem 5-2:[25]

For the completely controllable system (5-35) and the performance index (5-37), the necessary and sufficient condition for the control law (5-36) to be a stable optimal control law is that there exists a positive definite, symmetric matrix $\hat{\underline{P}}$ which uniquely satisfies the algebraic relations

$$\hat{\underline{P}}\ \hat{\underline{B}} = \hat{\underline{K}} \tag{5-42}$$

and

$$-\hat{\underline{P}}\ (\hat{\underline{A}} - \hat{\underline{B}}\ \hat{\underline{K}}^T) - (\hat{\underline{A}} - \hat{\underline{B}}\ \hat{\underline{K}}^T)^T \hat{\underline{P}} = \hat{\underline{H}}^T\ \hat{\underline{H}} + \hat{\underline{K}}\ \hat{\underline{K}}^T \ . \tag{5-43}$$

Restated, if an arbitrary symmetric positive definite matrix $\hat{\underline{P}}$ determined by

$$\hat{\underline{P}}\ \hat{\underline{B}} = \hat{\underline{K}} \tag{5-44}$$

also allows a solution for $\hat{H}$ from (5-43) which also satisfies (5-38), then the resulting performance index (5-37) is optimized by the given control law (5-36).

This study of Kalman was the first published on the inverse problem. The results revealed the positive definiteness of $\hat{V}^o(\underline{Y})$ for the loss function of (5-37) with the condition of (5-38). Although the problem assumptions are relatively simple, others were subsequently encouraged to attempt to generalize them. In Kalman's original paper,[25] the solution to the inverse problem was also discussed from the viewpoint of the frequency domain techniques of synthesizing optimal feedback control systems.

5.3.2 Study of Suga[30]

Suga considered an inverse problem under the assumptions that:

(i) the system is given by

$$\dot{\underline{Y}} = \hat{\underline{A}}(t)\ \underline{Y} + \hat{\underline{B}}(t)\ \underline{V}\ , \qquad (5\text{-}45)$$

where $\hat{\underline{A}}(t)$ and $\hat{\underline{B}}(t)$ have continuous first derivatives and

$$\text{rank}\ \hat{\underline{B}}(t) = r = m \leq n\ \text{(full rank)}\ , \qquad (5\text{-}46)$$

(ii) the control law is given by

$$\underline{V}(\underline{Y},\ t) = -\ \hat{\underline{K}}^{T}(t)\ \underline{Y}\ , \qquad (5\text{-}47)$$

where $\hat{\underline{K}}(t)$ has a continuous first derivative,

(iii) the form of the performance index is restricted to

$$\int_{t_o}^{T} \{\hat{L}(\underline{Y},\ t) + \underline{V}^{T}\ \hat{\underline{R}}(t)\ \underline{V}\}dt\ , \qquad (5\text{-}48)$$

where T is fixed and $\hat{\underline{R}}(t)$ is a given $r \times r$ positive definite symmetric matrix with a continuous time derivative at every $t \in R^{1}$. The Hamiltonian is normal from these

assumptions. Thus, the author stated the following.

Theorem 5-3:

Suppose that $\hat{\underline{K}}(t)$ is specified so that $\hat{\underline{R}}(t)\ \hat{\underline{K}}(t)\ \hat{\underline{B}}(t)$ is a symmetric matrix. Then the performance index is optimized if and only if it is given by

$$\hat{L}(\underline{Y},t) = \underline{Y}^T[\hat{\underline{K}}(t)\ \hat{\underline{R}}(t)\ \hat{\underline{K}}^T(t) - \hat{\underline{A}}^T(t)\ \hat{\underline{P}}(t) - \hat{\underline{P}}(t)\ \hat{\underline{A}}(t) - \hat{\underline{P}}(t)]\underline{Y}$$

$$+ \left[\frac{\partial \hat{\Gamma}(\underline{Y},t)}{\partial \underline{Y}}\right]^T \hat{\underline{A}}(t)\ \underline{Y} + \frac{\partial \hat{\Gamma}(\underline{Y},t)}{\partial t}, \qquad (5\text{-}49)$$

where $\hat{\underline{P}}(t)$ is an n x n symmetric matrix of class C_2 satisfying

$$\hat{\underline{P}}(t)\ \hat{\underline{B}}(t) = \hat{\underline{K}}(t)\ \hat{\underline{R}}(t) \qquad (5\text{-}50)$$

and

$$\hat{\underline{P}}(T) = [0], \qquad (5\text{-}51)$$

and $\hat{\Gamma}(\underline{Y}, t)$ is an arbitrary scalar function of class C_2 in all arguments satisfying

$$\hat{\underline{B}}^T(t)\ \frac{\partial \hat{\Gamma}(\underline{Y},t)}{\partial \underline{Y}} = \underline{0} \qquad (5\text{-}52)$$

and

$$\left.\frac{\partial \hat{\Gamma}(\underline{Y},t)}{\partial \underline{Y}}\right|_{t=T} = \underline{0} \ . \qquad (5\text{-}53)$$

Then the resulting optimal performance index function is

$$\hat{V}^{o}(\underline{Y},t) = \underline{Y}^{T}\,\hat{\underline{P}}(t)\,\underline{Y} - \hat{\Gamma}(\underline{Y},t) + \hat{\Gamma}(\underline{Y}_{1},T) \ , \qquad (5\text{-}54)$$

where $\underline{Y}_1$ is a final state of the system and the last term is constant, say $\gamma(T)$, because of (5-53) and the fixed T.

If a stable control law is a suitable control law which provides asymptotic stability in the large for the synthesized feedback control system relative to the origin, then the following corollary exists.

Corollary 5-1:

Assume that a stable control law is given and $T\to\infty$. Then the theorem is still valid with a change of (5-51) to

$$\lim_{t\to\infty}\ \hat{\underline{P}}(t)\ \hat{\underline{\Phi}}_{\underline{V}}(t;\underline{Y}_{o},t_{o}) = \underline{0} \ , \quad \text{for every } (\underline{Y}_{o},t_{o}) \in R^{n} \times R^{1} \ . \qquad (5\text{-}55)$$

Based on this work, Suga observed the following points for his problem.

(1). The symmetry of $\hat{\underline{R}}(t)\ \hat{\underline{K}}^T(t)\ \hat{\underline{B}}(t)$ is necessary for the inverse problem to be meaningful. In the study of Kalman, this condition is trivial because of the single input.

(2). The additive terms of $\hat{L}(\underline{Y}, t)$ which are associated with the function $\hat{\Gamma}(\underline{Y}, t)$ don't affect the optimal feedback control law. To show this, the value of the optimal performance index for each initial condition $(\underline{Y}_o, t_o) \in R^n \times R^1$ becomes, from (5-54),

$$\hat{V}^o(\underline{Y}_o, t_o) = \underline{Y}_o^T\ \hat{\underline{P}}(t_o)\ \underline{Y}_o - \hat{\Gamma}(\underline{Y}_o, t_o) - \gamma(T) \quad . \quad (5\text{-}56)$$

Only the first term is sensitive to the control function; the last two are independent to $\underline{V}(t)$. If $r = n$, $\hat{B}(t)$ is nonsingular and there can be no $\hat{\Gamma}(\underline{Y}, t)$ because of the restrictions of (5-52). The larger the value of $n - r$, the more flexibility of $\hat{L}(\underline{Y}, t)$ exists through this $\hat{\Gamma}$. Suga expressed this idea as a flexibility of loss functions.

As an example, consider a system given by

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} v \quad , \qquad (5\text{-}57)$$

a feedback control law given by

$$\mathcal{U}(\underline{Y}) = - [1, \sqrt{3}] \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} , \tag{5-58}$$

and a performance index with the form

$$\int_0^\infty \{\hat{L}(\underline{Y}, t) + \mathcal{U}^2\}dt . \tag{5-59}$$

(5-58) is a stable control law and the integral interval of the performance index is $[0,\infty)$; thus Corollary 5-1 can be applied. From (5-51), it follows that

$$\hat{\underline{P}}(t) = \begin{bmatrix} g(t) & 1 \\ 1 & \sqrt{3} \end{bmatrix} , \tag{5-60}$$

where g(t) is undetermined. Subsequently (5-49) becomes

$$\hat{L}(\underline{Y},t) = \underline{Y}^T \begin{bmatrix} 1-\dot{g}(t) & \sqrt{3}-g(t) \\ \sqrt{3}-g(t) & 1 \end{bmatrix} \underline{Y} + \frac{\partial \hat{\Gamma}(\underline{Y},t)}{\partial y_1} y_2 + \frac{\partial \hat{\Gamma}(\underline{Y},t)}{\partial t} \tag{5-61}$$

and (5-54) is

$$\hat{V}^{o}(\underline{Y},t) = \underline{Y}^{T} \begin{bmatrix} g(0) & 1 \\ 1 & \sqrt{3} \end{bmatrix} \underline{Y} - \hat{\Gamma}(\underline{Y},0) - \hat{\Gamma}(0,\infty) \ . \qquad (5\text{-}62)$$

From Corollary 5-1,

$$\lim_{t\to\infty} \begin{bmatrix} g(t) & 1 \\ 1 & \sqrt{3} \end{bmatrix} \begin{bmatrix} \hat{\Phi}_1(t;\underline{Y}_o,t_o) \\ \hat{\Phi}_2(t;\underline{Y}_o,t_o) \end{bmatrix} = \underline{0} \ , \text{ for every } (\underline{Y}_o,t_o) \in R^n \times R^1 \ . \qquad (5\text{-}63)$$

From (5-52) and (5-53),

$$\hat{B}^{T}(t) \frac{\partial \hat{\Gamma}(\underline{Y},t)}{\partial \underline{Y}_1} = \frac{\partial \hat{\Gamma}(\underline{Y},t)}{\partial Y_2} = 0 \ , \quad \text{for all } t \in R^1 \ , \quad (5\text{-}64)$$

and

$$\left.\frac{\partial \hat{\Gamma}(\underline{Y},t)}{\partial \underline{Y}}\right|_{t\to\infty} = \underline{0} \ . \qquad (5\text{-}65)$$

By various choices of $g(t)$ and $\hat{\Gamma}(\underline{Y}, t)$, the following optimized $\hat{L}(\underline{Y}, t)$ were found.

	$g(t)$	$\hat{\Gamma}(t)$	$\hat{L}(\underline{Y},t)$	
(a)	$\sqrt{3}-\frac{\alpha}{2}$	0	$y_1^2+\alpha y_1 y_2+y_2^2$	α: arbitrary real
(b)	$\sqrt{3}-e^{-t}$	0	$(1-e^{-t})y_1^2+2e^{-t}y_1y_2+y_2^2$	
(c)	$\sqrt{3}+t$	0	$y_2^2-2ty_1y_2$	
(d)	$\sqrt{3}-1$	$\frac{y_1^{m+1}}{m+1}$	$(y_1+y_2)^2+y_1^m y_2$	m: positive integer

By generalizing the assumptions of Kalman, Suga discovered the fundamental composition of $\hat{L}(\underline{Y}, t)$ in relation to the dummy function $\hat{\Gamma}(\underline{Y}, t)$ and the symmetry condition of $\hat{\underline{R}}(t)$ $\hat{\underline{K}}(t)$ $\hat{\underline{B}}(t)$. However, as in the demonstrated examples, the calculation does not increase a nonnegative $\hat{L}(\underline{Y}, t)$ on $R^n \times R^1$. From a practical viewpoint for the forward optimal control problem, $\hat{L}(\underline{Y}, t)$ is isusually assumed to be nonnegative as a penalty function with respect to regulating errors. Thus Suga's examples tend to be unrealistic.

5.3.3 Study of Thau[31]

Thau considered an inverse problem such that

(i) a system is time invariant and given by

$$\dot{\underline{Y}} = \hat{\underline{F}}(\underline{Y}) + \hat{\underline{B}}\,\underline{V}\,, \tag{5-66}$$

where $\hat{\underline{F}}(\underline{Y})$ is of class C_2,

(ii) a control law is given by $\underline{V}(\underline{Y})$, a stable control law of class C_2,

(iii) the form of performance indices is restricted to

$$\int_0^\infty \{\hat{L}(\underline{Y}) + \hat{R}(\underline{V})\}dt\,, \tag{5-67}$$

where $\hat{L}(\underline{Y})$ and $\hat{R}(\underline{V})$ are of class C_2 and

(a) $\hat{L}(\underline{0}) = 0$ (5-68)

(b) $\hat{R}(\underline{0}) = 0$ (5-69)

(c) $\frac{\partial \hat{R}(\underline{V})}{\partial \underline{V}}$ has a one-to-one correspondence from R^m to R^m.

(d) $[\frac{\partial^2 \hat{R}(\underline{V})}{\partial \underline{V} \partial \underline{V}}]$ is positive definite for a normal Hamiltonian.

It was then asserted, based on Theorem 5-1, that $\hat{V}^o(\underline{Y})$ of class C_2 is the optimal performance index function and $\underline{V}(\underline{Y})$ is the optimal feedback control law for

$$\hat{L}(\underline{Y}) = - \hat{R}(\underline{V}(\underline{Y})) - [\frac{\partial \hat{V}^o}{\partial \underline{Y}}]^T \{\hat{\underline{F}}(\underline{Y}) + \hat{\underline{B}}\, \underline{V}(\underline{Y})\} \quad (5\text{-}70)$$

if and only if $\hat{V}^o(\underline{Y})$ satisfies the Hamilton-Jacobi equation of

$$\hat{H}(\underline{Y}, \frac{\partial \hat{V}^o}{\partial \underline{Y}}, \underline{V}(\underline{Y}), t) = 0 , \quad \text{for all } (\underline{Y},t) \in R^n \times R^1 \quad (5\text{-}71)$$

and $\underline{V}(\underline{Y})$ is given as

$$\underline{V}(\underline{Y}) = \hat{\underline{\mathcal{R}}}(- B^T \frac{\partial \hat{V}^o}{\partial \underline{Y}}) , \quad (5\text{-}72)$$

where $\hat{\underline{\mathcal{R}}}(\ )$ is the H-minimal of the Hamiltonian.

Further investigations were directed to restrict the resulting $\hat{V}^{o}(\underline{Y})$ to a quadratic form $\underline{Y}^{T}\,\hat{\underline{P}}\,\underline{Y}$. Then (5-70) and (5-71) become

$$\hat{L}(\underline{Y}) = -\,\hat{R}(\underline{V}(\underline{Y})) - \underline{Y}^{T}\,\hat{\underline{P}}\,\{\hat{\underline{F}}(\underline{Y}) + \hat{\underline{B}}\,\underline{V}(\underline{Y})\} \qquad (5\text{-}73)$$

and

$$\underline{V}(\underline{Y}) = \hat{\underline{\mathcal{R}}}(-\,\underline{B}^{T}\,\underline{P}\,\underline{Y})\ . \qquad (5\text{-}74)$$

The following three cases were investigated.

(1) If (a) the system (5-67) is linear,

$$\dot{\underline{Y}} = \hat{\underline{A}}\,\underline{Y} + \hat{\underline{B}}\,\underline{V}\ , \qquad (5\text{-}75)$$

(b) the form of the loss function is restricted to

$$\hat{L}(\underline{Y}) = \underline{Y}^{T}\,\hat{\underline{H}}^{T}\,\hat{\underline{H}}\,\underline{Y}$$
$$\hat{R}(\underline{V}) = \underline{V}^{T}\,\underline{V}\ , \qquad (5\text{-}76)$$

where $\hat{\underline{H}}$ is any n_1 x n matrix with the rank $n_1 \leq n$,

(c) a linear control law is assumed

$$\underline{V}(\underline{Y}) = -\,\hat{\underline{K}}^{T}\,\underline{Y}\ , \qquad (5\text{-}77)$$

then (5-73) and (5-74) reduce to

$$\hat{\underline{K}} = \hat{\underline{P}}\,\hat{\underline{B}} \tag{5-78}$$

and

$$-\hat{\underline{P}}\,(\hat{\underline{A}} - \hat{\underline{B}}\,\hat{\underline{K}}^T) - (\hat{\underline{A}} - \hat{\underline{B}}\,\hat{\underline{K}}^T)^T\,\hat{\underline{P}} = \hat{\underline{H}}^T\,\hat{\underline{H}} + \hat{\underline{K}}\,\hat{\underline{K}}^T\ , \tag{5-79}$$

which are generalizations of the equations of Kalman, (5-42) and (5-43). Theoretical justifications are not given, however.

(2) Assume in addition to (1) that (5-75) has a single input with

$$\hat{\underline{B}} = [0,\ 0,\ \ldots,\ 0,\ 1]^T\ . \tag{5-80}$$

Then for a positive definite $\hat{\underline{P}}$, it is necessary for $\hat{k}_n$ $\hat{\underline{K}} = [\hat{k}_1,\ \hat{k}_2,\ \ldots,\ \hat{k}_n]^T$ to be positive. In fact, from (5-78)

$$\hat{k}_n = \hat{P}_{nn} \tag{5-81}$$

of the matrix $\hat{\underline{P}}$. If $\hat{k}_n = 0$, then the sign-definiteness of $\hat{\underline{P}}$ fails. More generally, if the elements of $\hat{\underline{B}}$ are

either 0 or 1, for a positive definite $\hat{\underline{P}}$ it is necessary to have positive elements $\hat{\underline{K}}$ which correspond to the unit elements of $\hat{\underline{B}}$.

(3) For a single input linear system, consider a control law such that

$$\mathcal{U}(\hat{\underline{K}}^T \underline{Y}) , \tag{5-82}$$

where the $\hat{\underline{K}}$ is a n x 1 matrix and the $\mathcal{U}(\sigma)$ is given by

$$\mathcal{U}(\sigma) = \sum_{\substack{i=1 \\ i \text{ odd}}}^{\infty} a_i \sigma^i , \quad a_i > 0 \text{ for all } i . \tag{5-83}$$

It was shown that if the inverse function of $\mathcal{U}(\sigma)$ can be expressed as a power series

$$\tilde{\mathcal{U}}(\sigma) = \sum_{\substack{i=1 \\ i \text{ odd}}}^{\infty} C_i \sigma^i , \tag{5-84}$$

and if $\hat{\underline{P}}$ is positive definite, then each coefficient C_i can be determined explicitely in terms of the coefficients a_i and the components of $\hat{\underline{P}}$, and $\hat{R}(\mathcal{U}(\sigma))$ can be calculated as

$$\hat{R}(\mathcal{V}(\sigma)) = \sum_{\substack{i=2 \\ \text{even}}}^{\infty} d_i \sigma^i \ . \qquad (5\text{-}85)$$

Substituting,

$$\hat{L}(\underline{Y}) = -\frac{1}{2}\underline{Y}^T(\hat{\underline{P}}\,\hat{\underline{A}} + \hat{\underline{A}}^T\hat{\underline{P}})\underline{Y} - \sum_{\substack{i=1 \\ i \text{ odd}}}^{\infty}\{\underline{Y}^T\hat{\underline{P}}\,\hat{\underline{B}}\,a_i(\hat{\underline{K}}^T\underline{Y})^i + d_{i+1}(\hat{\underline{K}}^T\underline{Y})^{i+1}\} \ . \qquad (5\text{-}86)$$

Practically, two examples were demonstrated. One considered the system

$$\begin{bmatrix}\dot{y}_1 \\ \dot{y}_2\end{bmatrix} = \begin{bmatrix}0 & 1 \\ -1 & -a\end{bmatrix}\begin{bmatrix}y_1 \\ y_2\end{bmatrix} + \begin{bmatrix}0 \\ 1\end{bmatrix}\mathcal{V} \qquad (5\text{-}87)$$

with $a > 0$ and

$$\mathcal{V}(\sigma) = \sigma^3 \ , \qquad (5\text{-}88)$$

with

$$\sigma = \hat{\underline{K}}^T\underline{Y} = -(y_1 + \frac{a}{2}y_2) \ . \qquad (5\text{-}89)$$

Then it can be shown that

$$\hat{\underline{R}}(\sigma) = \frac{3a\hat{P}_{22}}{8} \quad \mathcal{U}^{4/3} \tag{5-90}$$

and

$$\hat{L}(\underline{Y}) = \frac{1}{2}\,\underline{Y}^T \begin{bmatrix} a\hat{P}_{22} & -\hat{P}_{11}+\hat{P}_{22}+\frac{a^2}{2}\hat{P}_{22} \\ -\hat{P}_{11}+\hat{P}_{22}+\frac{a^2}{2}\hat{P}_{22} & a\hat{P}_{22} \end{bmatrix} \underline{Y}$$

$$+ \frac{a}{2}\hat{P}_{22}\left(\frac{y_1^4}{4} + \frac{2}{a}y_1^3 y_2 + \frac{6}{a}y_1^2 y_2^2 + \frac{8}{a^3}y_1 y_2^3\right) + \frac{2\hat{P}_{22}}{a^3}y_2^4 , \tag{5-91}$$

with

$$\hat{\underline{P}} = \begin{bmatrix} \hat{P}_{11} & \frac{a}{2}\hat{P}_{22} \\ \frac{a}{2}\hat{P}_{22} & \hat{P}_{22} \end{bmatrix} . \tag{5-92}$$

Thau pointed out that the result, case (3), could be applied to the construction of a Liapunov function for a Lur'e system

$$\dot{\underline{Y}} = \hat{\underline{A}}\,\underline{Y} + \hat{\underline{B}}\,\mathcal{U}(\hat{\underline{K}}^T\underline{Y}) . \tag{5-93}$$

Considering (5-93) as a synthesized feedback control system for a control law $\mathcal{U}(\hat{\underline{K}}^T\underline{Y})$ in the inverse problem,

$\hat{\underline{P}}$, $\hat{L}(\underline{Y})$ and $\hat{R}(\underline{U})$ can be calculated according to the method of the inverse problem. If there exists a $\hat{\underline{P}}$, positive definite, such that

$$\hat{L}(\underline{Y}) + \hat{R}(\underline{U}(\underline{K}^T\underline{Y})) \tag{5-94}$$

is positive definite, then the origin of (5-93) is asymptotically stable in the large from Theorem 2-8, and a Liapunov function is

$$\hat{V}^o(\underline{Y}) = \underline{Y}^T \hat{\underline{P}}\, \underline{Y} . \tag{5-95}$$

In fact, from (5-70), $\dot{\hat{V}}^o(\underline{Y})$ is the negative of (5-94).

Summarizing Thau's work, the form of his loss function was less restrictive than that considered by Kalman and Suga. However, the nonnegative character of $\hat{L}(\underline{Y})$ was not discussed. The application of the results to construct Liapunov functions is a unique contribution, although the necessary assumptions are quite restrictive.

5.4 Comments

Based on Kalman's study of the inverse problem, Suga generalized the assumptions to include time varying multi-input linear systems (5-45), a performance index with any time interval, a time-varying loss function and a

time-varying linear control law. Thau further generalized the assumptions to include multi-input nonlinear systems (5-66), a broader class of loss functions and control laws. Although relevant characteristics of the optimal feedback control system are revealed by this work, the nonnegative property of either the loss function or the optimal performance index function was not adequately considered. Accordingly the results sometimes seem unrealistic from the viewpoint of optimal control theory, as illustrated by examples.

For the inverse problem considered in the next two chapters, generalizations of Kalman's assumptions are made with respect to the nonlinear, multi-input systems, the form of loss function and the nonlinear control law. Furthermore, nonnegativity of an optimized loss function and an optimized performance index function are considered. The canonical form developed in Chapter 4 contributes to the efficient analyses and the compact descriptions of the development.

Chapter 6

INVERSE PROBLEM OF THE OPTIMAL REGULATOR

In this chapter, the inverse problem of the optimal regulator is considered in a general context, i.e., for a class of multi-input systems with an unspecified nonlinearity and feedback control law. Following a precise description of the problem in Section 6.1, an equivalent problem is defined in Section 6.2, using the canonical form developed in Chapter 4. Fundamental lemmas for the analysis are given in Section 6.3. Based on the Hamilton-Jacobi theory and Caratheodory's lemma, a principal theorem for the inverse problem is stated in Section 6.4. Section 6.5 has a discussion of the relevant aspects of this theorem to optimal feedback control systems and to work by other authors as special cases. Two practical examples of the application of the theorem are demonstrated in Section 6.6.

6.1 Statement of the Inverse Problem

The inverse problem of the optimal regulator is considered in this chapter under the assumptions such that

(i) the system equation is given by

$$\dot{\underline{Y}} = \hat{\underline{F}}(\underline{Y}) + \hat{\underline{B}}\,\underline{V} \,, \qquad (6\text{-}1)$$

where $\hat{\underline{F}}(\underline{Y})$ is an n-dimensional vector valued function of class C_2 satisfying $\hat{\underline{F}}(\underline{0}) = \underline{0}$ and $\hat{\underline{B}}$ is an n x m matrix such that

$$0 < \text{rank } \hat{\underline{B}} = r \leq m \leq n \,, \tag{6-2}$$

(ii) the desired final condition of the system is $\underline{Y} = \underline{0}$ in $R^n \times R^1$, with a stable feedback control law given by an m-dimensional vector valued function $\underline{V}(\underline{Y})$ of class C_2 with $\underline{V}(\underline{0}) = \underline{0}$ (thus the origin of the synthesized system

$$\dot{\underline{Y}} = \hat{\underline{F}}(\underline{Y}) + \hat{\underline{B}}\,\underline{V}(\underline{Y}) \tag{6-3}$$

is asymptotically stable in the large),

(iii) the form of performance index is restricted to

$$\int_{t_o}^{\infty} \{\hat{L}(\underline{Y}) + \underline{V}^T \hat{\underline{R}}\,\underline{V}\}\, dt \,, \tag{6-4}$$

where $\hat{\underline{R}}$ is an m x m symmetric positive definite matrix and $\hat{L}(\underline{Y})$ is of class C_2 satisfying

$$\hat{L}(\underline{0}) = 0 \,. \tag{6-5}$$

Equivalently an initial condition $(\underline{Y}_o, t_o)$ and a suitable control function $\underline{V}(t)$ defined on $[t_o, \infty)$ are assumed such that

$$\int_{t_o}^{\infty} \{\hat{L}(\underline{\hat{\phi}}_{\underline{V}}(t; \underline{Y}_o, t_o)) + \underline{V}^T(t)\, \hat{\underline{R}}\, \underline{V}(t)\}\, dt$$

$$= \lim_{\tau \to \infty} \int_{t_o}^{\tau} \{\hat{L}(\underline{\hat{\phi}}_{\underline{V}}(t; \underline{Y}_o, t_o) + \underline{V}^T(t)\, \hat{\underline{R}}\, \underline{V}(t)\}dt \quad , \tag{6-6}$$

(iv) M_s and A.S.R. are the whole $R^n \times R^1$ and A.C.R. is the whole $R^m \times R^1$.

The inverse problem is to find performance indices (6-4) or, equivalently, loss functions that are optimized by the assumed control law, under assumptions (i)-(iv). This inverse problem is a generalization of the inverse problems considered by other authors, as reviewed in Chapter 5, i.e., the assumptions of the problem in Section 6.1 are less restrictive than those previously made. Specifically:

(a) (6-1) is a nonlinear multi-input system with $\hat{\underline{B}}$ of a general rank, in comparison with (5-35) used by Kalman and (5-45) by Suga;

(b) the control law (6-3) is unspecified, in comparison with (5-36) and (5-47) assumed by Kalman and Suga;

(c) the performance index (6-4) has a general penalty function $\hat{L}(\underline{Y})$, in comparison with the quadratic form

used by Kalman, (5-37).

Consequently, Kalman's problem is completely generalized in this chapter. However, all assumptions are limited to time invariancy in comparison to those of Suga. Also the quadratic form in $\underline{V}$ in the performance index (6-4) is more restrictive than the $\hat{R}(\underline{V})$ used by Thau (5-67).

6.2 An Equivalent Inverse Problem

The analysis of the inverse problem can be facilitated by using the canonical form given in Chapter 4. Consider the canonical form of (6-1) as

$$\dot{\underline{X}} = \underline{F}(\underline{X}) + \underline{B}\ \underline{U} \tag{6-7}$$

for the transformation

$$\begin{aligned} \underline{X} &= \underline{N}^{-1}\ \underline{Y} \\ \underline{U} &= \underline{M}^{-1}\ \underline{V}\ , \end{aligned} \tag{6-8}$$

and specifically define

$$\underline{U}(\underline{X}) \triangleq M^{-1}\ \underline{V}(\underline{N}\ \underline{X}) \tag{6-9}$$

as the feedback control law, equivalent to that given in (6-3).

The problem assumptions given in Section 6.1 are invarient with this transformation (6-8) and can be expressed in terms of the transformed variables. This is shown in the remainder of the section.

The rank of $\underline{B}$ is r from (4-193) and

$$\underline{F}(\underline{0}) = \underline{N}^{-1}\,\hat{\underline{F}}\,(\underline{N}\ \underline{0}) = \underline{0}\ . \qquad (6\text{-}10)$$

Also $\underline{F}(\underline{X})$ is of class C_2, as each function $f_i(\underline{X})$, $\frac{\partial f_i(\underline{X})}{\partial x_j}$ or $\frac{\partial^2 f_i(\underline{X})}{\partial x_k \partial x_j}$, i,j,k = 1,2,...n, is a linear combination of $\hat{f}_i(N\ \underline{X})$, $\frac{\partial \hat{f}_i(\underline{N}\ \underline{X})}{\partial x_j}$ and $\frac{\partial^2 \hat{f}_i(\underline{N}\ \underline{X})}{\partial x_k \partial x_j}$ respectively, and Theorem 2-1 can be applied. As $\underline{V}(\underline{0}) = \underline{0}$, from (6-8)

$$\underline{U}(\underline{0}) = \underline{M}^{-1}\,\underline{V}(\underline{N}\ \underline{0}) = \underline{0}\ , \qquad (6\text{-}11)$$

and $\underline{U}(\underline{X})$ can be similarly shown to be of class C_2.

To establish $\underline{U}(\underline{X})$ as a stable control law, a lemma is introduced.

Lemma 6-1:

If $\underline{V}(\underline{Y})$ is a stable control law for (6-1), then $\underline{U}(\underline{X})$ is a stable control law for the transformed system (6-7).

Proof: From Lemma 4-4, a solution for (6-7) with an initial condition $(\underline{X}_o, t_o)$ is

$$\underline{\Phi}_f(t; \underline{X}_o, t_o) = \underline{N}^{-1} \hat{\underline{\Phi}}_f(t; \underline{N}\,\underline{X}_o, t_o) \ , \qquad (6\text{-}12)$$

where $\hat{\underline{\Phi}}_f(t; \underline{N}\,\underline{X}_o, t_o)$ is a solution of (6-3) with an initial condition $(\underline{N}\,\underline{X}_o, t_o)$. Since the origin of (6-3) is asymptotically stable in the large by assumption, for arbitrary $\varepsilon > 0$, $\mu > 0$ and t_o, there exist a $\hat{\delta}(\varepsilon, t_o) > 0$ and a $\hat{T}(\delta, \mu, t_o) > 0$ such that if

$$||\underline{Y}_o|| \leq \hat{\delta}(\varepsilon, t_o) \ , \qquad (6\text{-}13)$$

then

(a) $$||\hat{\underline{\Phi}}_f(t; \underline{Y}_o, t_o)|| \leq \varepsilon \ , \text{ for all } t \geq t_o \ , \qquad (6\text{-}14)$$

and

(b) $$||\hat{\underline{\Phi}}_f(t; \underline{Y}_o, t_o)|| \leq \mu \ , \text{ for all } t \geq t_o + \hat{T} \ . \qquad (6\text{-}15)$$

Define

$$\delta(\varepsilon, t_o) \triangleq \frac{\hat{\delta}\left(\dfrac{\varepsilon}{\lambda_{max}((\underline{N}^{-1})^T\,\underline{N}^{-1})}\, t_o\right)}{\lambda_{max}(\underline{N}^T\,\underline{N})} \qquad (6\text{-}16)$$

and

$$T(\delta,\ \mu,\ t_o) = \hat{T}\ (\delta, \frac{\mu}{\lambda_{max}((\underline{N}^{-1})^T\ \underline{N}^{-1})}\ ,\ t_o)\ , \quad (6\text{-}17)$$

where $\lambda_{max}(\ )$ is defined in Theorem 2-5 with $\underline{Q} = \underline{N}^T\ \underline{N}$. Then for the $\varepsilon > 0$, $\mu > 0$ and t_o chosen, consider

$$||\underline{X}_o|| \leq \delta(\varepsilon,\ t_o)\ . \quad (6\text{-}18)$$

From (6-16), (6-18) and Theorem 2-5,

$$\hat{\delta}(\frac{\varepsilon}{\lambda_{max}((\underline{N}^{-1})^T\ \underline{N}^{-1})}\ ,\ t_o) \geq \lambda_{max}\ (\underline{N}^T\ \underline{N})\,||\underline{X}_o||$$

$$\geq ||\underline{N}\ \underline{X}_o|| = ||\underline{Y}_o||\ . \quad (6\text{-}19)$$

Therefore, from (6-13) and (6-19), (6-14) becomes

$$||\hat{\underline{\Phi}}_f(t;\ \underline{Y}_o,\ t_o)|| \leq \frac{\varepsilon}{\lambda_{max}((\underline{N}^{-1})^T\ \underline{N}^{-1})}\ ,\ \text{for all } t \geq t_o\ , \quad (6\text{-}20)$$

or

$$\varepsilon \geq \lambda_{max}((\underline{N}^{-1})^T\ \underline{N}^{-1})\,||\hat{\underline{\Phi}}_f(t;\underline{Y}_o,t_o)|| \geq ||N^{-1}\hat{\underline{\Phi}}_f(t;\underline{N}\ \underline{X}_o,t_o)||$$

$$= ||\underline{\Phi}_f(t;\underline{X}_o,t_o)||\ ,\ \text{for all } t \geq t_o\ , \quad (6\text{-}21)$$

using Theorem 2-5 and Lemma 4-4. Also, from (6-15) and (6-17), it follows that

$$||\hat{\underline{\Phi}}_f(t;\ \underline{Y}_o,\ t_o)|| \leq \frac{\mu}{\lambda_{max}((\underline{N}^{-1})^T\ N^{-1})}\ ,\ \text{for all } t \geq t_o + T, \tag{6-22}$$

or

$$\mu \geq \lambda_{max}((\underline{N}^{-1})^T\ \underline{N}^{-1})\ ||\hat{\underline{\Phi}}_f(t;\underline{Y}_o,t_o)|| \geq ||\underline{\Phi}_f(t;\underline{X}_o,t_o)||\ ,$$

$$\text{for all } t \geq t_o + T\ , \tag{6-23}$$

again using Theorem 2-5 and Lemma 4-4. By (6-18), (6-21) and (6-23), the lemma is proved.

The value of the performance index given by (6-5) for a suitable control $\underline{V}(t)$ and from an initial condition $(\underline{Y}_o,\ t_o)$ is

$$\hat{J}\ [\underline{Y}_o,t_o,\underline{V}(t)] = \int_{t_o}^{\infty} \{\hat{L}(\hat{\underline{\Phi}}_{\underline{V}}(t;\underline{Y}_o,t_o)) + \underline{V}^T(t)\ \hat{\underline{R}}\ \underline{V}(t)\}dt\ . \tag{6-24}$$

This reduces by Lemma 4-4 for (6-8), to

$$\hat{J}\ [\underline{N}\ \underline{X}_o,t_o,\underline{M}\ \underline{U}(t)] \triangleq \int_{t_o}^{\infty} \{\hat{L}(\underline{N}\ \underline{\Phi}_{\underline{U}}(t;\underline{X}_o,t_o)) + \underline{U}^T(t)\underline{M}^T\ \hat{\underline{R}}\ \underline{M}\ \underline{U}(t)\}dt\ . \tag{6-25}$$

Define

$$J[\underline{X}_o, t_o, \underline{U}(t)] \triangleq \hat{J}\,[\underline{N}\ \underline{X}_o, t_o, \underline{M}\ \underline{U}(t)]$$

$$L(\underline{X}) \triangleq \hat{L}(\underline{N}\ \underline{X}) \qquad (6\text{-}26)$$

$$\underline{R} \triangleq \underline{M}^T\ \hat{\underline{R}}\ \underline{M}\ .$$

$L(\underline{X})$ is of class C_2 because $\hat{L}(\underline{Y})$ is of class C_2 and Theorem 2-1. Also $L(\underline{0}) = \hat{L}(\underline{N}\ \underline{0}) = 0$, and $\underline{R}$ is symmetric and positive definite, from part (iii) of Theorem 2-4. Then the performance index (6-4) is equivalently written as

$$J[\underline{X}_o, t_o, \underline{U}(t)] = \int_{t_o}^{\infty} \{L(\underline{X}) + \underline{U}^T\ \underline{R}\ \underline{U}\}dt \qquad (6\text{-}27)$$

for the canonical form (6-7), and assumption (iii) for $\hat{L}(\underline{Y})$ and $\hat{\underline{R}}$ is completely preserved for $L(\underline{X})$ and $\underline{R}$.

Since M_s, A.S.R. and A.C.R. in the $\underline{X}$ and $\underline{U}$ coordinates are whole $R^n \times R^1$ and $R^m \times R^1$ because of the bijective mappings of (6-8), assumption (iv) is also preserved. Thus, the original inverse problem stated in Section 6.1 can be considered in the canonical form (6-7) under the same mathematical assumptions without loss of generality. The recovery of the solution for the original system follows from the inverse transformations of (6-8) and (6-26).

For convenience of analysis, the following notation is used for (6-7), (6-9) and (6-27).

(1) $\underline{X}$ is decomposed into

$$\underline{X}_1 \triangleq [x_1, x_2, \ldots, x_{n-r}]^T$$
$$\underline{X}_2 \triangleq [x_{n-r+1}, \ldots, x_n]^T , \qquad (6\text{-}28)$$

and, if $n = r$, $\underline{X} = \underline{X}_2$.

(2) $\underline{U}$ is decomposed into

$$\underline{U}_d \triangleq [u_1, u_2, \ldots, u_{m-r}]^T$$
$$\underline{U}_e \triangleq [u_{m-r+1}, \ldots, u_m]^T , \qquad (6\text{-}29)$$

and, if $m = r$, $\underline{U} = \underline{U}_e$.

(3) $\underline{F}(\underline{X})$ is decomposed into

$$\underline{F}_1(\underline{X}) \triangleq [f_1(\underline{X}), f_2(\underline{X}), \ldots, f_{n-r}(\underline{X})]^T$$
$$\underline{F}_2(\underline{X}) \triangleq [f_{n-r+1}(\underline{X}), \ldots, f_n(\underline{X})]^T , \qquad (6\text{-}30)$$

with

$$\underline{F}(\underline{X}) = \underline{A}\,\underline{X} + \underline{F}'(\underline{X}) , \qquad (6\text{-}31)$$

where $\underline{A}\ \underline{X}$ defines the first degree homogeneous function in $\underline{F}(\underline{X})$, and $\underline{\Gamma}(\underline{X})$ the remainder. $\underline{A}$ and $\underline{\Gamma}(\underline{X})$ are further defined as

$$\underline{A} \triangleq \begin{bmatrix} \underline{A}_1 \\ \underline{A}_2 \end{bmatrix} \triangleq \begin{bmatrix} \underline{A}_{11} & \underline{A}_{12} \\ \underline{A}_{21} & \underline{A}_{22} \end{bmatrix} \begin{matrix} \} \ n-r \\ \} \ r \end{matrix} \qquad (6\text{-}32)$$

$$\qquad\qquad\quad \underbrace{\qquad}_{n-r} \quad \underbrace{\qquad}_{r}$$

and

$$\underline{\Gamma}(\underline{X}) = \begin{bmatrix} \underline{\Gamma}_1(\underline{X}) \\ \underline{\Gamma}_2(\underline{X}) \end{bmatrix} \begin{matrix} \} \ n-r \\ \} \ r \end{matrix} \qquad (6\text{-}33)$$

Subsequently (6-7) can be written as

$$\begin{aligned} \dot{\underline{X}}_1 &= \underline{F}_1(\underline{X}) \\ \dot{\underline{X}}_2 &= \underline{F}_2(\underline{X}) + \underline{U}_e \end{aligned} \qquad (6\text{-}34)$$

from (6-30), or

$$\begin{aligned} \dot{\underline{X}}_1 &= \underline{A}_1\ \underline{X} + \underline{\Gamma}_1(\underline{X}) \\ \dot{\underline{X}}_2 &= \underline{A}_2\ \underline{X} + \underline{\Gamma}_2(\underline{X}) + U_e \end{aligned} \qquad (6\text{-}35)$$

from (6-32) and (6-33).

(4) Let

$$\underline{R} \triangleq \begin{bmatrix} \underline{R}_{11} & \underline{R}_{12} \\ \underline{R}_{12}^T & \underline{R}_{22} \end{bmatrix} \begin{matrix} \} \, m-r \\ \} \, r \end{matrix} \qquad (6\text{-}36)$$

$$\begin{matrix} \underbrace{} & \underbrace{} \\ m-r & r \end{matrix}$$

(5) If $\nabla^o(\underline{X})$ is a scalar function of class C_2, then define

$$\frac{\partial \nabla^o}{\partial \underline{X}_1} \triangleq [w_1(\underline{X}), w_2(\underline{X}), \ldots w_{n-r}(\underline{X})]^T$$

$$\frac{\partial \nabla^o}{\partial \underline{X}_2} \triangleq [w_{n-r+1}(\underline{X}), \ldots w_n(\underline{X})]^T \qquad (6\text{-}37)$$

and

$$\frac{\partial \nabla^o}{\partial \underline{X} \partial \underline{X}} = \begin{bmatrix} \frac{\partial^2 \nabla^o}{\partial \underline{X}_1 \partial \underline{X}_1} & \frac{\partial^2 \nabla^o}{\partial \underline{X}_2 \partial \underline{X}_1} \\ \frac{\partial^2 \nabla^o}{\partial \underline{X}_1 \partial \underline{X}_2} & \frac{\partial^2 \nabla^o}{\partial \underline{X}_2 \partial \underline{X}_2} \end{bmatrix},$$

$$= \begin{bmatrix} \frac{\partial w_1}{\partial x_1} & \frac{\partial w_2}{\partial x_2} & \cdots & \frac{\partial w_1}{\partial x_n} \\ \frac{\partial w_2}{\partial x_1} & & & \vdots \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ \frac{\partial w_n}{\partial x_1} & \cdots & & \frac{\partial w_n}{\partial x_n} \end{bmatrix} . \qquad (6\text{-}38)$$

6.3 Fundamental Lemmas

Lemma 6-2:

Let m and r be integers such that $m > r > 0$ and $\underline{R}$ $m \times m$ symmetric matrix. Defining $\underline{R}$ as (6-36), $\underline{R}$ is positive definite if and only if

(i) $\underline{R}_{11}$ and $\underline{R}_{22} - \underline{R}_{12}^T \underline{R}_{11}^{-1} \underline{R}_{12}$ are positive definite,

or

(ii) $\underline{R}_{22}$ and $\underline{R}_{11} - \underline{R}_{12} \underline{R}_{22}^{-1} \underline{R}_{12}^T$ are positive definite.

Proof: Only (i) is proved; (ii) is proved similarly. For $\underline{R}$ to be positive definite, $\underline{R}_{11}$ must be positive definite by (i) of Corollary 2-1. Thus $\underline{R}_{11}^{-1}$ exists. Define a matrix

$$\underline{S} \triangleq \begin{bmatrix} I_{m-r} & -\underline{R}_{11}^{-1}\underline{R}_{12} \\ [0] & I_r \end{bmatrix} \tag{6-39}$$

which is nonsingular. According to (iii) of Theorem 2-4, $\underline{R}$ is positive definite if and only if

$$\underline{S}^T \underline{R}\, \underline{S} = \begin{bmatrix} \underline{R}_{11} & [0] \\ [0] & \underline{R}_{22} - \underline{R}_{12}^T \underline{R}_{11}^{-1} \underline{R}_{12} \end{bmatrix} \tag{6-40}$$

is positive definite. Since the characteristic equation of (6-40) is

$$0 = |\lambda I - \underline{S}^T \underline{R}\, \underline{S}| = |\lambda I - \underline{R}_{11}| \cdot |\lambda I - (\underline{R}_{22} - \underline{R}_{12}^T \underline{R}_{11}^{-1} \underline{R}_{12})|, \tag{6-41}$$

eigenvalues of (6-40) are those of $\underline{R}_{11}$ and of $(\underline{R}_{22} - \underline{R}_{12}^T \underline{R}_{11}^{-1} \underline{R}_{12})$. From Theorem 2-4, the assertion is immediate.

Lemma 6-3:

For $\underline{R}$ given as in (6-36) and for the rank of $\underline{R}_{22}$ as $r_1 \leq r$, $\underline{R}$ is positive semidefinite if and only if

(i) $\underline{R}_{22}$ is positive semidefinite,

(ii) $\underline{\Omega} \triangleq \underline{R}_{11} - \underline{R}_{12}\underline{D}_{22}^T \underline{D}_{22} \underline{R}_{22} \underline{D}_{22}^T \underline{D}_{22} \underline{R}_{12}^T$ is positive semi-definite,

where $\underline{D}_{22}$ is an r x r nonsingular congruent transformation matrix such that

$$\underline{D}_{22}^{T}\ \underline{R}_{22}\ \underline{D}_{22} = \begin{bmatrix} I_{r_1} & [0] \\ [0] & [0] \end{bmatrix} , \qquad (6\text{-}42)$$

(the existence of $\underline{D}_{22}$ follows from Theorem 2-4 and (i) above),

and

(iii) the last $r - r_1$ columns of $\underline{R}_{12}\ \underline{D}_{22}$ are null, i.e.,

$$\underline{R}_{12}\ \underline{D}_{22} \triangleq \left[\underbrace{\underline{\Sigma}}_{r_1} \quad \underbrace{\underline{\mathcal{R}}}_{r-r_1}\right] \Big\}\, m-r \ , \text{ with } \underline{\mathcal{R}} = [0] \ . \qquad (6\text{-}43)$$

Proof: Assume an m-dimensional vector $\underline{Z} \triangleq [z_1, z_2, \ldots, z_m]^T$ and

$$\begin{aligned} \underline{Z}_1 &= [z_1,\ z_2,\ \ldots,\ z_{m-r}]^T \\ \underline{Z}_2 &= [z_{m-r+1},\ z_{m-r+2},\ \ldots,\ z_{m-r+r_1}]^T \qquad (6\text{-}44) \\ \underline{Z}_3 &= [z_{m-r+r_1+1},\ z_{m-r+r_1+2},\ \ldots,\ z_m]^T \ . \end{aligned}$$

Then $\underline{R}$ is positive semidefinite if and only if the quadratic form $\underline{Z}^T\ \underline{R}\ \underline{Z}$ is positive semidefinite. Fix $\underline{Z}_1 = 0$ and then

$$\underline{Z}^T \underline{R} \underline{Z} = [\underline{Z}_2^T \quad \underline{Z}_3^T] \quad \underline{R}_{22} \begin{bmatrix} \underline{Z}_2 \\ \underline{Z}_3 \end{bmatrix} \tag{6-45}$$

and (i) is necessary. For (iii), define

$$D \triangleq \begin{bmatrix} I_{m-r} & [0] \\ -\underline{\Delta} & \underline{D}_{22} \end{bmatrix} \tag{6-46}$$

with

$$\underline{\Delta} \triangleq \underline{D}_{22} \underline{D}_{22}^T \underline{R}_{22} \underline{D}_{22} \underline{D}_{22}^T \underline{R}_{12}^T . \tag{6-47}$$

$\underline{D}$ is nonsingular because of the nonsingular $\underline{D}_{22}$ given by (6-42). From (iii) of Theorem 2-4, $\underline{R}$ is positive semi-definite if and only if

$$\underline{D}^T \underline{R} \underline{D} = \begin{bmatrix} \underline{R}_{11} - \underline{\Delta}^T \underline{R}_{12}^T \underline{\Delta} + \underline{\Delta}^T \underline{R}_{22} \underline{\Delta} & \underline{R}_{12} \underline{D}_{22} - \underline{\Delta}^T \underline{R}_{22} \underline{D}_{22} \\ \underline{D}_{22}^T \underline{R}_{12}^T - \underline{D}_{22}^T \underline{R}_{22} \underline{\Delta} & \underline{D}_{22}^T \underline{R}_{22} \underline{D}_{22} \end{bmatrix} \tag{6-48}$$

is positive semidefinite. But

$$\underline{R}_{11} - \underline{\Delta}^T\underline{R}_{12}^T - \underline{R}_{12}\underline{\Delta} + \underline{\Delta}^T\underline{R}_{22}\underline{\Delta} = \underline{R}_{11} - \underline{R}_{12}\underline{D}_{22}\underline{D}_{22}^T\underline{R}_{22}\underline{D}_{22}\underline{D}_{22}^T\underline{R}_{12}^T$$

$$= \underline{\Omega} \qquad (6\text{-}49)$$

and, from (6-43),

$$\underline{R}_{12}\underline{D}_{22} - \underline{\Delta}^T\underline{R}_{22}\underline{D}_{22} = \underline{R}_{12}\underline{D}_{22}\begin{bmatrix} [0] & [0] \\ [0] & I_{r-r_1} \end{bmatrix} = \begin{bmatrix} [0] & \underline{\mathcal{R}} \end{bmatrix} , \qquad (6\text{-}50)$$

where $\underline{\mathcal{R}}$ is the last $(r - r_1)$ columns of $\underline{R}_{12}\ \underline{D}_{22}$. Substituting (6-47), (6-49) and (6-50) into (6-48),

$$\underline{D}^T\ \underline{R}\ \underline{D} = \begin{bmatrix} \underline{\Omega} & [0] & \underline{\mathcal{R}} \\ [0] & I_{r_1} & [0] \\ \underbrace{\underline{\mathcal{R}}^T}_{m-r} & \underbrace{[0]}_{r_1} & \underbrace{[0]}_{r-r_1} \end{bmatrix} \begin{matrix} \} \ m-r \\ \} \ r_1 \\ \} \ r-r_1 \end{matrix} . \qquad (6\text{-}51)$$

In order for $\underline{D}^T\ \underline{R}\ \underline{D}$ to be positive semidefinite, $\underline{\mathcal{R}}$ must be the null matrix. If it is not, consider $\underline{Z}_2 = \underline{0}$ and $\underline{Z}$ satisfying $\underline{Z}_1\ \underline{\mathcal{R}} \neq \underline{0}$. Then

$$\underline{Z}^T\underline{D}^T\underline{R}\ \underline{D}\ \underline{Z} = \underline{Z}_1^T\ \underline{\Omega}\ \underline{Z}_1^T + 2\underline{Z}_1^T\ \underline{\mathcal{R}}\ \underline{Z}_3 . \qquad (6\text{-}52)$$

As $\underline{Z}_3$ can be chosen to provide a negative value for (6-52), the positive semidefiniteness of $\underline{D}^T\underline{R}\ \underline{D}$ fails. Therefore $\underline{\mathcal{R}} = [0]$ and (iii) is necessary. Alternately, fixing $\underline{Z}_2 = \underline{0}$ and $\underline{Z}_3 = \underline{0}$,

$$\underline{Z}^T\underline{D}^T\underline{R}\ \underline{D}\ \underline{Z} = \underline{Z}_1^T\ \underline{\Omega}\ \underline{Z}_1 \ , \tag{6-53}$$

and (ii) is required.

Conversely if (i)-(iii) are satisfied,

$$\underline{Z}^T\underline{D}^T\underline{R}\ \underline{D}\ \underline{Z} = \underline{Z}_1^T\underline{\Omega}\ \underline{Z}_1 + \underline{Z}_2^T\underline{Z}_2 \ , \tag{6-54}$$

and $\underline{D}^T\underline{R}\ \underline{D}$ are positive semidefinite as $\underline{\Omega}$ is positive semidefinite; by Theorem 2-4, $\underline{R}$ is then positive semi-definite.

For the loss function in (6-27), define

$$\nabla(\tau;\underline{X}_o,t_o) = \int_{t_o}^{\tau} \{L(\underline{\Phi}_f(t;\underline{X}_o,t_o)) + \underline{U}^T(\underline{\Phi}_f(t;\underline{X}_o,t_o))\underline{R}\ \underline{U}(\underline{\Phi}_f(t;\underline{X}_o,t_o))\}dt \tag{6-55}$$

and

$$\nabla(\underline{X}_o,\ t_o) = \lim_{\tau\to\infty}\ \nabla(\tau;\ \underline{X}_o,\ t_o) \ , \tag{6-56}$$

where $\underline{\Phi}_f(t;\ \underline{X}_o,\ t_o)$ is the solution of

$$\dot{\underline{X}} = \underline{F}(\underline{X}) + \underline{B}\ \underline{U}(\underline{X}) \qquad (6\text{-}57)$$

from $(\underline{X}_o,\ t_o)$.

Lemma 6-4:

Provided that $\nabla(\underline{X},\ t)$, (6-56), is well defined on $R^n \times R^1$, for an arbitrary $(\underline{X}_o,\ t_o) \in R^n \times R^1$ and $\alpha > 0$,

$$\nabla(\underline{X}_o,\ t_o) = \nabla(\underline{X}_o,\ t_o + \alpha)\ , \qquad (6\text{-}58)$$

that is, $\nabla(\underline{X},\ t)$ is independent of t.

Proof: From (6-56), the statement is justified by proving that for each $(\underline{X}_o,\ t_o)$ and $\alpha > 0$,

$$\lim_{\tau \to \infty} \{\nabla(\tau;\ \underline{X}_o,\ t_o) - \nabla(\tau;\ \underline{X}_o,\ t_o + \alpha)\} = 0\ , \qquad (6\text{-}59)$$

or, for an arbitrary $\epsilon > 0$, that there exists a $T > 0$ such that

$$|\nabla(\tau;\ \underline{X}_o,\ t_o) - \nabla(\tau;\ \underline{X}_o,\ t_o + \alpha)| \leq \epsilon, \quad \text{for all } \tau > T\ . \qquad (6\text{-}60)$$

As (6-57) is autonomous, it is known that[13]

$$\underline{\phi}_f(t + \alpha, \underline{X}_o, t_o + \alpha) = \underline{\phi}_f(t, \underline{X}_o, t_o) \ . \qquad (6\text{-}61)$$

Then, for arbitrary $\tau > t_o + \alpha$, it follows from (6-55) that

$$V(\tau; \underline{X}_o, t_o) = \int_{t_o+\alpha}^{\tau+\alpha} \{L(\underline{\phi}_f(t; \underline{X}_o, t_o + \alpha)$$

$$+ \underline{U}^T(\underline{\phi}_f(t;\underline{X}_o,t_o+\alpha))\underline{R}\ \underline{U}(\underline{\phi}_f(t;\underline{X}_o,t_o+\alpha))\}dt$$

$$= V(\tau;\underline{X}_o,t_o+\alpha) + \int_{\tau}^{\tau+\alpha} \{L(\underline{\phi}_f(t;\underline{X}_o,t_o+\alpha)$$

$$+ \underline{U}^T(\underline{\phi}_f(t;\underline{X}_o,t_o+\alpha))\underline{R}\ \underline{U}(\underline{\phi}_f(t;\underline{X}_o,t_o+\alpha))\}dt \ . \qquad (6\text{-}62)$$

As $L(\underline{X})$ and $\underline{U}(\underline{X})$ are continuous and zero at $\underline{X} = \underline{0}$ by the assumptions, i.e., (6-5) and (6-11), there exist $\mu_1 > 0$ and $\mu_2 > 0$ satisfying

$$|L(\underline{X})| < \frac{\varepsilon}{2\alpha} \ , \text{ if } ||\underline{X}|| < \mu_1 \ , \qquad (6\text{-}63)$$

and

$$||\underline{U}(\underline{X})|| < \sqrt{\frac{\varepsilon}{2\alpha\lambda_{max}(\underline{R})}} \ , \text{ if } ||\underline{X}|| < \mu_2 \ , \qquad (6\text{-}64)$$

where $\lambda_{max}(\underline{R})$ is given in Theorem 2-5. Since the origin of (6-59) is assumed asymptotically stable in the large, for the given $(\underline{X}_o, t_o)$, there exists a $T_1 > 0$ such that

$$||\underline{\Phi}_f(t; \underline{X}_o, t_o + \alpha)|| \leq \text{Min}(\mu_1, \mu_2) \text{ , for all } t \geq t_o + T_1. \tag{6-65}$$

From (6-62∿65), it follows that

$$\begin{aligned} &|\nabla(\tau;\underline{X}_o,t_o) - \nabla(\tau;\underline{X}_o,t_o+\alpha)| = \\ &\leq \left|\int_{\tau}^{\tau+\alpha} L(\underline{\phi}_f(t;\underline{X}_o,t_o+\alpha))\,dt\right| \\ &+ \left|\int_{\tau}^{\tau+\alpha} \underline{U}^T(\underline{\phi}_f(t;\underline{X}_o,t_o+\alpha))\,\underline{R}\,\underline{U}(\underline{\phi}_f(t;\underline{X}_o,t_o+\alpha))dt\right| \\ &\leq \frac{\varepsilon}{2\alpha}\cdot\alpha + \frac{\varepsilon}{2\alpha}\cdot\alpha = \varepsilon, \text{ for all } \tau \geq t_o + \alpha + T_1 . \end{aligned} \tag{6-66}$$

Thus (6-60) follows if

$$T \triangleq t_o + \alpha + T_1 . \tag{6-67}$$

Accordingly, $\nabla(\underline{X}, t)$ in (6-56) can be simply described as $\nabla(\underline{X})$.

Consider a class of nonlinear systems given by

$$\dot{\underline{X}} = \underline{A}\,\underline{X} + \underline{B}\,\underline{U} + \underline{B}_e\,\underline{F}_2(\underline{X}) \quad , \tag{6-68}$$

that is, $\underline{F}_1(\underline{X})$ in (6-35) is identically zero and $\underline{B}_e$ given by (4-114). The controllability of this particular class of systems can be established by the application of Lemma 6-5, based on Theorem 2-9.

Lemma 6-5:

A system (6-68) is completely controllable if and only if the system

$$\dot{\underline{X}} = \underline{A}\,\underline{X} + \underline{B}\,\underline{U} \tag{6-69}$$

is completely controllable.

Proof: Assume that (6-68) is completely controllable. Then, from Definition 2-11, for arbitrary $\underline{X}_o$, $\underline{X}_1 \in R^n$ and $t_o \in R^1$, there exists a control function $\underline{\mathcal{U}}(t)$ to provide a solution

$$\underline{\Phi}_{\underline{\mathcal{U}}}(t;\ \underline{X}_o,\ t_o) \tag{6-70}$$

satisfying

$$\dot{\underline{\Phi}}_{\underline{u}}(t;\ \underline{X}_o,\ t_o) = \underline{A}\ \underline{\Phi}_{\underline{u}}(t;\ \underline{X}_o,\ t_o)$$

$$+ \underline{B}\,\underline{u}(t) + \underline{B}_e\ \underline{F}_2(\underline{\Phi}_{\underline{u}}(t;\ \underline{X}_o,\ t_o)) \qquad (6\text{-}71)$$

and at some $t_1 \geq t_o$

$$\underline{\Phi}_{\underline{u}}(t_1;\ \underline{X}_o,\ t_o) = \underline{X}_1 \ . \qquad (6\text{-}72)$$

Consider a control function

$$\underline{u}(t) + \begin{bmatrix} \underline{0} \\ \underline{F}_2(\underline{\Phi}_{\underline{u}}(t;\underline{X}_o,t_o)) \end{bmatrix} \qquad (6\text{-}73)$$

for (6-69), i.e.,

$$\dot{\underline{X}} = \underline{A}\ \underline{X} + \underline{B}\,\underline{u}(t) + \underline{B}_e\underline{F}_2(\underline{\Phi}_{\underline{u}}(t;\underline{X}_o,t_o)\ , \qquad (6\text{-}74)$$

where $\underline{0}$ is $(m - r)$ dimensional. Then the solution from $(\underline{X}_o,\ t_o)$ is (6-70) and (6-69) is completely controllable if (6-68) is completely controllable.

The same arguments show that (6-68) is completely controllable if (6-69) is completely controllable.

Thus, according to Theorems 2-9 and 4-3, the controllability of (6-68) can be simply identified by the structure of $\underline{A}_{(1,1)}$ and $\underline{A}_{(1,2)}$ through the application of Corollary 4-1.

Lemma 6-6:

Consider $\underline{A}_1$ as defined in (6-32) with $\underline{A}$ given by (4-9) for $v \geq 2$, and assume an arbitrary scalar function $\varphi(\underline{X}_1)$ of class C_2 and $\varphi(\underline{0}) = 0$ with $\underline{X}_1$ defined by (6-28). Then for

$$-\left[\frac{\partial \varphi(\underline{X}_1)}{\partial \underline{X}_1}\right]^T \underline{A}_1 \underline{X} \tag{6-75}$$

to be positive semidefinite in R^n, $\varphi(\underline{X}_1)$ must be

(i) identically zero if $\underline{A}_1 \underline{X}$ is from a completely controllable system,

(ii) a function of only $\underline{X}_{(1)}$ defined by (4-116) if $\underline{A}_1 \underline{X}$ is from an uncontrollable system.

Proof: Define

$$\underline{\hat{w}}(\underline{X}_1) \triangleq \left[\frac{\partial \varphi(\underline{X}_1)}{\partial \underline{X}_1}\right] . \tag{6-76}$$

From Theorem 2-2,

$$\frac{\partial}{\partial \underline{X}_1} \{\underline{\hat{w}}(\underline{X}_1)\} \tag{6-77}$$

must be symmetric. Then, using (4-9),

$$-\hat{\underline{w}}^T(\underline{X}_1)\underline{A}_1\underline{X} = -[\hat{\underline{w}}^T_{(1)}(\underline{X}_1), \hat{\underline{w}}^T_{(2)}(\underline{X}_1), \ldots, \hat{\underline{w}}^T_{(v-1)}(\underline{X}_1)]$$

$$\begin{bmatrix} \underline{A}_{(1,1)} & \underline{A}_{(1,2)} & & & \\ & & \underline{A}_{(2,3)} & & \\ & & & \ddots & \\ & & & & \underline{A}_{(v-1,v)} \end{bmatrix} \begin{bmatrix} \underline{X}_{(1)} \\ \underline{X}_{(2)} \\ \vdots \\ \underline{X}_{(v)} \end{bmatrix}$$

$$= -\{\hat{\underline{w}}^T_{(1)}(\underline{X}_1)\underline{A}_{(1,1)}\underline{X}_{(1)} + \hat{\underline{w}}^T_{(1)}(\underline{X}_1)\underline{A}_{(1,2)}\underline{X}_{(2)}$$

$$+ \hat{\underline{w}}^T_{(3)}(\underline{X}_1)\underline{A}_{(2,3)}\underline{X}_{(3)} + \ldots + \hat{\underline{w}}^T_{(v-1)}(\underline{X}_1)\underline{A}_{(v-1,v)}\underline{X}_{(v)}\} , \quad (6\text{-}78)$$

where

$$\hat{\underline{w}}_{(i)}(\underline{X}_1) = [\hat{w}_{p_i+1}(\underline{X}_1), \hat{w}_{p_i+2}(\underline{X}_1), \ldots, \hat{w}_{p_i+\ell_i}(\underline{X}_1)]^T \quad (6\text{-}79)$$

with

$$p_i = \sum_{j=1}^{i-1} \ell_j . \quad (6\text{-}80)$$

(1) Assume v = 2. Then it follows, from (6-28) and (4-116), that

$$\underline{X}_1 = \underline{X}_{(1)}$$
$$\underline{X}_2 = \underline{X}_{(2)}, \tag{6-81}$$

If the system is completely controllable, $A_{(1,1)} = [0]$ from Corollary 4-1, and (6-78) becomes

$$-\hat{\underline{W}}^T_{(1)}(\underline{X}_{(1)})\underline{A}_{(1,2)}\underline{X}_{(2)} = -\hat{\underline{W}}^T_{(1)}(\underline{X}_1)\left[[0] \quad I_{\ell_1}\right]\underline{X}_{(2)} . \tag{6-82}$$

Therefore, if $\hat{\underline{W}}_{(1)}(\underline{X}_{(1)})$ is not identically zero, $\underline{X}_{(1)}$ can be selected to provide a nonzero value to $\hat{\underline{W}}_{(1)}(\underline{X}_{(1)})$. Then (6-82) becomes a linear function of $\underline{X}_{(2)}$ with nonzero coefficients and can have negative values for a proper choice of $\underline{X}_{(2)}$. Therefore $\hat{\underline{W}}_{(1)}(\underline{X}_{(1)})$ must be identically zero. As $\varphi(\underline{0}) = 0$, then $\varphi(\underline{X}_{(1)}) = 0$.

(2) Assume a general case of $v > 2$. Then for (6-78) to be positive semidefinite it similarly follows that $\hat{\underline{W}}_{(v-1)}(\underline{X}_{(1)})$ must be identically zero. According to the symmetry in (6-76), it must follow that

$$\left[\frac{\partial \hat{\underline{W}}_{(i)}(\underline{X}_1)}{\partial \underline{X}_{(v-1)}}\right] = \left[\frac{\partial \hat{\underline{W}}_{(v-1)}(\underline{X}_1)}{\partial \underline{X}_{(i)}}\right]^T = [0], \text{ for } i = 1,2,\ldots,v-1 . \tag{6-83}$$

Therefore, $\hat{\underline{W}}(\underline{X}_1)$ cannot be a function of $\underline{X}_{(v-1)}$ for the

positive semidefiniteness of (6-78).

A similar process can be repeated for each subvector $\underline{X}_{(v-1)}$, $\underline{X}_{(v-2)}$, ..., $\underline{X}_{(1)}$ succeedingly. Then it is concluded that, for (6-78) to be positive semidefinite, $\hat{\underline{w}}_{(i)}(\underline{X}_1)$, i = v-2,v-3,...,2, must be identically zero and $\hat{\underline{w}}_{(1)}(\underline{X}_1)$ must be a function of only $\underline{X}_{(1)}$, say $\hat{\underline{w}}_{(1)}(\underline{X}_{(1)})$. If the system is completely controllable, (6-78) reduces to

$$-\hat{\underline{w}}^T_{(1)}(\underline{X}_{(1)})\underline{A}_{(1,2)}\underline{X}_{(2)} \cdot \qquad (6\text{-}84)$$

Then applying the results of (1), $\hat{\underline{w}}_{(1)}(\underline{X}_{(1)})$ must be identically zero for the positive semidefiniteness of (6-75).

If the system is uncontrollable, the same argument follows, except, from Corollary 4-1, (6-78) reduces to

$$-\hat{\underline{w}}^T_{(1)}(\underline{X}_{(1)})\underline{A}_{(1,1)}\underline{X}_{(1)} \cdot \qquad (6\text{-}85)$$

6.4 Analysis of the Inverse Problem

6.4.1 Hamilton-Jacobi Equation

Assume that (a) a specific loss function is given as

$$L(\underline{X}) + \underline{U}^T \underline{R}\, \underline{U} \qquad (6\text{-}86)$$

for the inverse problem and (b) the resulting optimal performance index function $V^o(\underline{X})$ is of class C_2. The Hamilton-Jacobi equation becomes from (5-29) and (6-7),

$$0 = \min_{\underline{U}} \{ [\frac{\partial V^o}{\partial \underline{X}}]^T \underline{B}\ \underline{U} + \underline{U}^T \underline{R}\ \underline{U}\} + [\frac{\partial V^o}{\partial \underline{X}}]^T \underline{F}(\underline{X}) + L(\underline{X}) \ . \tag{6-87}$$

As $\underline{R}$ is positive definite by the assumption, the Hamiltonian is normal from Definition 5-1. Thus, its minimum at each $\underline{X} \in \underline{R}^n$ is uniquely provided by $\underline{U}$ satisfying

$$\underline{0} = \frac{\partial}{\partial \underline{U}} \{ [\frac{\partial V^o}{\partial \underline{X}}]^T \underline{B}\ \underline{U} + \underline{U}^T \underline{R}\ \underline{U}\}$$

$$= [\frac{\partial V^o}{\partial \underline{X}}]^T \underline{B} + 2\underline{U}^T \underline{R} \ , \tag{6-88}$$

from (5-33). Identifying this $\underline{U}$ in a closed form as a function of $\underline{X}$, the Hamilton-Jacobi equation is realized by a feedback control law such that

$$\underline{U}(\underline{X}) = -\frac{1}{2} \underline{R}^{-1} \underline{B}^T [\frac{\partial V^o}{\partial \underline{X}}] \ . \tag{6-89}$$

Define

$$\underline{\underline{R}}^{-1} \triangleq \underline{\underline{\tilde{R}}} \triangleq \begin{bmatrix} \underline{\underline{\tilde{R}}}_{11} & \underline{\underline{\tilde{R}}}_{12} \\ \underline{\underline{\tilde{R}}}_{12}^{T} & \underline{\underline{\tilde{R}}}_{22} \end{bmatrix} , \tag{6-90}$$

where the dimensions of each $\underline{\underline{\tilde{R}}}_{ij}$ are the same to those of $\underline{\underline{R}}_{ij}$ in (6-36). Then it follows,[32] that

$$\begin{aligned} \underline{\underline{\tilde{R}}}_{11} &= [\underline{\underline{R}}_{11} - \underline{\underline{R}}_{12}^{T}\underline{\underline{R}}_{22}^{-1}\underline{\underline{R}}_{12}^{T}]^{-1} , \\ \underline{\underline{\tilde{R}}}_{12} &= -\underline{\underline{R}}_{11}^{-1}\underline{\underline{R}}_{12} [\underline{\underline{R}}_{22} - \underline{\underline{R}}_{12}^{T}\underline{\underline{R}}_{11}^{-1}\underline{\underline{R}}_{12}]^{-1} , \\ \underline{\underline{\tilde{R}}}_{22} &= [\underline{\underline{R}}_{22} - \underline{\underline{R}}_{12}^{T}\underline{\underline{R}}_{11}^{-1}\underline{\underline{R}}_{12}]^{-1} . \end{aligned} \tag{6-91}$$

For $\underline{B}$ given by (4-99), (6-89) can be reduced to

$$\begin{aligned} \underline{U}_d(\underline{X}) &= -\frac{1}{2}\underline{\underline{\tilde{R}}}_{12} [\frac{\partial V^o}{\partial \underline{X}_2}] \\ \underline{U}_e(\underline{X}) &= -\frac{1}{2}\underline{\underline{\tilde{R}}}_{22} [\frac{\partial V^o}{\partial \underline{X}_2}] . \end{aligned} \tag{6-92}$$

Since $\underline{\underline{\tilde{R}}}_{22}$ must be nonsingular to insure the positive definiteness of $\underline{\underline{R}}$ from Lemma 6-2, it follows from (6-91) that

$$[\frac{\partial V^o}{\partial \underline{X}_2}] = -2\underline{\underline{\tilde{R}}}_{22}^{-1}\underline{U}_e(\underline{X}) = -2[\underline{\underline{R}}_{22} - \underline{\underline{R}}_{12}^{T}\underline{\underline{R}}_{11}^{-1}\underline{\underline{R}}_{12}]\underline{U}_e(\underline{X}) \tag{6-93}$$

and

$$\underline{U}_d(\underline{X}) = \widetilde{\underline{R}}_{12}\,\widetilde{\underline{R}}_{22}^{-1}\,\underline{U}_e(\underline{X}) = -\underline{R}_{11}^{-1}\,\underline{R}_{12}\,\underline{U}_e(\underline{X}) \ . \qquad (6\text{-}94)$$

Substituting (4-91), (6-30), (6-92), (6-93) and (6-94) into (6-89),

$$0 = \left[0,0,\ldots 0,\left[\frac{\partial \nabla^o}{\partial \underline{X}_2}\right]\right]\underline{U}(\underline{X}) + \underline{U}^T(\underline{X})\begin{bmatrix}\underline{R}_{11} & \underline{R}_{12}\\ \underline{R}_{12}^T & \underline{R}_{22}\end{bmatrix}\underline{U}(\underline{X})$$

$$+ \begin{bmatrix}\left[\frac{\partial \nabla^o}{\partial \underline{X}_1}\right]^T & \left[\frac{\partial \nabla^o}{\partial \underline{X}_2}\right]^T\end{bmatrix}\begin{bmatrix}\underline{F}_1(\underline{X})\\ \underline{F}_2(\underline{X})\end{bmatrix} + L(\underline{X}) \qquad (6\text{-}95)$$

or

$$L(\underline{X}) = \underline{U}_e^T(\underline{X})\,[\underline{R}_{22} - \underline{R}_{12}^T\underline{R}_{11}^{-1}\underline{R}_{12}]\,\underline{U}_e(\underline{X}) + 2\underline{U}_e^T(\underline{X})\,[\underline{R}_{22} - \underline{R}_{12}^T\underline{R}_{11}^{-1}\underline{R}_{12}]\,\underline{F}_2(\underline{X})$$

$$- \left[\frac{\partial \nabla^o}{\partial \underline{X}_1}\right]^T \underline{F}_1(\underline{X}) \ . \qquad (6\text{-}96)$$

6.4.2 Concerning $\nabla^o(\underline{X})$

Since $\nabla^o(\underline{X})$ is a scalar of class C_2, the functional matrix $\dfrac{\partial^2 \nabla^o(\underline{X})}{\partial \underline{X}\partial \underline{X}}$ is symmetric (by Theorem 2-2). From (6-38),

$$w_{ij}(\underline{X}) \triangleq \frac{\partial w_i(\underline{X})}{\partial x_j} = \frac{\partial w_j(\underline{X})}{\partial x_i} \triangleq w_{ji}(\underline{X}) \quad , \text{ for all } i,j = 1,2,\ldots,n, \tag{6-97}$$

and, with (6-92), it follows that

$$\{\frac{\partial^2 V^o}{\partial \underline{X}_1 \partial \underline{X}_1}\}^T = \{\frac{\partial^2 V^o}{\partial \underline{X}_1 \partial \underline{X}_1}\} \quad , \tag{6-98}$$

$$\{\frac{\partial^2 V^o}{\partial \underline{X}_2 \partial \underline{X}_1}\}^T = \{\frac{\partial^2 V^o}{\partial \underline{X}_1 \partial \underline{X}_2}\} = -2[\underline{R}_{22} - \underline{R}_{12}^T \underline{R}_{11}^{-1} \underline{R}_{12}] \; [\frac{\partial \underline{U}_e(\underline{X})}{\partial \underline{X}_1}] \tag{6-99}$$

and

$$\{\frac{\partial^2 V^o}{\partial \underline{X}_2 \partial \underline{X}_2}\}^T = \{\frac{\partial^2 V^o}{\partial \underline{X}_2 \partial \underline{X}_2}\} = -2[\underline{R}_{22} - \underline{R}_{12}^T \underline{R}_{11}^{-1} \underline{R}_{12}] \; [\frac{\partial \underline{U}_e(\underline{X})}{\partial \underline{X}_2}] \; . \tag{6-100}$$

From (2-10), $w_i(\underline{X})$ can be calculated from its gradient as

$$w_i(\underline{X}) = \int_o^{\underline{X}} [\frac{\partial w_i(\underline{X})}{\partial \underline{X}}]^T \, d\underline{X}$$

$$= \int_o^{x_1} w_{1i}(\gamma_1, 0, \ldots 0) d\gamma_1 + \int_o^{x_2} w_{2i}(x_1, \gamma_2, 0 \ldots 0) d\gamma_2 + \cdots$$

$$+ \int_o^{x_{n-r}} w_{n-ri}(x_1, x_2, \ldots, x_{n-r-1}, \gamma_{n-r}, 0 \ldots 0) d\gamma_{n-r}$$

$$+\int_0^{x_{n-r+1}} w_{n-r+1\,i}(x_1,x_2,\ldots x_{n-r},\gamma_{n-r+1},0,\ldots 0)\,d\gamma_{n-r+1} + \cdots$$

$$+\int_0^{x_n} w_{ni}(x_1,x_2,\ldots,x_{n-1},\gamma_n)\,d\gamma_n \,. \qquad (6\text{-}101)$$

Let conveniently describe the sum of the first (n-r) integrals (6-101) as $w_i(\underline{X}_1)$, a function of $\underline{X}_1$ only. From (6-97), the last r integrals can be calculated with $w_{ij}(\underline{X})$, instead of $w_{ji}(\underline{X})$. Thus, from (6-98),

$$\frac{\partial \nabla^o}{\partial \underline{X}_1} = \begin{bmatrix} w_1(\underline{X}) \\ w_2(\underline{X}) \\ \vdots \\ w_{n-r}(\underline{X}) \end{bmatrix} = \begin{bmatrix} \tilde{w}_1(\underline{X}_1) \\ \tilde{w}_2(\underline{X}_1) \\ \vdots \\ \tilde{w}_{n-r}(\underline{X}_1) \end{bmatrix}$$

$$+\int_0^{\underline{X}_2} \begin{bmatrix} w_{n-r+1\ 1}(\underline{X}), & w_{n-r+2\ 1}(\underline{X}), & \cdots & , w_{n1}(\underline{X}) \\ w_{n-r+1\ 2}(\underline{X}), & w_{n-r+2\ 2}(\underline{X}), & \cdots & , w_{n2}(\underline{X}) \\ \vdots & & & \vdots \\ w_{n-r+1\ n-r}(\underline{X}), & \cdots & \cdots & , w_{n\ n-r}(\underline{X}) \end{bmatrix} d\underline{X}_2$$

$$\triangleq \tilde{\underline{w}}(\underline{X}_1) + \int_0^{\underline{X}_2} \left\{\frac{\partial^2 \nabla^o}{\partial \underline{X}_1 \partial \underline{X}_2}\right\}^T d\underline{X}_2$$

$$= -2\int_{o}^{\underline{X}_2} [\frac{\partial \underline{U}_e(\underline{X})}{\partial \underline{X}_1}]^T [\underline{R}_{22}-\underline{R}_{12}^T\underline{R}_{11}^{-1}\underline{R}_{12}]d\underline{X}_2 + \tilde{\underline{w}}(\underline{X}_1) \ . \qquad (6\text{-}102)$$

Then the gradient of $V^o(\underline{X})$ becomes, from (6-93) and (6-102),

$$[\frac{\partial V^o}{\partial \underline{X}}] = \begin{bmatrix} \frac{\partial V^o}{\partial \underline{X}_1} \\ \vdots \\ \frac{\partial V^o}{\partial \underline{X}_n} \end{bmatrix}$$

$$= \begin{bmatrix} -2\int_{o}^{\underline{X}_2} [\frac{\partial \underline{U}_e(\underline{X})}{\partial \underline{X}_1}]^T [\underline{R}_{22}-\underline{R}_{12}^T\underline{R}_{11}^{-1}\underline{R}_{12}]d\underline{X}_2 + \tilde{\underline{w}}(\underline{X}_1) \\ \\ -2[\underline{R}_{22} - \underline{R}_{12}^T\underline{R}_{11}^{-1}\underline{R}_{12}]\underline{U}_e(\underline{X}) \end{bmatrix} \qquad (6\text{-}103)$$

and the Hamilton-Jacobi equation (6-96) becomes

$$L(\underline{X}) = \underline{U}_e^T(\underline{X})[\underline{R}_{22}-\underline{R}_{12}^T\underline{R}_{11}^{-1}\underline{R}_{12}]\underline{U}_e(\underline{X}) + 2\underline{U}_e^T[\underline{R}_{22}-\underline{R}_{12}^T\underline{R}_{11}^{-1}\underline{R}_{12}]\underline{F}_2(\underline{X})$$

$$+ 2\{\int_{o}^{\underline{X}_2} [\frac{\partial \underline{U}_e(\underline{X})}{\partial \underline{X}_1}]^T [\underline{R}_{22}-\underline{R}_{12}^T\underline{R}_{11}^{-1}\underline{R}_{12}]d\underline{X}_2\}^T \underline{F}_1(\underline{X}) - \tilde{\underline{w}}^T(\underline{X}_1)\underline{F}_1(\underline{X}) \ . \qquad (6\text{-}104)$$

6.4.3 Principal Theorem of the Inverse Problem

With the preceding results, the inverse problem can be investigated to determine combinations of $L(\underline{X})$ and $\underline{R}$ to satisfy the Hamilton-Jacobi equation (6-104). The fundamental result can be stated as the following.

Theorem 6-1: The Inverse Problem.

For the inverse problem as described in Section 6.2 such that the optimal performance index function is of class C_2, a performance index can be optimized by the given $\underline{U}(\underline{X})$ if and only if

$$\text{(i)} \quad \underline{U}_d(\underline{X}) = -\underline{R}_{11}^{-1}\underline{R}_{12}\underline{U}_e(\underline{X}) \ , \tag{6-105}$$

$$\text{(ii)} \quad [\underline{R}_{22}-\underline{R}_{12}^{T}\underline{R}_{11}^{-1}\underline{R}_{12}] \ [\frac{\partial \underline{U}_e(\underline{X})}{\partial \underline{X}_2}] \tag{6-106}$$

is symmetric, and

(iii) there exists an (n - r) dimensional vector valued function $\underline{w}(\underline{X}_1)$ of class C_1 and insuring symmetry in

$$-2 \ \frac{\partial}{\partial \underline{X}_1}\{\int_0^{\underline{X}_2} [\frac{\partial \underline{U}_e(\underline{X})}{\partial \underline{X}_1}]^T [\underline{R}_{22}-\underline{R}_{12}^{T}\underline{R}_{11}^{-1}\underline{R}_{12}]d\underline{X}_2\} + [\frac{\partial \underline{\tilde{w}}(\underline{X}_1)}{\partial \underline{X}_1}] \ . \tag{6-107}$$

The corresponding $V^o(\underline{X})$ is given from (6-103) as

$$\nabla^{o}(\underline{X}) = \int_{o}^{\underline{X}} [\frac{\partial \nabla^{o}}{\partial \underline{X}}]^{T} \, d\underline{X} \, . \qquad (6\text{-}108)$$

Proof: The necessity of the conditions has been shown by the previous work of the chapter. That is, (i) comes from the absolute minimum condition of the Hamiltonian (6-94). The symmetry in the functional matrix of (6-100) corresponds to (ii), and (iii) follows from the Hamilton-Jacobi equation and the symmetry in (6-98), using (6-102).

According to Theorem 5-1, the sufficiency of the conditions can be proved by showing that the existence of a unique function $\underline{\tilde{w}}(\underline{X}_1)$ satisfying the conditions of the theorem can exist for each combination of $L(\underline{X})$ and $\underline{R}$ as the Hamiltonian is normal. Assuming the contrary, that there exist two different functions, say $\underline{\tilde{w}}_a(\underline{X}_1)$ and $\underline{\tilde{w}}_b(\underline{X}_1)$, satisfying (iii) for a combination of $L(\underline{X})$ and $\underline{R}$. Necessarily, from (6-103) and (6-96),

$$\underline{\tilde{w}}_a^{T}(\underline{X}_1)\underline{F}_1(\underline{X}) = \underline{\tilde{w}}_b^{T}(\underline{X}_1)\underline{F}_1(\underline{X}) \, . \qquad (6\text{-}109)$$

Describe the resulting optimal performance index functions as $\nabla_a^{o}(\underline{X})$ and $\nabla_b^{o}(\underline{X})$. Then from (6-103) and (6-108), it follows that

$$\nabla_b^o(\underline{X}) = \nabla_a^o(\underline{X}) + \varphi(\underline{X}_1) \ , \qquad (6\text{-}110)$$

where $\varphi(\underline{X}_1)$ is such that

$$\varphi(\underline{X}_1) = \int_o^{\underline{X}_1} \{-\underline{\tilde{w}}_a(\underline{X}_1) + \underline{\tilde{w}}_b(\underline{X}_1)\}^T \, d\underline{X}_1 \ . \qquad (6\text{-}111)$$

This is not identically zero by the contrary hypothesis. However $\varphi(\underline{0}) = 0$ as $\nabla_a^o(\underline{0}) = \nabla_b^o(\underline{0}) = 0$. Thus there exists a specific $(n - r)$ dimensional vector $\underline{X} = [\underline{X}_1, \underline{X}_2, \cdots \underline{X}_{n-r}]^T$ satisfying

$$\varphi(\underline{X}_1) = k, \text{ a nonzero constant} \ . \qquad (6\text{-}112)$$

Consider the hypersurface $\varphi(\underline{X}_1) = k$ in $R^n \times R^1$ which does not include $\underline{X} = \underline{0}$. However, consider the time derivative of $\varphi(\underline{X}_1)$ governed by the synthesized system equation. This follows from (6-111), (6-34), and (6-109),

$$\dot{\varphi}(\underline{X}_1) = \left[\frac{\partial \varphi(\underline{X}_1)}{\partial \underline{X}}\right]^T \{\underline{F}(\underline{X}) + \underline{B}\,\underline{U}(\underline{X})\}$$

$$= \left[\left[-\underline{\tilde{w}}_a(\underline{X}_1) + \underline{\tilde{w}}_b(\underline{X}_1)\right]^T \quad \underline{0}\right] \begin{bmatrix} \underline{F}_1(\underline{X}) \\ \underline{F}_2(\underline{X}) + \underline{U}_e(\underline{X}) \end{bmatrix}$$

$$= \left[-\underline{\tilde{w}}_a(\underline{X}_1) + \underline{\tilde{w}}_b(\underline{X}_1)\right]^T \underline{F}_1(\underline{X}) = 0 \qquad (6\text{-}113)$$

from (6-109). This implies that every motion of the synthesized system from points on $\varphi(\underline{X}_1) = k$ stays on this hypersurface and can never approach to origin. This contradicts the asymptotic stability in the large of the origin of the synthesized system, Lemma 6-1 and assumption (ii) of Section 6.1.

6.5 Discussion

6.5.1 On the General Method of Solution of the Inverse Problem

A solution to the inverse problem is obtained by determining all combinations of $L(\underline{X})$ and $\underline{R}$ satisfying the conditions of Theorem 6-1. $\underline{R}$ must be determined to meet conditions (i) and (ii) with respect to the given $\underline{U}(\underline{X})$. The corresponding $L(\underline{X})$ are then determined from (6-104) by choosing various $\underline{\tilde{w}}(\underline{X}_1)$ satisfying condition (iii). If no positive definite symmetric $\underline{R}$ exists for the given $\underline{U}(\underline{X})$, then the $\underline{U}(\underline{X})$ cannot be an optimal control law.

6.5.2 Dependency within $\underline{U}(\underline{X})$

As $\underline{U}_d(\underline{X})$ is dependent upon $\underline{U}_e(\underline{X})$, (6-105), there can exist at most r independent elements in an optimal feedback control law, where r is the rank of $\hat{\underline{B}}$. If m = r, i.e., $\hat{\underline{B}}$ is full rank, condition (i) of the theorem is nonexistent because $\underline{U}_d(\underline{X})$ is reduced to dimension zero.

In Thau's problem statement, Section 5.3.3, the rank of $\hat{\underline{B}}$ is not mentioned. If it is assumed to be either full rank or less than full rank, an additional condition corresponding to (6-105) must be given.

6.5.3 Consideration of the Variety of $L(\underline{X})$

For an $\underline{R}$ satisfying Theorem 6-1, a variety of $L(\underline{X})$ may exist for which the given $\underline{U}(\underline{X})$ is an optimal feedback control law. These are associatedwith various $\tilde{\underline{w}}(\underline{X}_1)$.

Corollary 6-1:

For an $\underline{R}$ satisfying Theorem 6-1, assume an $L_a(\underline{X})$ can be found from (6-104) as an optimized performance index for the given $\underline{U}(\underline{X})$. Let the resulting optimal performance index function be $V_a^o(\underline{X})$. Then an $L(\underline{X})$ can also be found from (6-104) for $\underline{R}$ if and only if

$$L(\underline{X}) = L_a(\underline{X}) - \left[\frac{\partial \varphi(\underline{X}_1)}{\partial \underline{X}_1}\right]^T \underline{F}_1(\underline{X}) \ , \qquad (6\text{-}114)$$

where $\varphi(\underline{X}_1)$ is an arbitrary scalar function of class C_2 and $\varphi(\underline{0}) = 0$. Necessarily, the resulting optimal performance index function is

$$V^o(\underline{X}) = V_a^o(\underline{X}) + \varphi(\underline{X}_1) \ . \qquad (6\text{-}115)$$

Proof: $L_a(\underline{X})$ corresponds to $\tilde{\underline{w}}_a(\underline{X}_1)$ in (6-104). Then from (6-96), it follows that

$$\tilde{\underline{w}}(\underline{X}_1) = \tilde{\underline{w}}_a(\underline{X}_1) + \left[\frac{\partial \varphi(\underline{X}_1)}{\partial \underline{X}_1}\right] \ , \qquad (6\text{-}116)$$

and the assertion follows directly from Theorem 6-1.

Possibilities for a nonnegative $L(\underline{X})$ can be considered by trying various $\varphi(\underline{X}_1)$ in (6-114). In the next chapter, the nonnegative property of $L(\underline{X})$ is examined in some detail under additional problem assumptions.

6.5.4 Uniqueness of $L(\underline{X})$

Consider the case of $n = m = r$ in (6-2), (6-28) and (6-29). Then the dimension of $\underline{X}_1$ is zero and the general method of solution stated in Section 6.5.1 can be simplified.

Condition (iii) of Theorem 6-1 is nonexistent because $\underline{\tilde{w}}(\underline{X}_1)$ is reduced to zero dimension. Thus, a corollary follows directly from Theorem 6-1.

Corollary 6-2:

For the inverse problem, if n = m = r, then a unique $L(\underline{X})$ corresponds to each $\underline{R}$ and, from (6-104),

$$L(\underline{X}) = \underline{U}^T(\underline{X})\,\underline{R}\,\underline{U}(\underline{X}) + 2\underline{U}^T(\underline{X})\,\underline{R}\,\underline{F}(\underline{X}) \ . \qquad (6\text{-}117)$$

6.5.5 Linear Control Law

Let $\underline{U}(\underline{X})$ be specified as a linear feedback control law,

$$\underline{U}(\underline{X}) = -\underline{K}^T\underline{X} \ , \qquad (6\text{-}118)$$

where $\underline{K}$ as an n x m matrix decomposed as

$$\underline{K} \triangleq \begin{bmatrix} \underline{K}_{11} & \underline{K}_{12} \\ \underline{K}_{21} & \underline{K}_{22} \\ \underbrace{\quad}_{m-r} & \underbrace{\quad}_{r} \end{bmatrix} \begin{matrix} \} n-r \\ \} r \end{matrix} \ . \qquad (6\text{-}119)$$

Thus by (6-29), (6-118) is

$$\underline{U}_d(\underline{X}) = -\underline{K}_{11}^T\underline{X}_1 - \underline{K}_{21}^T\underline{X}_2$$

$$\underline{U}_e(\underline{X}) = -\underline{K}_{12}^T\underline{X}_1 - \underline{K}_{22}^T\underline{X}_2 \qquad (6\text{-}120)$$

and

$$\frac{\partial \underline{U}_e(\underline{X})}{\partial \underline{X}_1} = -\underline{K}_{12}^T$$

$$\frac{\partial \underline{U}_e(\underline{X})}{\partial \underline{X}_2} = -\underline{K}_{22}^T \qquad (6\text{-}121)$$

The conditions in Theorem 6-1 can be identified directly with these submatrices.

Assuming a linear feedback control law (6-118), (6-104) becomes

$$L(\underline{X}) = \underline{X}^T\begin{bmatrix}\underline{K}_{12}\\ \underline{K}_{22}\end{bmatrix}[\underline{R}_{22}-\underline{R}_{12}^T\underline{R}_{11}^{-1}\underline{R}_{12}][\underline{K}_{12}^T \quad \underline{K}_{22}^T]\underline{X}$$

$$- 2\underline{X}^T\begin{bmatrix}\underline{K}_{12}\\ \underline{K}_{22}\end{bmatrix}[\underline{R}_{22}-\underline{R}_{12}^T\underline{R}_{11}^{-1}\underline{R}_{12}]\underline{F}_2(\underline{X})$$

$$- 2\underline{X}_2^T[\underline{R}_{22}-\underline{R}_{12}^T\underline{R}_{11}^{-1}\underline{R}_{12}]\underline{K}_{12}^T\underline{F}_1(\underline{X}) - \tilde{\underline{W}}^T(\underline{X}_1)\underline{F}_1(\underline{X}) \; . \qquad (6\text{-}122)$$

For the equivalent statement of condition (i), by (6-105), it can be shown from (6-120) that

(i)'

$$-\underline{K}_{11}^{T} = \underline{R}_{11}^{-1}\underline{R}_{12}\underline{K}_{12}^{T}$$

$$-\underline{K}_{21}^{T} = \underline{R}_{11}^{-1}\underline{R}_{12}\underline{K}_{22}^{T} \cdot \qquad (6\text{-}123)$$

Conditions (ii) and (iii) follow directly, i.e.,

(ii)'

$$-[\underline{R}_{22} - \underline{R}_{12}^{T}\underline{R}_{11}^{-1}\underline{R}_{12}]\underline{K}_{22}^{T} \qquad (6\text{-}124)$$

is symmetric, and

(iii)'

$$[\frac{\partial \underline{\tilde{w}}(\underline{X}_1)}{\partial \underline{X}_1}] \qquad (6\text{-}125)$$

is symmetric.

From (6-118) and (6-93), it follows that

$$\frac{\partial V^{o}}{\partial \underline{X}_2} = 2[\underline{R}_{22} - \underline{R}_{12}^{T}\underline{R}_{11}^{-1}\underline{R}_{12}]\ [\underline{K}_{12}^{T}\ \ \underline{K}_{22}^{T}]\underline{X} \cdot \qquad (6\text{-}126)$$

Therefore

$$\frac{\partial^2 V^o}{\partial \underline{X}_1 \partial \underline{X}_2} = \{\frac{\partial^2 V^o}{\partial \underline{X}_2 \partial \underline{X}_1}\}^T = 2[\underline{R}_{22} - \underline{R}_{12}^T \underline{R}_{11}^{-1} \underline{R}_{12}]\underline{K}_{12}^T$$

$$\frac{\partial^2 V^o}{\partial \underline{X}_2 \partial \underline{X}_2} = 2[\underline{R}_{22} - \underline{R}_{12}^T \underline{R}_{11}^{-1} \underline{R}_{12}]\underline{K}_{22}^T \qquad (6\text{-}127)$$

$$\frac{\partial^2 V^o}{\partial \underline{X}_1 \partial \underline{X}_1} = \frac{\partial \underline{\tilde{w}}(\underline{X}_1)}{\partial \underline{X}_1} \quad .$$

Consequently, it follows that

$$V^o(\underline{X}) = \underline{X}^T \begin{bmatrix} [0] & \underline{K}_{12}[\underline{R}_{22} - \underline{R}_{12}^T \underline{R}_{11}^{-1} \underline{R}_{12}] \\ [\underline{R}_{22} - \underline{R}_{12}^T \underline{R}_{11}^{-1} \underline{R}_{12}]\underline{K}_{12}^T & [\underline{R}_{22} - \underline{R}_{12}^T \underline{R}_{11}^{-1} \underline{R}_{12}]\underline{K}_{22}^T \end{bmatrix} \underline{X} + \varphi(\underline{X}_1) \quad , \qquad (6\text{-}128)$$

where

$$\varphi(\underline{X}_1) = \int_o^{\underline{X}_1} \underline{\tilde{w}}^T(\underline{X}_1) d\underline{X}_1 \quad . \qquad (6\text{-}129)$$

From (6-128), the structure of $V^o(\underline{X})$ is a sum of an arbitrary scalar function $\varphi(\underline{X}_1)$ and a quadratic form in $\underline{X}$ determined by $\underline{K}$ and $\underline{R}$. If $n = m = r$, then only one $V^o(\underline{X})$, a quadratic form, can exist for each $\underline{R}$, according to Corollary 6-2.

Considering the results of Suga for a linear control law, Section 5.3.2, from the viewpoint of the above results, the symmetry in (6-124) corresponds to that of $\hat{\underline{R}}(t)\ \hat{\underline{K}}(t)\ \hat{\underline{R}}(t)$ in Theorem 5-3. It is interesting to note that the structure of $\hat{V}^{o}(\underline{Y}, t)$ in (5-56), i.e., a quadratic in $\underline{Y}$ determined by the given $\underline{V}(\underline{Y})$ plus an arbitrary function, is invariant for the general nonlinear system, (6-128). Thus the flexibility of $L(\underline{X})$ due to the function $\tilde{\underline{W}}(\underline{X}_1)$ corresponds to that of $\hat{L}(\underline{Y}, t)$ due to $\hat{\Gamma}(\underline{Y}, t)$ in (5-49).

6.5.6 Nonnegative $\nabla^{o}(\underline{X})$ for a Linear Control Law

For a linear control law, a definitive statement is possible for the sign definiteness of $\nabla^{o}(\underline{X})$.

Theorem 6-2:

For a linear feedback control law (6-118) in the inverse problem, the resulting $\nabla^{o}(\underline{X})$, (6-128), is positive semi-definite in R^n if and only if

(i) the last $r - r_1$ columns of $\underline{R}_{12}\underline{D}_{22}$ are null, r_1 the rank of $[\underline{R}_{22} - \underline{R}_{12}^{T}\underline{R}_{11}^{-1}\underline{R}_{12}]$, and

(ii) a function

$$\underline{x}_1 \tilde{\underline{\Omega}} \underline{x}_1 + \varphi(\underline{x}_1) \qquad (6\text{-}130)$$

is positive semidefinite in R^{n-r}, where

$$\tilde{\underline{\Omega}} \triangleq -\underline{K}_{12}\underline{R}_o\underline{D}_{22}\underline{D}_{22}^T\underline{R}_o\underline{K}_{22}^T\underline{D}_{22}\underline{D}_{22}^T\underline{R}_o\underline{K}_{12}^T , \qquad (6\text{-}131)$$

$$\underline{R}_o \triangleq [\underline{R}_{22} - \underline{R}_{12}^T\underline{R}_{11}^{-1}\underline{R}_{12}] , \qquad (6\text{-}132)$$

and $\underline{D}_{22}$ is an r x r nonsingular matrix for the congruent transformation

$$\underline{D}_{22}^T[\underline{R}_{22} - \underline{R}_{12}^T\underline{R}_{11}^{-1}\underline{R}_{12}]\underline{K}_{22}^T\underline{D}_{22} = \begin{bmatrix} I_{r_1} & [0] \\ [0] & [0] \end{bmatrix} . \qquad (6\text{-}133)$$

Proof: For (6-128) to be semidefinite, $\underline{R}_o$ must be positive semidefinite, from

$$\nabla^o([\underline{0}^T, \ \underline{x}_2^T]) = \underline{x}_2^T\underline{R}_o\underline{K}_{22}^T\underline{x}_2 . \qquad (6\text{-}134)$$

Then the matrix $\underline{D}_{22}$ satisfying (6-133) exists, from Theorem 2-4. Define a nonsingular matrix

$$\underline{D} \triangleq \begin{bmatrix} I_{n-r} & [0] \\ -\underline{\Delta} & \underline{D}_{22} \end{bmatrix} \qquad (6\text{-}135)$$

and a function

$$\tilde{\nabla}^{o}(\underline{X}) = \nabla^{o}(\underline{D}\ \underline{X}) \ , \tag{6-136}$$

where

$$\underline{\Delta} = \underline{D}_{22}\underline{D}_{12}^{T}\underline{R}_{o}\underline{K}_{22}^{T}\underline{D}_{22}\underline{D}_{22}^{T}\underline{R}_{o}\underline{K}_{12}^{T} \ . \tag{6-137}$$

Then

(a) $\tilde{\nabla}^{o}(\underline{0}) = \nabla^{o}(\underline{0}) = 0$,

(b) if $\tilde{\nabla}^{o}(\underline{X})$ is positive semidefinite, then

$$\nabla^{o}(\underline{X}) = \tilde{\nabla}^{o}(\underline{D}^{-1}\underline{X}) \geq 0 \ , \text{ for all } \underline{X} \in \underline{R}^{n} \ ,$$

and (c) if $\nabla^{o}(\underline{X})$ is positive semidefinite, then

$$\tilde{\nabla}^{o}(\underline{X}) = \nabla^{o}(\underline{D}\ \underline{X}) \geq 0 \ , \text{ for all } \underline{X} \in R^{n} \ .$$

Thus $\nabla^{o}(\underline{X})$ is positive semidefinite if and only if $\tilde{\nabla}^{o}(\underline{X})$ is positive semidefinite. Therefore, from (6-128), it follows that

$$V^{o}(D\,\underline{X}) = \underline{X}^{T}\begin{bmatrix} \underline{\tilde{\Omega}} & [0] & \underline{\mathcal{R}} \\ [0] & I_{r_1} & [0] \\ \underline{\mathcal{R}}^{T} & [0] & [0] \end{bmatrix}\underline{X} + \varphi(\underline{X}_1)$$

$$= \underline{X}_1^{T}\,\underline{\tilde{\Omega}}\,\underline{X}_1 + 2\underline{X}_1^{T}\begin{bmatrix}[0] & \underline{\mathcal{R}}\end{bmatrix}\underline{X}_2$$

$$+ \underline{X}_2^{T}\begin{bmatrix} I_{r_1} & [0] \\ [0] & [0] \end{bmatrix}\underline{X}_2 + \varphi(\underline{X}_1)\,, \qquad (6\text{-}138)$$

where $\underline{\mathcal{R}}$ is defined by the last $r - r_1$ column of $K_{12}\underline{R}_o D_{22}$. Applying a similar argument as in Lemmas 6-3 for (6-138), the necessity of the conditions can be justified. The sufficiency is apparent from (6-138) if $\underline{\mathcal{R}} = [\underline{0}]$.

Theorem 6-3:

For a linear feedback control law (6-118) in the inverse problem, the resulting $V^{o}(\underline{X})$, (6-128), is positive definite if and only if

(i) $\underline{R}_o K_{22}^{T}$ is positive definite, and

$$\text{(ii)} \quad -\underline{X}_1^T \underline{K}_{12} \underline{K}_{22}^{-1} \underline{R}_o \underline{K}_{12}^T \underline{X}_1 + \varphi(\underline{X}_1) \tag{6-139}$$

is positive definite in R^{n-r} and $\underline{R}_o = [\underline{R}_{22} - \underline{R}_{12}^T \underline{R}_{11}^{-1} \underline{R}_{12}]$.

Proof: The necessity of (i) follows directly from the positive definiteness of $V^o(\underline{X})$ for $\underline{X}_1 = \underline{0}$ or

$$V^o([\underline{0}^T, \underline{X}_2^T]) = \underline{X}_2^T \underline{R}_o \underline{K}_{22}^T \underline{X}_2 \, . \tag{6-140}$$

For the necessity of (ii), define

$$\underline{D}_1 \triangleq \begin{bmatrix} I_{n-r} & [0] \\ -(\underline{K}_{22}^{-1})^T \underline{K}_{12}^T & I_r \end{bmatrix} . \tag{6-141}$$

The inverse of $\underline{R}_o \underline{K}_{22}^T$ exists and is symmetric by (6-124). Then with the same argument as used in the proof of Theorem 6-2, $V^o(\underline{X})$ is positive definite if and only if $V^o(\underline{D}_1 \underline{X})$ is positive definite. It follows that

$$V^o(\underline{D}_1 \underline{X}) = \underline{X}^T \begin{bmatrix} -\underline{K}_{12} \underline{K}_{22}^{-1} \underline{R}_o \underline{K}_{12}^T & [0] \\ [0] & \underline{R}_o \underline{K}_{22}^T \end{bmatrix} \underline{X} + \varphi(\underline{X}_1)$$

$$= -\underline{X}_1^T \underline{R}_{12} \underline{K}_{22}^{-1} \underline{R}_o \underline{K}_{12}^T \underline{X}_1 + \varphi(\underline{X}_1) + \underline{X}_2^T \underline{R}_o \underline{K}_{22}^T \underline{X}_2 \, , \tag{6-142}$$

and (ii) is necessary. The sufficiency of the conditions follows from (6-142).

The nonnegative characteristic of $\nabla^{o}(\underline{X})$ is completely established by the above analysis. This is in contrast to Thau's partial consideration of the topic, as discussed in Section 5.3.3.

6.5.7 Necessity of Control Action

The problem of whether optimal control can exist for "No control action" for a nonnegative $L(\underline{X})$ in the performance index can be considered through the corresponding inverse problem. For the class of systems given by (6-68), let $\underline{U}(\underline{X})$ be identically zero, that is, the equation of the synthesized feedback control system is

$$\dot{\underline{X}} = \underline{A}\,\underline{X} + \underline{B}_{e}\,\underline{F}_{2}(\underline{X}) \; . \qquad (6\text{-}143)$$

Also assume that the origin of this system is asymptotically stable in the large. For $v \geq 2$, substituting $\underline{U}(\underline{X}) = \underline{0}$ into (6-104), $L(\underline{X})$ for the optimized performance index is

$$L(\underline{X}) = -\,\underline{\tilde{w}}^{T}(\underline{X}_{1})\;\underline{A}_{1}\;\underline{X} \; . \qquad (6\text{-}144)$$

Then, from (6-136) and Lemma 6-6, for (6-144) to be positive semidefinite,

(a) $\underline{\tilde{w}}(\underline{X}_1)$ must be identically zero, if the given system is completely controllable, or

(b) $\underline{\tilde{w}}(\underline{X}_1)$ must be a function of $\underline{X}_{(1)}$ only, if the given system is uncontrollable, say $\underline{\tilde{w}}(\underline{X}_{(1)})$.

Subsequently, (6-144) reduces to

$$L(\underline{X}) = \begin{cases} 0, \text{ if the given system is completely controllable,} \\ L(\underline{X}_{(1)}), \text{ if the given system is uncontrollable,} \\ \quad \text{where the state variables in } \underline{X}_{(1)} \text{ are the} \\ \quad \text{uncontrollable state variables, as discussed} \\ \quad \text{in Section 4.3.3.} \end{cases}$$

If $v = 1$, then $[\frac{\partial V}{\partial \underline{X}}]^{o} = [\frac{\partial V}{\partial \underline{X}_2}]^{o} = \underline{0}$ from (6-103), and $L(\underline{X})$ is identically zero.

Observing these results from the viewpoint of the forward problem, an important characteristic of an optimal feedback control system is evident.

Principle of Necessary Control Action:

Consider the optimal regulator problem such that

(i) the system is given as (6-68),

(ii) the desired final condition of the system is $\underline{X} = \underline{0}$,

(iii) the performance index is

$$\int_0^\infty \{L(\underline{X}) + \underline{U}^T \underline{R}\ \underline{U}\}dt \ ,$$

where $\underline{R}$ is positive definite, and $L(\underline{X})$ is a function of the controllable state variables and is positive semi-definite, then some control action is necessary for optimality, i.e., the optimal feedback control law cannot be identically zero.

6.5.8 Asymptotic Stability of the Synthesized Feedback Control System

The feedback control law $\underline{U}(\underline{X})$ by assumption (ii) of Section 6.1 requires that the origin of the synthesized system be asymptotically stable in the large. No definitive criterion exists to verify this, except for linearly synthesized systems. Practically, if $\underline{R}$, $L(\underline{X})$ and $\nabla^o(\underline{X})$ are

determined according to the procedure of Section 6.5.1 for some given $\underline{U}(\underline{X})$, it is then necessary to check the synthesized system for asymptotic stability. It may be possible to do this by applying the Liapunov direct method.

Theorem 6-4:

Let $\underline{R}$, $L_a(\underline{X})$ and $\nabla_a^o(\underline{X})$ be calculated for some $\underline{U}(\underline{X})$, following the procedure of Section 6.5.1. If there is a scalar function $\tilde{\varphi}(\underline{X}_1)$ such that it is of class C_2 satisfying

$$\nabla^o(\underline{X}) \triangleq \nabla_a^o(\underline{X}) + \tilde{\varphi}(\underline{X}_1) \ , \qquad (6\text{-}145)$$

and $\nabla^o(\underline{X})$ and

$$L_a(\underline{X}) + \underline{U}^T(\underline{X})\,\underline{R}\,\underline{U}(\underline{X}) - \left[\frac{\partial\, \tilde{\varphi}(\underline{X}_1)}{\partial \underline{X}_1}\right]^T \underline{F}_1(\underline{X}) \qquad (6\text{-}146)$$

are positive definite, then the origin of the synthesized system is asymptotically stable in the large.

Proof: From Corollary 6-1, $L(\underline{X})$ and $\nabla^o(\underline{X})$ in (6-114) and (6-115) can be determined as a function of $\underline{R}$ and $\underline{U}(\underline{X})$. Since the time derivative of (6-145) as governed by the synthesized system is

$$\dot{V}^{o}(\underline{X}) = \frac{d}{dt}\{V_a^{o}(\underline{X}) + \tilde{\varphi}(\underline{X}_1)\} = -L_a(\underline{X}) - \underline{U}^{T}(\underline{X})\,\underline{R}\,\underline{U}(\underline{X})$$

$$+ \left[\frac{\partial \tilde{\varphi}(\underline{X}_1)}{\partial \underline{X}_1}\right]^{T} \underline{F}_1(\underline{X}) \ , \qquad (6\text{-}147)$$

using (6-103), the proof follows directly from Theorem 2-8.

This theorem provides a sufficient condition only; accordingly a failure of the conditions does not necessarily mean that the origin of the synthesized system is not asymptotically stable in the large.

6.5.9 Miscellaneous Comments

As a generalization of the inverse problem first considered by Kalman (Chapter 5), Theorem 6-1 and the succeeding developments of this chapter are shown to include results of other authors. In addition, this work reveals new important characteristics of optimal feedback control systems, i.e., Sections 6.5.2 and 6.5.7. Moreover these results are presented very compactly as a result of the developed canonical form of Chapter 4.

6.6 Examples

Example 6-1: Consider a system given by

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\\ \dot{x}_3\end{bmatrix} = \begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 0\\ 0 & 0 & -2\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix} + \begin{bmatrix}0\\ 0\\ -x_3^3\end{bmatrix} + \begin{bmatrix}0 & 0\\ 1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}u_1\\ u_2\end{bmatrix} \qquad (6\text{-}148)$$

and assume a feedback control law

$$\begin{bmatrix}u_1(\underline{X})\\ u_2(\underline{X})\end{bmatrix} = \begin{bmatrix}-1 & -2 & 0\\ 0 & 0 & -1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix} + \begin{bmatrix}0\\ -x_3^3\end{bmatrix} . \qquad (6\text{-}149)$$

As n = 3 and m = r = 2, it follows for the canonical form, from (6-28∿30), that

$$\begin{aligned}\underline{X}_1 &= x_1 ,\\ \underline{X}_2 &= [x_2, x_3]^T ,\end{aligned} \qquad (6\text{-}150)$$

$$\begin{aligned}\underline{U}_d(\underline{X}) &\quad \text{nonexisting} ,\\ \underline{U}_e(\underline{X}) &= \underline{U}(\underline{X}) ,\end{aligned} \qquad (6\text{-}151)$$

$$\underline{F}_1(\underline{X}) = x_2 \ ,$$

and

$$\underline{F}_2(\underline{X}) = \begin{bmatrix} 0 \\ -2x_3 - x_3^3 \end{bmatrix} \ . \qquad (6\text{-}152)$$

Thus,

$$\left[\frac{\partial \underline{U}_e}{\partial \underline{X}_1}\right] = \left[\frac{\partial u_1(\underline{X})}{\partial x_1} \quad \frac{\partial u_2(\underline{X})}{\partial x_1}\right]^T = [-1 \quad 0]^T \qquad (6\text{-}153)$$

and

$$\left[\frac{\partial \underline{U}_e}{\partial \underline{X}_2}\right] = \begin{bmatrix} \dfrac{\partial u_1(\underline{X})}{\partial x_2} & \dfrac{\partial u_1(\underline{X})}{\partial x_3} \\ \dfrac{\partial u_2(\underline{X})}{\partial x_2} & \dfrac{\partial u_2(\underline{X})}{\partial x_3} \end{bmatrix} = \begin{bmatrix} -2 & 0 \\ 0 & -1-3x_3^2 \end{bmatrix} . \qquad (6\text{-}154)$$

Since $\underline{X}_1$ is one dimensional, define

$$\underline{\tilde{w}}(\underline{X}_1) \triangleq \tilde{w}(x_1) \ . \qquad (6\text{-}155)$$

Substituting (6-149~155) into (6-103) and (6-104),

$$\frac{\partial \nabla^o}{\partial \underline{X}} = \begin{bmatrix} \dfrac{\partial \nabla^o}{\partial x_1} \\ \dfrac{\partial \nabla^o}{\partial x_2} \\ \dfrac{\partial \nabla^o}{\partial x_3} \end{bmatrix} = \begin{bmatrix} 2\displaystyle\int_o^{\underline{X}_2} [1, \ 0]\underline{R} \ d\underline{X}_2 + \tilde{w}(x_1) \\ 2\underline{R} \begin{bmatrix} x_1 + 2x_2 \\ x_3 + x_3^3 \end{bmatrix} \end{bmatrix} \qquad (6\text{-}156)$$

and

$$L(\underline{X}) = [x_1 + 2x_2,\ x_3 + x_3^3]\underline{R}\begin{bmatrix} x_1 + 2x_2 \\ 5x_3 + 3x_3^3 \end{bmatrix}$$

$$- 2\{\int_0^{\underline{X}_2} [1,\ 0]\underline{R}\ d\underline{X}_2\}\ x_2 - \{\tilde{w}(x_1)\}x_2\ , \qquad (6\text{-}157)$$

where $\underline{R} = \underline{R}_{22} - \underline{R}_{12}\underline{R}_{11}^{-1}\underline{R}_{12}^{T} = \underline{R}_{22}$ because $\underline{B}$ is of full rank in (6-148). Referring to the statement of Theorem 6-1, (i) is nonexistent and the symmetry for (6-108) is satisfied since $\underline{X}_1$ is one dimensional. Consequently, for the system (6-148), a performance index

$$\int_{t_o}^{\infty} \{L(\underline{X}) + \underline{U}^T\underline{R}\ \underline{U}\}\ dt \qquad (6\text{-}158)$$

can be optimized by the feedback control law (6-149) if and only if symmetry exists for

$$\underline{R}\begin{bmatrix} -2 & 0 \\ 0 & -1-3x_3^2 \end{bmatrix}, \qquad (6\text{-}159)$$

and $\tilde{w}(x_1)$ is of class C_1.

Arbitrarily choose $\tilde{w}(x_1) = 2x_1$ and

$$\underline{R} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} , \qquad (6\text{-}160)$$

which satisfy the above conditions. Then (6-156 and 157) become

$$[\frac{\partial V^o}{\partial \underline{X}}] = \begin{bmatrix} 2x_1 + 2x_2 \\ 2x_1 + 4x_2 \\ 2x_3 + 2x_3^3 \end{bmatrix} \qquad (6\text{-}161)$$

and

$$L(\underline{X}) = x_1^2 + 2x_1x_2 + 2x_2^2 + 5x_3^2 + 8x_3^4 + 3x_3^6$$

$$= (x_1+x_2)^2 + x_2^2 + x_3^2(5+8x_3^2+3x_3^4) \quad . \qquad (6\text{-}162)$$

From (6-161) and (6-108), the optimal performance index function is

$$V^o(\underline{X}) = x_1^2 + 2x_1x_2 + 2x_2^2 + x_3^2 + \frac{x_3^4}{2}$$

$$= (x_1+x_2)^2 + x_2^2 + x_3^2 + \frac{x_3^4}{2} \quad . \qquad (6\text{-}163)$$

From Corollary 6-1, the performance indices from (6-114), that is

$$\int_{t_o}^{\infty} \{(x_1+x_2)^2 + x_2^2 + x_3^2(5+8x_3^2+3x_3^4) - \{\frac{\partial \tilde{\varphi}(x_1)}{\partial x_1}\}x_2 + u_1^2 + u_2^2\}dt, \tag{6-164}$$

can be optimized by the feedback control law (6-149). The optimal performance index function from (6-115) then becomes

$$V^o(\underline{X}) = (x_1+x_2)^2 + x_2^2 + x_3^2 + \frac{x_3^4}{2} + \tilde{\varphi}(x_1) , \tag{6-165}$$

where $\tilde{\varphi}(x_1)$ is any function of class C_2, e.g., $x_1 + x_1^3$.

To check the asymptotic stability of the synthesized system, combine (6-148) and (6-149) for

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} x_2 \\ -x_1 - 2x_2 \\ -3x_3 - 2x_3^3 \end{bmatrix} . \tag{6-166}$$

Then $V^o(\underline{X})$ for $\tilde{\varphi}(x_1) = 0$ is positive definite in R^n and it follows, from (6-164), that

$$\dot{V}^o(\underline{X}) = [\frac{\partial V^o}{\partial \underline{X}}]^T \{\underline{F}(\underline{X}) + \underline{B}\ \underline{U}(\underline{X})\} = -2\{(x_1+\tfrac{3}{2}x_2)^2 + \tfrac{3}{4}x_2^2 + x_3^2(3+5x_2^2+2x_3^4)\} , \tag{6-167}$$

which is negative definite in R^n. From Theorem 6-4, the origin of (6-166) is asymptotically stable in the large.

Example 6-2: Consider a system given by

$$\begin{bmatrix}\dot{x}_1\\ \dot{x}_2\end{bmatrix} = \begin{bmatrix}x_1+x_2-x_1^3-x_1x_2^2\\ -x_1+x_2-x_1^2x_2-x_2^3\end{bmatrix} + \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}u_1\\ u_2\end{bmatrix} \qquad (6\text{-}168)$$

and assume a linear feedback control law given by

$$\underline{U}(\underline{X}) = \begin{bmatrix}u_1(\underline{X})\\ u_2(\underline{X})\end{bmatrix} = \begin{bmatrix}-1 & 0\\ 0 & -1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix}, \qquad (6\text{-}169)$$

As n = m = r, it follows for the canonical form that $\underline{X}_1$ is nonexistent and $\underline{X} = \underline{X}_2$,

$$[\frac{\partial \underline{U}}{\partial \underline{X}}] = \begin{bmatrix}-1 & 0\\ 0 & -1\end{bmatrix}, \qquad (6\text{-}170)$$

$$[\frac{\partial \nabla^o}{\partial \underline{X}}] = \begin{bmatrix}\frac{\partial \nabla^o}{\partial x_1}\\ \frac{\partial \nabla^o}{\partial x_2}\end{bmatrix} = \underline{R}\begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix}\underline{X}, \qquad (6\text{-}171)$$

and

$$L(\underline{X}) = [x_1, x_2]\ \underline{R} \begin{bmatrix} -x_1-2x_2+2x_1^3+2x_1x_2^2 \\ 2x_1-x_2+2x_1^2x_2+2x_2^3 \end{bmatrix}, \qquad (6\text{-}172)$$

from (6-106). Arbitrarily assume

$$\underline{R} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \qquad (6\text{-}173)$$

Then the symmetry of $\underline{R}[\frac{\partial \underline{U}(\underline{X})}{\partial \underline{X}}]$ is satisfied. From Corollary 6-2, a performance index given by

$$\int_{t_o}^{\infty} \{L(\underline{X}) + u_1^2 + u_2^2\}\ dt \qquad (6\text{-}174)$$

can be optimized by the linear feedback control law (6-169) only if

$$L(\underline{X}) = -(x_1^2+x_2^2) + 2(x_1^2+x_2^2)^2 . \qquad (6\text{-}175)$$

Then the optimal performance index function becomes

$$V^o(\underline{X}) = x_1^2 + x_2^2 . \qquad (6\text{-}176)$$

To check the asymptotic stability of the synthesized feedback control system, calculate

$$\dot{V}^{o}(\underline{X}) = [\frac{\partial V^{o}}{\partial \underline{X}}]^{T} \{\underline{F}(\underline{X}) + \underline{B}\ \underline{U}\} = -(x_1^2 + x_2^2)^2 \qquad (6\text{-}177)$$

from (6-168), (6-169) and (6-171). Since $V^{o}(\underline{X})$ and $-\dot{V}^{o}(\underline{X})$ are positive definite, the origin of the synthesized system are asymptotically stable in the large.

Chapter 7

THE INVERSE PROBLEM OF LINEARLY SYNTHESIZED FEEDBACK CONTROL SYSTEMS

The inverse problem considered in this chapter is a subclass of the inverse problem of the previous chapter. This subclass is identified basically by the additional assumption that the synthesized feedback control system is linear. The precise statement of this problem, called the modified inverse problem, is given in Section 7.1. After a lemma is presented in Section 7.2, the results of Chapter 6 are restated for the specific subclass considered in this chapter. Finally the nonnegative property of the loss functions in an optimized performance index is discussed. The purpose for considering this modified inverse problem is to establish more general conclusions about optimal feedback control systems synthesized as linear systems, than have been presented in the literature.

7.1 Statement of the Modified Inverse Problem

The modified inverse problem considered in this chapter has additional assumptions to those stated for the inverse problem. In addition to assumptions (i)-(iv) in Section 6.1, it is assumed for the modified inverse problem that:

(i)' For the system equation given in the canonical form (6-7),

(a) $\hat{\underline{B}}$ is of full rank, i.e., m = r,

(b) $\underline{F}_1(\underline{X})$ is identically zero, and

(c) $\underline{F}_2(\underline{X})$ is an r-dimensional, vector valued, finite degree polynomial function given as

$$\underline{F}_2(\underline{X}) = \underline{F}_2^{(2)}(\underline{X}) + \underline{F}_2^{(3)}(\underline{X}) + \ldots\ldots + \underline{F}_2^{(\psi)}(\underline{X}) \ , \tag{7-1}$$

where each $\underline{F}_2^{(i)}(\underline{X})$ is i^{th} degree homogeneous and $\underline{F}_2^{(\psi)}(\underline{X})$ is not identically zero. (If a linear system is given, it is convenient to set $\psi = 1$ in (7-1) by identifying $\underline{A}_2\underline{X} = \underline{F}_2^{(1)}(\underline{X})$ in (6-35).) Thus the system (6-7) can be written as

$$\dot{\underline{X}} = \underline{A}\,\underline{X} + \underline{B}\,\{\underline{F}_2(\underline{X}) + \underline{U}\} \ ,$$

or, from (4-192) and (4-106),

$$\begin{bmatrix} \dot{\underline{X}}_{(1)} \\ \dot{\underline{X}}_{(2)} \\ \cdot \\ \cdot \\ \cdot \\ \dot{\underline{X}}_{(v)} \end{bmatrix} = \begin{bmatrix} \underline{A}_{(1,1)} & \underline{A}_{(1,2)} & & & & \\ & & \underline{A}_{(2,3)} & & & \\ & & & \cdot & & \\ & & & & \cdot & \\ & & & & & \underline{A}_{(v-1,v)} \\ \underline{A}_{(v,1)} & \cdot & \cdot & \cdot & \cdot & \underline{A}_{(v,v)} \end{bmatrix} \begin{bmatrix} \underline{X}_{(1)} \\ \underline{X}_{(2)} \\ \cdot \\ \cdot \\ \cdot \\ \underline{X}_{(v)} \end{bmatrix}$$

$$+ \begin{bmatrix} [0] \\ \\ \underline{F}_2(\underline{X}) \end{bmatrix} + \begin{bmatrix} [0] \\ \\ I_r \end{bmatrix} \underline{U} \text{ , all other entries zero,} \qquad (7\text{-}2)$$

which has the nonlinear functions in the last r equations corresponding to those equations which also have independent control variables,

(ii)' For $\underline{U}(\underline{X})$ of (6-9)

$$\underline{U}(\underline{X}) = -\underline{K}^T \underline{X} - \underline{F}_2(\underline{X}) \text{ ,} \qquad (7\text{-}3)$$

where $\underline{K}$ is an n x r matrix such that

$$\underline{K} = \begin{bmatrix} \underline{K}_{12} \\ \underline{K}_{22} \end{bmatrix} \begin{matrix} \} \, n - r \\ \} \, r \end{matrix} \qquad (7\text{-}4)$$

to provide negative real parts for all eigenvalues of $[\underline{A} - \underline{B}\ \underline{K}^T]$. The synthesized feedback control system, therefore, is the linear autonomous system

$$\dot{\underline{X}} = [\underline{A} - \underline{B}\ \underline{K}^T]\underline{X} \triangleq \underline{A}_K\ \underline{X} \qquad (7\text{-}5)$$

and the origin is asymptotically stable in the large. If the system (7-2) is uncontrollable, then all diagonal elements of $\underline{A}_{(1,1)}$ must have negative real parts for asymptotic stability, i.e., state variables in $\underline{X}_{(1)}$ behave as $\dot{\underline{X}} = \underline{A}_{(1,1)}\ \underline{X}_{(1)}$.

(iii)' $L(\underline{X})$, given by (6-27), is restricted to the form

$$L(\underline{X}) = L^{(2)}(\underline{X}) + L^{(3)}(\underline{X}) + \ldots + L^{(\sigma)}(\underline{X}) \ , \qquad (7\text{-}6)$$

where each $L^{(i)}(\underline{X})$ is i^{th} degree homogeneous and $\sigma \geq 2$ is a given integer.

7.2 Fundamental Lemma

Lemma 7-1:

The form of the optimal performance index function

$V^{o}(\underline{X})$ for the modified inverse problem is

$$V^{o}(\underline{X}) = V^{o(2)}(\underline{X}) + V^{o(3)}(\underline{X}) + \ldots + V^{o(\xi)}(\underline{X}) \ , \tag{7-7}$$

where

$$\xi \triangleq \max(\sigma, 2\psi) \ , \tag{7-8}$$

and σ and ψ are given by (7-1) and (7-6).

Proof: From (7-3) and (7-6), the loss function with a feedback control law is

$$\begin{aligned} L(\underline{X}) + \underline{U}^{T}(\underline{X})\,\underline{R}\,\underline{U}(\underline{X}) &= L^{(2)}(\underline{X}) + L^{(3)}(\underline{X}) + \ldots . \\ &\quad + L^{(\sigma)}(\underline{X}) + \{-K^{T}\underline{X} - \underline{F}_{2}(\underline{X})\}^{T}\,\underline{R}\{-\underline{K}^{T}\underline{X} - \underline{F}_{2}(\underline{X})\} \\ &\triangleq \mathcal{L}^{(2)}(\underline{X}) + \mathcal{L}^{(3)}(\underline{X}) + \ldots + \mathcal{L}^{(\xi)}(\underline{X}) \\ &\triangleq \mathcal{L}(\underline{X}) \ , \end{aligned} \tag{7-9}$$

where each $\mathcal{L}^{(i)}(\underline{X})$ is i^{th} degree homogeneous. Assume a solution for the synthesized system (7-5) for an arbitrary $(\underline{X}_{o}, t_{o})$,

$$\underline{\Phi}_{f}(t;\ \underline{X}_{o},\ t_{o}) = e^{\underline{A}_{K}(t-t_{o})}\,\underline{X}_{o} \ . \tag{7-10}$$

The value of the performance index from $(\underline{X}_o, t_o)$ is then

$$V^o(\underline{X}_o) = \int_{t_o}^{\infty} \mathcal{L}(e^{\underline{A}_K(t-t_o)}\underline{X}_o)dt \ . \qquad (7\text{-}11)$$

But the integral

$$\int_{t_o}^{\infty} \mathcal{L}^{(i)}(e^{\underline{A}_K(t-t_o)}\underline{X}_o)dt \ , \text{ for } i = 2,3,\ldots,\xi \ , \qquad (7\text{-}12)$$

is i^{th} degree homogeneous in $\underline{X}_o$ because the integrand is i^{th} degree in $\underline{X}_o$. Defining (7-12) to be $V^{o(i)}(\underline{X}_o)$, the lemma is proved.

7.3 Solution of the Modified Inverse Problem

According to the assumptions of the modified inverse problem, $\frac{\partial V^o}{\partial \underline{X}}$, (6-103), and the Hamilton-Jacobi equation, (6-104), are reduced to

$$\left[\frac{\partial V^o}{\partial \underline{X}}\right] = \begin{bmatrix} \frac{\partial V^o}{\partial \underline{X}_1} \\ \frac{\partial V^o}{\partial \underline{X}_2} \end{bmatrix} = \begin{bmatrix} 2\underline{K}_{12}R\,\underline{X}_2 + 2\int_o^{\underline{X}_2} \left[\frac{\partial \underline{F}_2(\underline{X})}{\partial \underline{X}_1}\right]^T \underline{R}d\underline{X}_2 + \underline{\tilde{w}}(\underline{X}_1) \\ 2\underline{R}\,\underline{K}^T\underline{X} + 2\underline{R}\,\underline{F}_2(\underline{X}) \end{bmatrix} , \qquad (7\text{-}13)$$

and

$$L(\underline{X}) = \underline{X}^T \underline{K}\ \underline{R}\ \underline{K}^T \underline{X} - 2\underline{X}^T \underline{K}\ \underline{R}\ \underline{A}_2 \underline{X} - 2\underline{X}_2^T \underline{R}\ \underline{K}_{12}^T \underline{A}_1 \underline{X}$$

$$- 2\underline{F}_2^T(\underline{X})\ \underline{R}\ \underline{A}_2 \underline{X} - \underline{F}_2^T(\underline{X})\ \underline{R}\ \underline{F}_2(\underline{X})$$

$$- 2\{\int_o^{\underline{X}_2} [\frac{\partial \underline{F}_2(\underline{X})}{\partial \underline{X}_1}]^T \underline{R} d\underline{X}_2\}^T \underline{A}_1 \underline{X} - \underline{\tilde{w}}^T(\underline{X}_1)\ \underline{A}_1\ \underline{X} \ . \qquad (7\text{-}14)$$

Then Theorem 6-1 can be restated as follows.

Corollary 7-1:

For the modified inverse problem, a performance index $\int_{t_o}^{\infty} (L(\underline{X}) + \underline{U}^T\ \underline{R}\ \underline{U})dt$ is optimized by a given $\underline{U}(\underline{X})$ if and only if

$$\text{(i)} \quad R\ K_{22}^T + R\ [\frac{\partial \underline{F}_2(\underline{X})}{\partial \underline{X}_2}] \qquad (7\text{-}15)$$

is symmetric,

(ii) there exists an (n - r) dimensional vector valued function $\underline{\tilde{w}}(\underline{X}_1)$ such that

$$\text{(a)} \quad \underline{\tilde{w}}(\underline{X}) = \underline{\tilde{w}}^{(1)}(\underline{X}_1) + \underline{\tilde{w}}^{(2)}(\underline{X}_1) + \ldots + \underline{\tilde{w}}^{(\xi-1)}(\underline{X}_1) \ , \qquad (7\text{-}16)$$

where each $\underset{\sim}{w}^{(i)}(\underline{X}_1)$ is i^{th} degree homogeneous and ξ is given by (7-19),

$$(b) \quad 2\frac{\partial}{\partial \underline{X}_1}\left\{\int_0^{\underline{X}_2}\left[\frac{\partial \underset{\sim}{F}_2(\underline{X})}{\partial \underline{X}_1}\right]\underline{R}d\underline{X}_2\right\} + \left[\frac{\partial \underset{\sim}{\tilde{w}}(\underline{X}_1)}{\partial \underline{X}_1}\right] \qquad (7\text{-}17)$$

is symmetric.

Proof: As m = r, (i) of Theorem 6-1 is nonexistent. The required symmetry of (7-15) and (7-17) follow from the requirements on (6-106) and (6-107). Since the highest and lowest degrees in $\nabla^o(\underline{X})$ are 2 and ξ respectively from (7-7), $[\frac{\partial \nabla^o}{\partial \underline{X}}]$ must be the sum of homogeneous polynomial functions with degree from 1 to $(\xi - 1)$. As $\underset{\sim}{F}_2(\underline{X})$ is also a polynomial, from (7-13) condition (7-16) must be satisfied.

Corollary 7-2:

Let $\underline{R}$ and $L_a(\underline{X})$ be calculated for an optimized performance index for a given $\underline{U}(\underline{X})$ and a $\underset{\sim}{\tilde{w}}_a(\underline{X}_1)$ as (7-16), and let $\nabla_a^o(\underline{X})$ be the resulting optimal performance index function. Then, a performance index with $\underline{R}$ and

$$L(\underline{X}) = L_a(\underline{X}) - \left[\frac{\partial \varphi(\underline{X}_1)}{\partial \underline{X}_1}\right]^T \underline{A}_1 \underline{X} \qquad (7\text{-}18)$$

can be optimized by the same $\underline{U}(\underline{X})$, where $\varphi(\underline{X}_1)$ is an arbitrary ξ^{th} degree polynomial function from degree 2 to $(\xi - 1)$. Then the resulting optimal performance index function $V^o(\underline{X})$ is

$$V^o(\underline{X}) = V_a^o(\underline{X}) + \varphi(\underline{X}_1) \ . \qquad (7\text{-}19)$$

The proof follows directly from Corollary 6-1 with the additional assumptions given.

Consider $\sigma = 2$ in (7-6) and the given system is linear, i.e., $\underline{F}_2(\underline{X})$ in (7-2) is identically zero, and $\psi = 1$. Then the control law (7-3) is a linear control law and the optimal performance index function $V^o(\underline{X})$ is a quadratic form, from Lemma 7-1. Necessarily $\underline{\tilde{w}}(\underline{X}_1)$ of (7-16) is a linear function of $\underline{X}_1$, say

$$\underline{\tilde{w}}(\underline{X}_1) \triangleq 2\underline{P}_{11}\underline{X}_1 \ , \qquad (7\text{-}20)$$

where $\underline{P}_{11}$ is an (n-r) x (n-r) symmetric matrix according to (iib) of Corollary 7-1. Therefore, (7-13) is

$$\left[\frac{\partial V^o}{\partial \underline{X}}\right] = 2\begin{bmatrix} \underline{P}_{11} & \underline{K}_{12}\underline{R} \\ \underline{R}\,\underline{K}_{12}^T & \underline{R}\,\underline{K}_{22}^T \end{bmatrix}\begin{bmatrix} \underline{X}_1 \\ \underline{X}_2 \end{bmatrix} , \qquad (7\text{-}21)$$

which results in

$$V^{o}(\underline{X}) = \underline{X}^{T}\begin{bmatrix} \underline{P}_{11} & \underline{K}_{12}\underline{R} \\ \underline{R}\ \underline{K}_{12}^{T} & \underline{R}\ \underline{K}_{22}^{T} \end{bmatrix}\underline{X} \triangleq \underline{X}^{T}\ \underline{P}\ \underline{X}\ . \qquad (7\text{-}22)$$

Substituting (7-20) into (7-14), $L(\underline{X})$ can be written as a quadratic form,

$$L(\underline{X}) \triangleq \underline{X}^{T}\ \underline{Q}\ \underline{X} \triangleq [\underline{X}_1^{T}\ \ \underline{X}_2^{T}]^{T}\begin{bmatrix} \underline{Q}_{11} & \underline{Q}_{12} \\ \underline{Q}_{12}^{T} & \underline{Q}_{22} \end{bmatrix}\begin{bmatrix} \underline{X}_1 \\ \underline{X}_2 \end{bmatrix}, \qquad (7\text{-}23)$$

where

$$\underline{Q}_{11} = \underline{K}_{12}\underline{R}\ \underline{K}_{12}^{T} - \underline{K}_{12}\underline{R}\ \underline{A}_{21} - \underline{A}_{21}^{T}\underline{R}\ \underline{K}_{12}^{T} - \underline{P}_{11}\underline{A}_{11} - \underline{A}_{11}^{T}\underline{P}_{11}$$

$$\underline{Q}_{12} = \underline{K}_{12}\underline{R}\ \underline{K}_{22}^{T} - \underline{A}_{21}^{T}\underline{R}\ \underline{K}_{22}^{T} - \underline{K}_{12}\underline{R}\ \underline{A}_{22} - \underline{A}_{11}^{T}\underline{K}_{12}\underline{R} - \underline{P}_{11}\underline{A}_{12}$$

$$\underline{Q}_{22} = \underline{K}_{22}\underline{R}\ \underline{K}_{22}^{T} - \underline{K}_{22}\underline{R}\ \underline{A}_{22} - \underline{A}_{22}^{T}\underline{R}\ \underline{K}_{22}^{T} - \underline{R}\ \underline{K}_{12}^{T}\underline{A}_{12} - \underline{A}_{12}^{T}\underline{K}_{12}\underline{R}\ , \qquad (7\text{-}24)$$

and $\underline{A}_{ij}$ is given in (6-32). Then Corollary 7-1 can be restated for this case.

Corollary 7-3:

For the modified inverse problem with the given system

linear and $L(\underline{X})$ restricted to a quadratic form, a performance index is optimized if and only if

(i) $L(\underline{X})$ is given as (7-23 and 24),

(ii) $\underline{R}\,\underline{K}_{22}^{T}$ is symmetric, and

(iii) $\underline{P}_{11}$ is symmetric.

Then $V^{o}(\underline{X})$ is reduced to (7-22).

The sign definiteness of $L(\underline{X})$ and $V^{o}(\underline{X})$ can be identified from $\underline{Q}$ and $\underline{P}$, according to Lemmas 6-2 and 3. Thus

Corollary 7-4: The $V^{o}(\underline{X})$ of Corollary 7-3, (7-22), is positive definite if and only if $\underline{R}\,\underline{K}_{22}^{T}$ and $\underline{P}_{11}-\underline{K}_{12}\underline{R}\,\underline{K}_{22}^{-1}\underline{K}_{12}^{T}$ are positive definite.

Corollary 7-5:

The $L(\underline{X})$ in (7-23) is positive semidefinite if and only if in (7-24)

(i) $\underline{Q}_{22}$ is positive semidefinite, say of rank r_1,

(ii) the last $r - r_1$ columns of $\underline{Q}_{12}\underline{C}_{22}$ are null, and

(iii) $\underline{Q}_{11} - \underline{Q}_{12}\underline{C}_{22}^{T}\underline{C}_{22}\underline{Q}_{22}\underline{C}_{22}^{T}\underline{C}_{22}\underline{Q}_{12}^{T}$ is positive semidefinite,

where $\underline{C}_{22}$ is an $(n-r) \times (n-r)$ nonsingular for the congruent transformation

$$\underline{C}_{22}^{T}\ \underline{Q}_{22}\ \underline{C}_{22} = \begin{bmatrix} I_{r_1} & [0] \\ [0] & [0] \end{bmatrix} . \qquad (7\text{-}25)$$

Corollary 7-6:

The $L(\underline{X})$ in (7-23) is positive definite if and only if in (7-23)

(i) $\underline{Q}_{22}$ is positive definite, and

(ii) $\underline{Q}_{11} - \underline{Q}_{12}\underline{Q}_{22}^{-1}\underline{Q}_{12}^{T}$ is positive definite.

Thus the inverse problem that Kalman originally presented has been generalized. While Kalman considered a controllable, single input, linear system with a linear control law, the results of this section are also applicable to uncontrollable, multi-input and not necessarily linear systems with a more general control law.

7.4 Nonnegative Loss Function of the Optimized Performance Index

Generally the loss function in (6-28) is assumed nonnegative. As $\underline{R}$ in this equation is restricted to be positive definite, the nonnegative property of the loss function depends on that of $L(\underline{X})$. The nonnegative property

of $L(\underline{X})$ is considered in this section.

7.4.1 Nonnegative $L(\underline{X})$ for a Controllable System

Theorem 7-1:

For the modified inverse problem, assume that the system (7-2) is completely controllable. Then $L(\underline{X})$ in an optimized performance index can be positive semi-definite only as a quadratic form of $\underline{X}$.

Proof: Assume the contrary, that an $L(\underline{X})$ in an optimized performance index is positive semidefinite polynomial

$$L(\underline{X}) = L^{(2)}(\underline{X}) + L^{(3)}(\underline{X}) + \ldots + L^{(\beta)}(\underline{X}) , \quad (7\text{-}26)$$

where $L^{(\beta)}(\underline{X})$ is not identically zero for an arbitrary β, $2 < \beta \leq \sigma$ and σ is given by (7-6). For (7-26) to be positive semidefinite, β must be even, from Theorem 2-3.

(1) Assume for the canonical form that $v = 1$. Then $\underline{A}_1$, $\underline{\tilde{w}}(\underline{X}_1)$ and $\underline{X}_1$ are nonexistent. It follows from (7-2) and (7-14) that

$$L^{(\beta)}(\underline{X}) = - \underline{F}_2^{(\psi)^T}(\underline{X}) \; \underline{R} \; \underline{F}_2^{(\psi)}(\underline{X})$$

$$= (\underline{D}\ \underline{F}_2^{(\psi)}(\underline{X}))^T\ [-I_r]\ (\underline{D}\ \underline{F}_2^{(\psi)}(\underline{X}))\ , \qquad (7\text{-}27)$$

where $\underline{D}$ is an r x r nonsingular matrix satisfying

$$\underline{R} = \underline{D}^T\ \underline{D}\ . \qquad (7\text{-}28)$$

The existence of $\underline{D}$ is from (ii) of Corollary 2-1, as $\underline{R}$ is positive definite. However (7-27) can have negative values. Thus (7-26) can also have negative values (Theorem 2-3), contradicting the hypothesis.

(2) Assume $v \geq 2$ for the canonical form. Let $\underset{\sim}{w}^{(k)}(\underline{X}_1)$ be the highest degree, nonidentically zero, homogeneous function in (7-16).

(2a) For the case of

$$2 \leq k + 1 < 2\psi\ , \qquad (7\text{-}29)$$

it follows from (7-14) that

$$L^{(\beta)}(\underline{X}) = -\underline{F}_2^{(\psi)T}(\underline{X})\ \underline{R}\ \underline{F}_2^{(\psi)}(\underline{X})\ . \qquad (7\text{-}30)$$

Arguing as in (1), (7-26) cannot be positive semidefinite and the hypothesis fails.

(2b) For the case of

$$2 \leq 2\psi < k + 1 , \tag{7-31}$$

it follows from (7-13) and (7-14) that

$$L^{(\beta)}(\underline{X}) = -\underline{\tilde{w}}^{(k)^T} \underline{A}_1 \underline{X}, \quad (\beta = k + 1) , \tag{7-32}$$

and

$$\nabla^{o(k+1)}(\underline{X}) = \int_o^{\underline{X}_1} \underline{\tilde{w}}^{(k)^T}(\underline{X}_1) d\underline{X}_1 . \tag{7-33}$$

According to (7-33) and Lemma 6-6, (7-32) can have negative values since it is not identically zero by the assumption and the system is controllable. Thus (7-26) can have negative values (Theorem 2-3) and the hypothesis fails.

(2c) For the case of

$$2 \leq 2\psi = k + 1 , \tag{7-34}$$

it follows that

$$L^{(\beta)}(\underline{X}) = -\underline{F}_2^{(\psi)^T}(\underline{X}) \underline{R} \underline{F}_2^{(\psi)}(\underline{X}) - \underline{\tilde{w}}^{(k)^T}(\underline{X}_1) \underline{A}_1 \underline{X} \tag{7-35}$$

and

$$V^{o(k+1)}(\underline{X}_1) = \int_o^{\underline{X}_1} \underline{\tilde{w}}^{(k)T}(\underline{X}_1)\, d\underline{X}_1 \ . \qquad (7\text{-}36)$$

From the results of (2a) and (2b), (7-35) can have negative values if it is not identically zero and the hypothesis fails.

To provide a quadratic form for $L(\underline{X})$ consistent with Theorem 7-1, $\underline{F}_2(\underline{X})$ must be identically zero, from (7-30). Thus for a linearly synthesized feedback control system with a completely controllable system, a nonnegative $L(\underline{X})$ is possible only if $L(\underline{X})$ is a quadratic form.

Optimal controls may be designed to minimize measures both of errors and energy during the control action. A performance index often used for these designs has the form

$$\int_{t_o}^{\infty} (\underline{X}^T\ \underline{Q}\ \underline{X} + \underline{U}^T\ \underline{R}\ \underline{U})\, dt \ , \qquad (7\text{-}37)$$

where $\underline{Q}$ and $\underline{R}$ are at least positive semidefinite. This choice of a performance index is due primarily to practical aspects of the problem, e.g., for mathematical convenience. From Theorem 7-1, however, this choice is seen to be

particularly appropriate if a completely controllable system is to be synthesized as a linear feedback control system.

For Kalman's inverse problem, Section 5.3.1, $L(\underline{X})$ is restricted to a quadratic form. From Theorem 7-1, it follows that no other nonnegative polynomial $L(\underline{X})$ can exist if the given system is completely controllable. In contrast, the example in Section 5.3.3, illustrating Suga's work, showed optimized nonquadratic polynomial $L(\underline{X})$. Theorem 7-1 explains why these nonquadratic polynomials are not positive semidefinite.

7.4.2 Nonnegative $L(\underline{X})$ for an Uncontrollable System

Assume that the given system (7-2) is uncontrollable. Then it follows that $v \geq 2$ and $\underline{X}_{(1)}$ represents the uncontrollable state variables governed by

$$\dot{\underline{X}}_{(1)} = \underline{A}_{(1,1)} \underline{X}_{(1)} \cdot \qquad (7\text{-}38)$$

Corollary 7-7:

Provided that the given system is uncontrollable in the modified inverse problem, then it is necessary for $L^{(\beta)}(\underline{X})$ in (7-26) to be a function of only $\underline{X}_{(1)}$ if $L(\underline{X})$ is positive semidefinite for some $\beta > 0$.

Proof: For $L(\underline{X})$ to be positive semidefinite with some $\beta > 2$, $L^{(\beta)}(\underline{X})$ must be nonnegative from Theorem 2-3. For $2\psi > k+1 \geq 2$, $L^{(\beta)}(\underline{X})$ can have negative values, as shown in (2a) of the proof of Theorem 7-1. For $k + 1 \geq 2\psi \geq 2$, it follows that

$$v^{o(k+1)}(\underline{X}) = \int_{o}^{\underline{X}_1} \underline{\tilde{w}}^{(k)T}(\underline{X}_1)\, d\underline{X}_1 \qquad (7\text{-}39)$$

and either

$$\text{(a)} \quad L^{(\beta)}(\underline{X}) = -\underline{\tilde{w}}^{(k)T}(\underline{X}_1)\underline{A}_1\underline{X} \text{ , if } k + 1 > 2\psi \text{ ,} \qquad (7\text{-}40)$$

or

$$\text{(b)} \quad L^{(\beta)}(\underline{X}) = -\underline{F}_2^{(\psi)T}(\underline{X})\,\underline{R}\,\underline{F}_2^{(\psi)}(\underline{X}) - \underline{\tilde{w}}^{(k)T}(\underline{X}_1)\underline{A}_1\underline{X} \text{ ,}$$

$$\text{if } k + 1 = 2\psi \text{ .} \qquad (7\text{-}41)$$

The last terms of either (7-40) or (7-41) may be positive semidefinite if it is a function of $\underline{X}_{(1)}$, from (7-39) and Lemma 6-6. Therefore for (7-41) to be positive semidefinite, $\underline{F}_2^{(\psi)}(\underline{X})$ must be a function only of $\underline{X}_{(1)}$.

For the case of the uncontrollable system, the consequences of the nonnegative property of $L(\underline{X})$ are not as definitive as those for the completely controllable system.

Chapter 8

CONCLUSIONS AND SUGGESTIONS FOR FURTHER STUDIES

8.1 Conclusions

Systems considered throughout in this study belong to the class given by

$$\underline{\dot{Y}} = \underline{\hat{F}}(\underline{Y}) + \underline{\hat{B}}\,\underline{V} ,$$

as described in Section 4.1.

8.1.1 The Canonical Form

A new canonical form is presented for this class of systems through a nonsingular transformation such that

$$\underline{Y} = \underline{N}\,\underline{X}$$

$$\underline{V} = \underline{M}\,\underline{U}$$

following the definition of an equivalent system, Definition 4-1. This canonical form for a linear system, i.e., $\underline{\hat{F}}(\underline{Y}) = \underline{\hat{A}}\,\underline{Y}$, is

$$\dot{\underline{X}} = \underline{A}\,\underline{X} + \underline{B}\,\underline{U}$$

$$= \begin{bmatrix} \underline{A}_{(1,1)} & \underline{A}_{(1,2)} & & & & \\ & & \underline{A}_{(2,3)} & & & \\ & & & \ddots & & \\ & & & & \underline{A}_{(v-1,v)} & \\ \underline{A}_{(v,1)} & \cdots & \cdots & \cdots & \cdots & \underline{A}_{(v,v)} \end{bmatrix} \begin{bmatrix} \underline{X}_{(1)} \\ \underline{X}_{(2)} \\ \vdots \\ \vdots \\ \vdots \\ \underline{X}_{(v)} \end{bmatrix}$$

$$+ \begin{bmatrix} & \\ & I_r \end{bmatrix} \underline{U} \text{ , all other entries zero,}$$

as characterized by the following statements.

(i) Each $\underline{A}_{(i,j)}$, $(i,j = 1,2,\ldots,v)$, is an $\ell_i \times \ell_j$ submatrix (Theorem 4-2).

(ii) v and ℓ_i, $(i = 1,2,\ldots,v)$, for the decomposition are positive integers uniquely determined by $\hat{\underline{A}}$ and $\hat{\underline{B}}$ of the given system such that

$$\sum_{i=1}^{v} \ell_i = n \text{ ,}$$

where each ℓ_i and the ordered set $\{\ell_1,\ell_2,\ldots,\ell_v\}$ are called the i^{th} stage number and the stage distribution (Theorem 4-2).

(iii) It is possible to let v = 1 if and only if n = m = r when the system is completely controllable (Theorem 4-2).

(iv) If $v \geq 2$ and the system is completely controllable, then

$$\underline{A}_{(1,1)} = [0]$$

and

$$\underline{A}_{(1,2)} = \begin{bmatrix} [0] & I_{\ell_1} \end{bmatrix} ,$$

where

$$\ell_i \leq \ell_j , \quad 1 \leq i < j \leq v .$$

In addition, if $v \geq 3$, then

$$\underline{A}_{(i,i+1)} = \begin{bmatrix} [0] & I_{\ell_i} \end{bmatrix} , \text{ for } i = 2,3,\ldots,v-1 \text{ (Theorem 4-2).}$$

(v) If the system is uncontrollable, then $v \geq 2$, $\underline{A}_{(1,1)}$ is given in Jordan canonical form, and $\underline{A}_{(1,2)} = [0]$. In addition, if $v \geq 3$,

$$\underline{A}_{(i,i+1)} = \left[[0] \quad I_{\ell_i}\right] \text{ , for } i = 2,3,\ldots,v-1, \text{ and}$$

$$\ell_i \leq \ell_j \text{ , } 2 \leq i < j \leq v \text{ (Theorem 4-2).}$$

In this case, only state variables $x_1, x_2, \ldots, x_{\ell_i}$ are uncontrollable, as discussed in Section 4.3.3.

(vi) The property of controllability is invariant for the transformation to the canonical form (Theorem 4-3).

(vii) The stage distribution of the system, i.e., the ordered set $\{\ell_1, \ell_2, \ldots, \ell_v\}$, is a unique characteristic (Theorem 4-4).

(viii) Provided m = r, the canonical transformation is possible with $\underline{M} = I_r$ if the given system has the stage distributions {r} or $\{\ell_1, r, \ldots, r\}$ (Theorem 4-6).

(ix) If the given system has the stage distribution {r,r,...,r,}, then only one $\underline{N}$ can exist with each possible $\underline{M}$ for the canonical transformation (Theorem 4-5).

From these characteristics, a number of observations can be made. The structure of $\underline{B}$ in the canonical form discloses the fundamental fact that only r independent control variables out of the m control variables contained in $\underline{V}$ can be effective in the control action. The number

v and the structures of $\underline{A}_{(1,1)}$ and $\underline{A}_{(1,2)}$ simply identify the controllability property of the given system from (iii)-(vi) above. If a given system is completely controllable and has $\hat{\underline{B}}$ of full rank and a stage distribution {r, r,...r}, then a unique canonical transformation is possible with $\underline{M} = I_r$, from (viii) and (ix). This particular canonical transformation coincides with the more familiar canonical transformation of Definition 4-1. Furthermore a single input completely controllable system has a unique phase variable canonical form and this unique form is the familiar form proposed by earlier investigators.

For a nonlinear system, the canonical form can be applied to the linear part and the above 8 characteristics are preserved for the linear part of the transformed system.

In comparison with the other phase canonical forms described in Chapter 3, the new canonical form has the following advantages.

(1) It can be applied to the entire class of systems given by (6-1) while other suggested canoncial forms are applicable essentially to subclasses.

(2) The many elements of the matrices $\underline{A}$ and $\underline{B}$ describing the linear part of the new canonical form can be reduced to units and zeros arranged in a simple and unique order. These decouple the matrices uniquely

due to the uniqueness of the stage distribution. Thus the mathematical structure of a general class of systems is compactly presented.

This canonical form has value for simplifying studies of optimal control problems for multi-input systems. Other known canonical forms are essentially subclasses of this new one.

8.1.2 The Inverse Problem of the Optimal Regulator

The inverse problem of the optimal regulator is considered for the class of systems given by (6-1). It is shown that the problem can be equivalently considered through the new canonical form under similar mathematical assumptions and without loss of generality. The recovery of the results for the originally given system is possible by the inverse of the canonical transformation. The analysis of the problem is efficiently performed with the compact structure of the canonical form.

Restricting the form of performance indices to

$$\int_{t_o}^{\infty} \{L(\underline{X}) + \underline{U}^T \underline{R}\, \underline{U}\}\, dt$$

with positive definite $\underline{R}$, the necessary and sufficient

conditions for optimized performance indices corresponding to a given $\underline{U}(\underline{X})$ are presented in a theorem (Theorem 6-1). From this theorem, new aspects of optimal feedback control systems are disclosed.

(i) At most, r functions out of m in the feedback control law are independent in optimal feedback control, i.e., there are r effective control functions for the optimal control action.

(ii) Various $L(\underline{X})$ can be paired with an $\underline{R}$ for optimized performance indices based on given $\underline{U}(\underline{X})$ (Corollary 6-1). However, if n = m = r, the $L(\underline{X})$ is unique for an $\underline{R}$ and given $\underline{U}(\underline{X})$ (Corollary 6-2).

(iii) When a linear feedback control law is given, the structure of the resulting optimal performance index function $V^{o}(\underline{X})$ is the sum of a quadratic form determined by the given feedback control law $\underline{U}(\underline{X})$ and an arbitrary function of $\underline{X}_1$. State variables in $\underline{X}_1$ are not exposed to $\underline{U}$ directly. The nonnegative property of these $V^{o}(\underline{X})$ is detailed (Theorems 6-2 and 6-3).

(iv) The controllability of nonlinear systems given by (6-68) is determined by the linear part of the nonlinear systems as (6-128). For this class of systems, the Principle of Necessity of Control Action is introduced.

Specially for the problem of the optimal regulator such that: the desired final condition of the system as $\underline{X} = \underline{0}$, the performance index is given as

$$\int_{t_o}^{\infty} \{L(\underline{X}) + \underline{U}^T \underline{R}\, \underline{U}\}\, dt$$

with a positive definite $\underline{R}$, $L(\underline{X})$ is nonnegative definite and a nonzero function of the controllable variables, then some control action is necessary for the optimality, i.e., a feedback control law not identically zero is necessary.

(v) When a feedback control law is given, the Liapunov direct method can be applied to identify the property of asymptotic stability in the large (Theorem 6-4).

With additional assumptions to the inverse problem, the modified inverse problem allows observation of general characteristics of linearly synthesized feedback control systems with a polynomial $L(\underline{X})$.

(vi) If the given system is controllable, a nonnegative $L(\underline{X})$ can exist only as a quadratic form (Theorem 7-1).

(vii) If the given system is uncontrollable, it is necessary for a nonnegative $L(\underline{X})$ that the lowest and highest degree homogeneous functions in $L(\underline{X})$ be functions

only of uncontrollable state variables (Corollary 7-7).

Thus this statement of the inverse problem and its solution is a generalization of the problem proposed by Kalman. It is also a generalization of Suga's work excepting his time varying assumption, and of Thau's results excepting his more general assumption of $\hat{\underline{R}}(\underline{V})$ in (5-67).

8.2 Suggestions for Further Studies

Throughout this work, problems are considered only under time invariant assumptions. The concepts and techniques appear to be extendable for time varying systems with some modification.

For the inverse problem, the matrix $\underline{R}$ is restricted to be positive definite to insure a normal Hamiltonian. Studies can be directed to attempt to relax this assumption, e.g., consider only positive semidefiniteness for R.

More generally, the inverse problem can be considered for basically different problem assumptions, e.g., more general systems descriptions and a different form of performance indices. Exhaustive studies of the inverse problem of the optimal regulator will reveal new characteristics of optimal feedback control systems.

REFERENCES

[1] Bellman, R., Dynamic Programming, Princeton University Press, Princeton, N. J., 1957.

[2] Pontryagin, L. S., et al., The Mathematical Theory of Optimal Processes, Interscience Pub. Inc., N. Y., 1962.

[3] Leitman, G., Optimization Technique, Academic Press, New York, N. Y., 1962.

[4] Drefus, S. E., Dynamic Programming and the Calculus of Variations, Academic Press, New York, N. Y., 1965.

[5] Athans, M. and P. L. Falb, Optimal Control, McGraw-Hill Book Co., Inc., New York, N. Y., 1966.

[6] Lee, E. B. and L. Markus, Foundations of Optimal Control Theory, John Wiley and Sons, Inc., New York, N. Y., 1967.

[7] Fleming, W. H., Functions of Several Variables, Addison-Wesley Pub. Co., Inc., Reading, Mass., 1965.

[8] Gibson, J. E., Nonlinear Automatic Control, McGraw-Hill Book Co., Inc., New York, N. Y., 1963.

[9] Hahn, W., Theory and Applications of Liapunov's Direct Method, (English Translation), Prentice-Hall, Inc., Englewood Cliffs, N. J., 1963.

[10] Thrill, R. M. and L. Tornheim, Vector Spaces and Matrices, John Wiley and Sons, Inc., New York, N. Y., 1957.

[11] Ayres, F., Jr., Theory and Problems of Matrices, Schaum Pub. Co., New York, N. Y., 1962.

[12] Kalman, R. E., "Mathematical Description of Linear Dynamical System," J. SIAM Control, Series A, Vol. 1, pp. 152-192, 1963.

[13] Pontryagin, L. S., Ordinary Differential Equations, (English Translation), Addison-Wesley Pub. Co., Inc., Reading, Mass., 1962.

[14] Kalman, R. E. and J. E. Bertram, "Control System Analysis and Design via the Second Method of Liapunov," Trans. ASME, Series D, J. Basic Eng., Vol. 82, pp. 371-393, 1960.

[15] Kalman, R. E., Y. C. Ho, and K. S. Narendra, Controllability of Linear Dynamical Systems, in Contribution to Differential Equations, Vol. 1, John Wiley and Sons, Inc., New York, N. Y., 1962.

[16] Wonham, W. M. and C. D. Johnson, "Optimal Bang-Bang Control with Quadratic Index of Performance," J. Basic Eng., Vol. 86, pp. 107-115, 1964.

[17] Johnson, C. D. and W. M. Wonham, "A Note on the Transformation to Canonical Form," Trans. IEEE, Vol. AC-9, pp. 312-313, 1964.

[18] Chidambara, M. R., "The Transformation to Canonical Form," Trans. IEEE, Vol. AC-10, pp. 206-207, 1965.

[19] Tuel, W. G., "On the Transformation to Canonical Form," Trans. IEEE, Vol. AC-11, pp. 607-608, 1966.

[20] Rane, D. S., A Simplified Transformation to Canonical Form," Trans. IEEE, Vol. AC-11, pp. 608, 1966.

[21] Tuel, W. G., "Canonical Forms for Linear Systems - I," IBM Research Paper RJ-375, 1966.

[22] Luenberger, D. G., "Canonical Forms for Linear Multi-Variable Systems," Trans. IEEE, Vol. AC-12, pp. 290-293, 1967.

[23] Asseo, S. J., "Phase-Variable Canonical Transformation of Multicontroller Systems," Trans. IEEE, Vol. AC-13, pp. 129-131, 1968.

[24] Dugundji, J., Topology, Allyn and Bacon, Inc., Boston, Mass., 1966.

[25] Kalman, R. E., "When is a System Optimal?", J. Basic Eng., Vol. 86, pp. 51-60, 1964.

[26] Kalman, R. E., "Contributions to the Theory of Optimal Control," Bol. Soc. Mat. Mex., Vol. 5, pp. 102-119, 1960.

[27] Snow, D. R., "Determining Reachable Regions and Optimal Controls," Advances in Control System, pp. 133-196, Academic Press, New York, N. Y., 1967.

[28] Bridgrand, T. F., Jr., "On the Existence of Optimal Feedback Controls," J. SIAM, Series A, Control Vol. 1, pp. 261-274, 1963.

[29] Boltyanskii, V. G., "Sufficient Conditions for Optimality and the Justification of the Dynamic Programming Method," J. SIAM, Series A, Control, Vol. 4, No. 2, pp. 326-361, 1966.

[30] Suga, I., "On Inverse Problem of Optimal Control," J. Soc. Inst. and Control Engn. (Japan), Vol. 6, No. 8, pp. 549-557, 1967.

[31] Thau, F. E., "On Inverse Optimal Control Problem for a Class of Non-linear Autonomous Systems," Trans. IEEE, Vol. AC-12, No. 6, pp. 674-681, 1967.

[32] Hsu, J. C. and A. U. Meyer, Modern Principles and Applications, pp. 696-697, McGraw-Hill Book Co., New York, N. Y., 1968.