SUBOPTIMAL DESIGN OF LINEAR REGULATOR SYSTEMS SUBJECT TO COMPUTER STORAGE LIMITATIONS

by

DAVID LEE KLEINMAN

B.E.E., The Cooper Union for the Advancement of Science and Art
1962

S.M.E.E., Massachusetts Institute of Technology
1963

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
January, 1967

Signature of Author _____
Department of Electrical Engineering, January 9, 1967

Certified by _____
Thesis Supervisor

Accepted by _____
Chairman, Departmental Committee on Graduate Students

SUBOPTIMAL DESIGN OF LINEAR REGULATOR SYSTEMS SUBJECT TO COMPUTER STORAGE LIMITATIONS

by

DAVID LEE KLEINMAN

Submitted to the Department of Electrical Engineering on January, 9 1967 in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

ABSTRACT

The feasibility of taking practical engineering constraints into consideration when designing optimal linear regulator systems is investigated. The study is conducted by prespecifying the structural form of time-varying feedback gains, while leaving various free parameters to be chosen optimally. In this manner, a suboptimal linear regulator problem is precisely formulated. Necessary conditions for its solution are obtained by introducing the concepts of a cost matrix and of a gradient matrix of a trace function. An algorithm is developed for computing the suboptimal control when the feedback gains are constrained to be piecewise constant. A numerical example illustrates the usefulness of the method.

Thesis Supervisor: Michael Athans

Title: Associate Professor of Electrical Engineering

ACKNOWLEDGEMENT

The author would like to express his sincere appreciation to Professors Michael Athans, Roger W. Brockett and William B. Davenport for offering constructive comments and criticisms while serving as thesis committee members, and to Professor George C. Newton who has been the author's Faculty Counselor during the doctoral program. A personal indebtedness is owed to Professor Michael Athans, not only for his acting as thesis committee chairman, but for his constant encouragement and moral support through difficult and trying years.

Thanks also go to the author's colleagues at the M.I.T. Electronic Systems Laboratory. The stimulating discussions held with Messrs. R. Canales, A. Debs, D. Gray, S. Greenberg, M. Gruber, A. Levis, W. Levine, J. Willems, Drs. H. Witsenhausen and P.L. Falb contributed greatly towards the report's exposition as well as towards a clearer understanding of suboptimal design problems.

Special acknowledgement is extended to Miss Judith Majeski (M.I.T. Lincoln Laboratory) for helping to program some of the results presented in Chapter V. The author wishes to also express his gratefulness to Miss Sandra Paul for preparing the technical illustrations, and to Miss Mary Sinclair and Mrs. Clara Conover for their typing of the final report.

The research presented in this document was made possible through the support extended by the National Aeronautics and Space Administration under Research Grants NsG-496 (M.I.T. Project DSR 79952) and NsG-22-009(124) (M.I.T. Project DSR 76265).

CONTENTS

CHAPTER I	INTRODUCTION page	1
CHAPTER II	THE OPTIMAL LINEAR REGULATOR PROBLEM	7
	A. Linear Dynamical Systems--Definition	7
	B. The Optimal Regulator Problem--Formulation	8
	C. The Optimal Regulator Problem--Solution	11
	D. A "Closed Form" Expression for $\underline{K}(t;T,\underline{F})$	16
	E. Controllability and Observability--Definitions	20
	F. The Optimal Linear Regulator Solution for $T = \infty$	24
CHAPTER III	RICCATI EQUATION COMPUTATIONS	32
	A. Implementation of the Optimal Feedback System	32
	B. Stability Properties of the Riccati Equation	35
	C. Off-Line, Numerical Solution of the Riccati Equation	40
	D. Computation of $\underline{K}(t;T,\underline{F})$ by Successive Approximations	49
CHAPTER IV	"SUBOPTIMAL" DESIGN TECHNIQUES	59
	A. Introduction	59
	B. Structure of the Suboptimal Control	64
	C. The Suboptimal Linear Regulator Problem	68
	D. Convergence of the Suboptimal Solution as $M \rightarrow \infty$	76
	E. Necessary Conditions for the Suboptimality of $\underline{L}^{o}(\cdot)$	82
CHAPTER V	SUBOPTIMAL PIECEWISE CONSTANT GAIN MATRICES	92
	A. Introduction	92
	B. Properties of the Suboptimal Solution as $N \rightarrow \infty$	92
	C. A Computational Scheme for Determining $\underline{L}^{o}(\cdot)$	98
	D. A Numerical Example	111

CONTENTS (Continued)

CHAPTER VI	TOPICS FOR FURTHER RESEARCH page	119
	A. Theoretical Studies	119
	B. Numero-Theoretic Investigations	122
CHAPTER VII	SUMMARY	129
CHAPTER VIII	CONCLUSIONS	132
APPENDIX A	THE EXISTENCE AND UNIQUENESS OF THE SOLUTION TO THE RICCATI EQUATION	133
APPENDIX B	PROOF OF THEOREM 6	138
APPENDIX C	COST MATRICES FOR LINEAR REGULATOR PROBLEMS	142
APPENDIX D	PROOF OF THEOREM 8 AND COROLLARY	147
APPENDIX E	ON THE RELATIONSHIP BETWEEN NEWTON'S METHOD AND THE METHOD OF SUCCESSIVE APPROXIMATIONS TO DETERMINE $\underline{K}(t;T,\underline{F})$	153
APPENDIX F	GRADIENT MATRICES OF TRACE FUNCTIONS	157
APPENDIX G	PROOF OF THEOREM 9	164
REFERENCES		170

to an impossible dream

CHAPTER I

INTRODUCTION

One of the most important and most widely-treated problems to date in the field of optimal control theory is the so-called "linear regulator problem."[1,2]† Historically, this problem of determining a control input to a linear system which minimizes the sum of integral squared error and control energy, finds its conception in Wiener's work on stationary time series and linear filtering and prediction problems.[3] The 1950's witnessed further contributions and extensions to the analysis of linear regulator systems[4,5,6] and today, among control theorists, we find a renewed interest in this area.

One of the primary reasons for this rebirth of interest in the linear regulator problem, besides the mathematical ease in which optimal control solutions are obtained in closed form, is that this study provides us with a strong correlation between the classical methods of analytic feedback system design via frequency domain methods and the more recent variational approach favoring analysis in the time domain.[7,8] The modern approach to the control problem, with a foundation resting on the concept of state variable descriptions

† Superscripts refer to numbered items in the References.

of dynamical systems and a structure molded by such tools as the minimum principle,[1,9] dynamic programming,[10] and the digital computer, is neither confined solely to time-invariant (stationary) problems nor is it confined to the consideration of only infinite-time control intervals.

The ability to consider the entire class of linear regulator problems in a general framework has not only unified the theory of optimal linear systems but has also served to uncover some of the underlying relationships that exist between the structure of the optimal system and such fundamental concepts as controllability[11] and observability.[12] Indeed, the linear regulator problem does not stand alone on a technical island, for the regard paid to its solution is matched in turn by the solution's importance and application to numerous segments of automatic control theory. The equations which appear in the study of the optimal linear regulator problem also appear in the study of every optimal tracking problem[1,2] and every linear filtering and prediction problem.[16,31] Therefore, results which are obtained from an investigation of the regulator problem are also pertinent to the tracking and filtering problems.

The elegant form in which the solution to the linear regulator problem may be expressed is well-known. The optimal control, $\underline{u}^*(t)$, for $t\epsilon[t_o, T]$ is simply a linear feedback control law

$$\underline{u}^*(t) = -\underline{L}^*(t)\,\underline{x}(t)$$

where $\underline{x}(t)$ is the current state of the system and $\underline{L}^*(t)$ is a matrix of feedback gains. The elements of $\underline{L}^*(t)$ are obtained from the solution of a nonlinear matrix differential equation (the Riccati equation) which lies at the heart of the optimization problem. However, this elegance gleams brighter in the eyes of the mathematician than in those of the engineer. Because of the computation instability of the Riccati equation solution in the forward time direction, it is not possible to accurately compute the elements of $\underline{L}^*(t)$, $t > t_o$, in an on-line manner by simply integrating the Riccati equation forward in time, starting from $t = t_o$.† It therefore becomes necessary to first solve the Riccati equation off-line, in the reverse time direction, by starting at $t = T$ with an appropriate boundary condition. Having accomplished this, the time-varying gains $\underline{L}^*(t)$ are then stored on tape in the feedback controller, to be played back upon command in real time. This method of implementing the optimal control is difficult and often impractical in many instances due to the circuitry requirements for synchronous playback of a large number of time-varying signals.

In this research we shall take these engineering problems into consideration, and propose a suboptimal control scheme for linear

---

† This is not the case in the filtering problem for which the Riccati equation solution is stable as $t \to +\infty$.[16] Nonetheless, the theoretical results of this report are still applicable to linear filtering problems.

regulator systems. Our goal is to determine a linear feedback control law which is relatively easy to implement, yet one which results in near optimal system performance. Our method of approach is to trade mathematical optimality in return for engineering simplicity and practical usefulness. This we shall accomplish by prespecifying the structural form of time-varying feedback gains, while leaving various free parameters to be chosen in an optimal fashion. In this manner, we shall precisely formulate, and subsequently analyze, a "suboptimal linear regulator problem."

Our initial task, which we undertake in Chapter II, is to explicitly define the optimal linear regulator problem. We discuss the solution to this optimal control problem in terms of the solution $\underline{K}(t; T, \underline{F})$ to the matrix Riccati differential equation. We show that the optimal control, $\underline{u}^*(t)$, may be expressed as a linear, time-varying feedback law.

$$\underline{u}^*(t) = -\underline{B}'(t)\, \underline{K}(t; T, \underline{F})\, \underline{x}(t) = -\underline{L}^*(t)\, \underline{x}(t)$$

and we present several well-known properties of $\underline{K}(t; T, \underline{F})$.

Having presented the reader with an understanding of the form of the optimal solution, we turn, in Chapter III, to methods for implementing the optimal control. We show that due to the computational instability of the Riccati equation solutions, one cannot accurately compute $\underline{K}(t; T, \underline{F})$ in an on line manner for $t > t_o$. This fact forces us to implement the optimal control by prestoring the elements of

$\underline{L}^*(t)$ on tape and playing the tape back upon command in real time to generate $\underline{u}^*(t)$. Therefore, $\underline{K}(t; T, \underline{F})$ is computed off-line, before the control system is placed into operation. This leads to our study of numerical techniques for the off-line computation of $\underline{K}(t; T, \underline{F})$, and we discuss three known algorithms in which the nonlinear Riccati differential equation is approximated by a nonlinear difference equation. We then develop an iterative scheme for determining $\underline{K}(t; T, \underline{F})$ which is an extension (to the matrix case) of Kalaba's method of successive approximations.[17] By introducing the concept of a "cost matrix," and solving a sequence of linear differential equations, we obtain a sequence of iterates which converge monotonically to $\underline{K}(t; T, \underline{F})$.

In Chapter IV, we discuss the engineering difficulties associated with storing the optimal feedback gain matrix $\underline{L}^*(t)$ on tape for $t \epsilon [t_o, T]$. Motivated by engineering feasibility, we then constrain the control input to our system to be of the form $\underline{u}(t) = -\underline{L}(t)\, \underline{x}(t)$, where we prescribe a time structure for the feedback gain matrix $\underline{L}(t)$. By leaving various free parameters in the description of $\underline{L}(t)$ it then becomes possible for us to choose a cost functional $\mu(\underline{L})$, and to develop the new concept of a "suboptimal linear regulator problem." Making use of gradient matrices, we then derive necessary conditions which the solution, $\underline{L}^o(t)$, of the suboptimal problem must satisfy, as well as various properties of the solution itself.

In Chapter V, we examine the important special case for which the feedback gain matrix $\underline{L}(t)$ is constrained to be piecewise constant over the control interval $[t_o, T]$. We discuss the implications of this constraint insofar as they relate to the storage limitations of a digital computer which may be used to implement the suboptimal control. We show that as the storage capacity is increased, the suboptimal control becomes arbitrarily close to the optimal control. We then apply the necessary conditions for suboptimality derived in Chapter IV, and develop an iterative scheme for computing the piecewise constant suboptimal gain matrix. A second-order example is included which illustrates the proposed method. Suggestions for further research comprise Chapter VI.

The major contribution of this report is the development and theoretical analysis of the suboptimal linear regulator problem, in particular, the piecewise constant problem of Chapter V. It is hoped that this research will disuade those critics of optimal control theory who argue that the gap between theory and practice has grown too wide. For it is possible to narrow that so-called "gap" by applying optimization techniques with one hand, while taking into account practical engineering constraints with the other. This research is but one such attempt.

CHAPTER II

THE OPTIMAL LINEAR REGULATOR PROBLEM

An essential prerequisite for any "sub-optimal" design of a linear regulator system is a thorough understanding of the optimal linear regulator problem itself. This knowledge provides a strong base upon which to build a theory of sub-optimization, and in the final analysis it is this knowledge which must be used to judge the merits of our sub-optimal design.

In this chapter we shall formulate the optimal linear regulator problem in a mathematical framework, and discuss its solution in terms of the solution to a matrix Riccati differential equation. We shall not attempt to be all inclusive, but merely present the salient features of the optimal solution and discuss several properties of the Riccati equation which appear elsewhere in the literature.

This chapter is an abridged version of Chapters II, IV and VI of Reference 13. In some cases, the proofs of certain results are omitted for the sake of brevity. For a more extensive investigation into the linear regulator problem and its associated Riccati equation the reader is urged to see References 1, 2, 13, 16.

A. LINEAR DYNAMICAL SYSTEMS--DEFINITION

In the sequel we confine our attention to dynamical systems which are characterized by the following elements:

(1) A time set $\{t\}$ which we shall take to be the real line, i.e., $\{t\} = (-\infty, \infty) = E_1$

(2) A set of states $\{\underline{x}\} = X = E_n$ called the state space, where E_n is an n-dimensional Euclidean vector space.

(3) A set of inputs or controls $\{\underline{u}\} = U = E_r$ called the input space.

(4) A function space Ω whose elements are bounded, measurable functions which map E_1 into U.

(5) A set of outputs $\{\underline{y}\} = Y = E_m$ called the output space.

(6) A linear differential equation which describes the evolution of the state of the system in time, i.e.,

$$\frac{d}{dt}\ \underline{x}(t) = \underline{A}(t)\,\underline{x}(t) + \underline{B}(t)\,\underline{u}(t) \qquad (2.1)$$

where the nxn and nxr matrices $\underline{A}(t)$ and $\underline{B}(t)$, respectively, are locally integrable.

(7) An algebraic equation which relates the output vector at time t to the state at time t, viz.,

$$\underline{y}(t) = \underline{C}(t)\,\underline{x}(t) \qquad (2.2)$$

where $\underline{C}(t)$ is an mxn matrix which is locally integrable.

A system, Σ, possessing the above properties is called a "continuous time, linear dynamical system."

B. THE OPTIMAL REGULATOR PROBLEM--FORMULATION

Let us suppose that we are given a linear dynamical system Σ satisfying conditions (1) - (7). Let us suppose further that t_o and $\underline{x}_o$ are <u>given</u> elements of $(-\infty, \infty)$ and X, respectively, and that T is a <u>given</u> element of (t_o, ∞), i.e., $T > t_o$.†

† $\underline{x}_o$, t_o and T are often referred to as the initial state, the initial time and the final or terminal time, respectively.

If $\underline{u}(\cdot)$ is a given element of Ω, let $\underline{x}_u(t) = \underline{\phi}(t; t_o, \underline{x}_o, \underline{u}(\cdot))$ denote the solution of the system equation (2.1) starting from $\underline{x}_o$ at time t_o (i.e., $\underline{x}(t_o) = \underline{x}_o$), and generated by the control $\underline{u}(\cdot)$. Let $\underline{y}_u(t) = \underline{C}(t)\underline{x}_u(t)$ be the corresponding output trajectory. The optimal linear regulator problem is then to determine the control $\underline{u}(\cdot) \in \Omega$ which minimizes the quadratic cost functional

$$J(\underline{x}_o, t_o, T, \underline{u}(\cdot)) = \frac{1}{2}<\underline{x}_u(T), \underline{F}\underline{x}_u(T)> + \frac{1}{2}\int_{t_o}^{T} [<\underline{y}_u(t), \underline{y}_u(t)> + <\underline{u}(t), \underline{u}(t)>]\,dt \tag{2.3}$$

where $\underline{F}$ is a positive semi-definite matrix, the "terminal state" $\underline{x}_u(T) \in X$ is unconstrained, and the terminal time T is fixed.

We shall denote by $\underline{u}^*(\cdot)$ the control which minimizes (2.3) and by $\underline{x}^*(\cdot)$ the trajectory in the state space X, generated by $\underline{u}^*(\cdot)$.

Note that there is no loss of generality in considering the cost functional (2.3). The most general form of the cost functional (2.3) for a given system Σ is[1]

$$J(\underline{x}_o, t_o, T, \underline{u}(\cdot)) = \frac{1}{2}<\underline{y}_u(T), \underline{F}\underline{y}_u(T)> + \frac{1}{2}\int_{t_o}^{T} [<\underline{y}_u(t), \underline{Q}(t)\underline{y}_u(t)> + <\underline{u}(t), \underline{R}(t)\underline{u}(t)>]\,dt \tag{2.4}$$

where $\underline{Q}(t)$ is an $m \times m$ positive semi-definite matrix and $\underline{R}(t)$ is an $r \times r$ positive definite matrix and $\underline{F}$ and $\underline{Q}(t)$ are not both identically zero. However, if we now define a new system $\tilde{\Sigma}$ which is charac-

terized by the equations

$$\tilde{\Sigma}: \quad \frac{d}{dt}\tilde{\underline{x}}(t) = \tilde{\underline{A}}(t)\tilde{\underline{x}}(t) + \tilde{\underline{B}}(t)\tilde{\underline{u}}(t) \; ; \quad \tilde{\underline{x}}_o = \underline{x}_o$$

$$\tilde{\underline{y}}(t) = \tilde{\underline{C}}(t)\,\tilde{\underline{x}}(t)$$

where $\tilde{\underline{A}}(t) = \underline{A}(t)$, $\tilde{\underline{B}}(t) = \underline{B}(t)\underline{R}^{-1/2}(t)$, $\tilde{\underline{C}}(t) = \underline{Q}^{1/2}(t)\underline{C}(t)$, $\tilde{\underline{u}}(t) = \underline{R}^{1/2}(t)\underline{u}(t)$† (so that $\tilde{\underline{x}}(t) = \underline{x}(t)$), it is clear that the cost functional (2.4) written with respect to the system $\tilde{\Sigma}$ becomes

$$\tilde{J}(\underline{x}_o, t_o, T, \tilde{\underline{u}}(\cdot)) = \frac{1}{2}\langle \tilde{\underline{x}}_{\tilde{u}}(T), \underline{C}'(T)\underline{F}\underline{C}(T)\tilde{\underline{x}}_{\tilde{u}}(T)\rangle$$

$$+ \frac{1}{2}\int_{t_o}^{T} [\langle \tilde{\underline{y}}_{\tilde{u}}(t), \tilde{\underline{y}}_{\tilde{u}}(t)\rangle + \langle \tilde{\underline{u}}(t), \tilde{\underline{u}}(t)\rangle]\, dt$$

which is of the same form as Eq. (2.3). Hence, without loss of generality, we shall consider the following optimization problem, which we summarize for convenience.

The Optimal Linear Regulator Problem

Given the linear dynamical system Σ, characterized by the equations

$$\dot{\underline{x}}(t) = \underline{A}(t)\underline{x}(t) + \underline{B}(t)\,\underline{u}(t) \; ; \quad \underline{x}(t_o) = \underline{x}_o$$

$$\underline{y}(t) = \underline{C}(t)\,\underline{x}(t)$$

and the cost functional.

---

† Since $\underline{R}^{-1}(t)$ is positive definite it possesses a unique positive definite square root, written as $\underline{R}^{-1/2}(t)$. Similarly, the positive semi-definiteness of $\underline{Q}(t)$ implies the existence of the unique positive semi-definite square root, $\underline{Q}^{1/2}(t)$.

$$J(\underline{x}_o, t_o, T, \underline{u}(\cdot)) = \frac{1}{2} <\underline{x}(T), \underline{F}\,\underline{x}(T)> + \frac{1}{2} \int_{t_o}^{T} [<\underline{y}(t), \underline{y}(t)> + <\underline{u}(t), \underline{u}(t)>]\, dt$$

where T is fixed. Find the control $\underline{u}^*(\cdot)$ over the interval $[t_o, T]$ such that $J(\underline{x}_o, t_o, T, \underline{u}(\cdot))$ is minimized.

C. THE OPTIMAL REGULATOR PROBLEM--SOLUTION[1]

The optimization problem posed above is solved most expeditiously by use of the Minimum Principle. Using this method we find that the optimal control $\underline{u}^*$ as a function of time may be obtained by solving the $2n \times 2n$ Hamiltonian system of equations

$$\begin{bmatrix} \dot{\underline{x}}(t) \\ \cdots \\ \dot{\underline{p}}(t) \end{bmatrix} = \begin{bmatrix} \underline{A}(t) & \vdots & -\underline{B}(t)\underline{B}'(t) \\ \cdots\cdots & \vdots & \cdots\cdots\cdots \\ -\underline{C}'(t)\underline{C}(t) & \vdots & -\underline{A}'(t) \end{bmatrix} \begin{bmatrix} \underline{x}(t) \\ \cdots \\ \underline{p}(t) \end{bmatrix} \tag{2.5}$$

subject to the boundary conditions

$$\begin{aligned} \underline{x}(t_o) &= \underline{x}_o \\ \underline{p}(T) &= \underline{F}\,\underline{x}(T) \end{aligned} \tag{2.6}$$

The optimal control for $t \in [t_o, T]$ is then given by

$$\underline{u}^*(t) = -\underline{B}'(t)\,\underline{p}(t) \tag{2.7}$$

The system of Eqs. (2.5) represents simply the Euler equations for our minimization problem. However, the solution of these equations is difficult due to the computational difficulties involved in solving time-varying equations with split boundary conditions.

Alternatively, a more manageable solution to the regulator problem is obtained by solving for $\underline{u}^*(t)$ as an instantaneous function of the time t and the state $\underline{x}(t)$, (i.e., solving for the optimal feedback control

law). In this case we find that the n-vector $\underline{p}(t)$ and the n-vector $\underline{x}(t)$ can be related by the linear transformation $\underline{K}(t)$, i.e.,

$$\underline{p}(t) = \underline{K}(t)\,\underline{x}(t) \tag{2.8}$$

so that the optimal control law may be written as a linear feedback law

$$\underline{u}^*(\underline{x}(t), t) = -\underline{B}'(t)\underline{K}(t)\,\underline{x}(t) = -\underline{L}^*(t)\underline{x}(t) \tag{2.9}$$

provided that $\underline{K}(t)$ is the unique solution of the matrix differential equation

$$\dot{\underline{K}}(t) = -\underline{A}'(t)\underline{K}(t) - \underline{K}(t)\underline{A}(t) - \underline{C}'(t)\underline{C}(t) + \underline{K}(t)\underline{B}(t)\underline{B}'(t)\underline{K}(t) \tag{2.10}$$

satisfying the boundary condition

$$\underline{K}(T) = \underline{F} \tag{2.11}$$

Under these conditions, the state $\underline{x}^*(t)$ of the optimal system is generated by the linear differential equation

$$\frac{d}{dt}\underline{x}^*(t) = \left[\underline{A}(t) - \underline{B}(t)\underline{B}'(t)\underline{K}(t)\right]\underline{x}^*(t)\ ;\ \underline{x}^*(t_o) = \underline{x}_o \tag{2.12}$$

Consequently, we see that any study of the optimal linear regulator problem is intrinsically tied to the study of the properties of the solution to Eq. 2.10. We call Eq. 2.10 the "matrix Riccati equation" or, for short, the "Riccati equation".

The first and most immediate property of the Riccati equation solution is

> Proposition 1: If $\underline{K}(t)$ is the unique solution of the Riccati equation (2.10) satisfying $\underline{K}(T) = \underline{F}$, then $\underline{K}(t)$ is symmetric, i.e., $\underline{K}(t) = \underline{K}'(t)$ for all $t \leq T$.

This proposition is well-known and is given here for the sake of completeness. Its proof, which is elementary if one takes the transpose of both sides of Eq. (2.10), is omitted.

Let us now define the functional, $J^*(\underline{x}, t, T)$, for arbitrary $t \in (-\infty, T)$ and $\underline{x} \in X$ as being the optimal "cost" relative to $(\underline{x}, t)$, i.e.,

$$J^*(\underline{x}, t, T) = \min_{\underline{u}(\cdot) \in \Omega} J(\underline{x}(t), t, T, \underline{u}(\cdot))$$

$$= \frac{1}{2} <\underline{x}^*(T), \underline{F}\underline{x}^*(T)> + \frac{1}{2} \int_t^T [<\underline{y}^*(\tau), \underline{y}^*(\tau)> + <\underline{u}^*(\tau), \underline{u}^*(\tau)>] d\tau$$

$$= J(\underline{x}, t, T, \underline{u}(\cdot))\Big|_{\underline{u} = \underline{u}^*} \qquad (2.13)$$

Having defined $J^*(\underline{x}, t, T)$ we now state and prove another well-known result. We present a more direct proof than those commonly found in the control literature[1,2] which rely upon Hamilton-Jacobi theory. We show,

Lemma 1: If $\underline{K}(t)$ is the unique solution of the Riccati equation satisfying $\underline{K}(T) = \underline{F}$, then

$$J^*(\underline{x}, t, T) = \frac{1}{2} <\underline{x}, \underline{K}(t)\underline{x}> \qquad (2.14)$$

Proof: In the proof we shall assume (without loss of generality) that $\underline{F} = \underline{0}$. Then, substituting

$$\underline{u}^*(\tau) = -\underline{B}'(\tau)\underline{K}(\tau)\underline{x}^*(\tau)$$

and

$$\underline{y}^*(\tau) = \underline{C}(\tau)\underline{x}^*(\tau)$$

into Eq. (2.13) we find that

$$J^*(\underline{x}, t, T) = \frac{1}{2} \int_t^T < \underline{x}^*(\tau), \left[\underline{C}'(\tau)\underline{C}(\tau)+\underline{K}(\tau)\underline{B}(\tau)\underline{B}'(\tau)\underline{K}(\tau)\right]\underline{x}^*(\tau) > d\tau$$

If we now define $\underline{\Phi}_c(\tau, t)$ as being the transition matrix of the optimal closed loop system 2.12, i.e. $\underline{\Phi}_c(\tau, t)$ satisfies the equation

$$\frac{d}{d\tau}\underline{\Phi}_c(\tau, t) = \left[\underline{A}(\tau) - \underline{B}(\tau)\underline{B}'(\tau)\underline{K}(\tau)\right]\underline{\Phi}_c(\tau, t) \; ; \; \underline{\Phi}_c(t, t) = \underline{I}$$

then

$$\underline{x}^*(\tau) = \underline{\Phi}_c(\tau, t)\,\underline{x}^*(t) = \underline{\Phi}_c(\tau, t)\,\underline{x} \qquad (2.15)$$

Substituting Eq. (2.15) into the above expression for $J^*(\underline{x}, t, T)$ we obtain

$$J^*(\underline{x}, t, T) = \frac{1}{2} < \underline{x}, \underline{V}(t)\,\underline{x} >$$

where

$$\underline{V}(t) = \int_t^T \underline{\Phi}_c'(\tau, t)\left[\underline{C}'(\tau)\underline{C}(\tau)+\underline{K}(\tau)\underline{B}(\tau)\underline{B}'(\tau)\underline{K}(\tau)\right]\underline{\Phi}_c(\tau, t)\, d\tau \qquad (2.16)$$

Note that $\underline{V}(t)$ is uniformly continuous in t. It remains only to show that $\underline{V}(t) = \underline{K}(t)$. To accomplish this we differentiate both sides of Eq. (2.16) with respect to t. Using the well-known relation

$$\frac{d}{dt}\underline{\Phi}_c'(\tau, t) = -\left[\underline{A}(t) - \underline{B}(t)\underline{B}'(t)\underline{K}(t)\right]' \underline{\Phi}_c'(\tau, t)$$

and its transpose

$$\frac{d}{dt}\underline{\Phi}_c(\tau, t) = -\underline{\Phi}_c(\tau, t)\left[\underline{A}(t) - \underline{B}(t)\underline{B}'(t)\underline{K}(t)\right]$$

we obtain, since $\underline{K}(t) = \underline{K}'(t)$, that

$$\frac{d}{dt}\underline{V}(t) = -\underline{A}'(t)\underline{V}(t) - \underline{V}(t)\underline{A}(t) + \underline{K}(t)\underline{B}(t)\underline{B}'(t)\underline{V}(t)$$

$$+ \underline{V}(t)\underline{B}(t)\underline{B}'(t)\underline{K}(t) - \underline{C}'(t)\underline{C}(t) - \underline{K}(t)\underline{B}(t)\underline{B}'(t)\underline{K}(t) \quad (2.17)$$

with $\underline{V}(T) = \underline{0}$. But Eq. (2.17) is a linear equation in $\underline{V}(t)$ and therefore possesses a unique solution. Consequently if $\underline{K}(t)$ is the unique solution of the Riccati equation with $\underline{K}(T) = \underline{0}$, substituting $\underline{V}(t) = \underline{K}(t)$ in (2.17) will result in an identity. ‖

In particular, since $J^*(\underline{x}, t, T)$ is non-negative (by virtue of the fact that $\underline{F}$ is positive semi-definite) we deduce from Lemma 1 that, for any element $\underline{x} \in X$,

$$<\underline{x}, \underline{K}(t)\underline{x}> = 2J^*(\underline{x}, t, T) \geq 0 \quad (2.18)$$

Therefore, $\underline{K}(t)$ is positive semi-definite for all $t \leq T$ provided that it is well defined.

The above discussion has been contingent upon the fact that the solution to the Riccati equation (2.10) is unique. However, this equation, being nonlinear, may not have any solution, much less a unique solution. Consequently, an investigation dealing with the existence and uniqueness of the Riccati equation solution becomes necessary. In Appendix A we show, by taking into account the nature of our specific optimization problem, that Eq. (2.10) does, indeed, possess a well-defined solution over the entire interval $(-\infty, T]$. Our main result is

Theorem 1: For all T and all positive semi-definite matrices $\underline{F}$, the equation

$$\dot{\underline{K}}(t) = -\underline{A}'(t)\underline{K}(t) - \underline{K}(t)\underline{A}(t) - \underline{C}'(t)\underline{C}(t) + \underline{K}(t)\underline{B}(t)\underline{B}'(t)\underline{K}(t)$$

has a unique, positive semi-definite solution defined over the entire interval $(-\infty, T]$ which satisfies $\underline{K}(T) = \underline{F}$.

Finally, for notational purposes in the sequel, we define:

Definition 1: Let $\underline{K}(t; T, \underline{F})$, $t \leq T$ denote the unique solution of the Riccati equation (2.10) satisfying the boundary condition $\underline{K}(T; T, \underline{F}) = \underline{F}$.

D. A "CLOSED FORM" EXPRESSION FOR $\underline{K}(t; T, \underline{F})$†

We can obtain the solution to the Riccati equation in a closed-form by considering the Euler equations (2.5) corresponding to the underlying minimization problem. These equations are repeated for convenience as

$$\begin{bmatrix} \dot{\underline{x}} \\ \cdots \\ \dot{\underline{p}} \end{bmatrix} = \underline{Z} \begin{bmatrix} \underline{x} \\ \cdots \\ \underline{p} \end{bmatrix} \tag{2.19}$$

where

$$\underline{Z} = \begin{bmatrix} \underline{A}(t) & \vdots & -\underline{B}(t)\underline{B}'(t) \\ \cdots\cdots & \vdots & \cdots\cdots \\ -\underline{C}'(t)\underline{C}(t) & \vdots & -\underline{A}'(t) \end{bmatrix} \tag{2.20}$$

Prior to deriving an expression for $\underline{K}(t; T, \underline{F})$, we shall investigate some properties associated with the $2n \times 2n$ matrix $\underline{Z}$. First of all, we note that $\underline{Z}$ satisfies the relation

---

† The results of this section have been derived by Kalman and may be found in Reference 16. We have repeated the proofs for the sake of completeness and to gain further insight to the structure of the matrix $\underline{K}(t; T, \underline{F})$.

$$\underline{Z} = \underline{J}\,\underline{Z}'\underline{J} \tag{2.21}$$

where

$$\underline{J} = \begin{bmatrix} \underline{0} & \vdots & -\underline{I} \\ \cdots & \vdots & \cdots \\ \underline{I} & \vdots & \underline{0} \end{bmatrix} \tag{2.22}$$

As a consequence of Eq. (2.21) we can immediately show

Proposition 2: If λ is an eigenvalue of $\underline{Z}$, then so is $-\lambda$.

Proof: $\underline{Z}\,\underline{\xi} = \lambda\,\underline{\xi}$ implies that $\underline{J}\underline{Z}'\underline{J}\underline{\xi} = \lambda\,\underline{\xi}$. But since $\underline{J}^{-1} = -\underline{J} = \underline{J}'$ we have $\underline{Z}'\underline{J}\underline{\xi} = -\lambda\,\underline{J}\,\underline{\xi}$. This implies that $-\lambda$ is an eigenvalue of $\underline{Z}'$ (with eigenvector $\underline{J}\,\underline{\xi}$). But a matrix and its transpose have the same eigenvalues and so $-\lambda$ is an eigenvalue of $\underline{Z}$. ‖

If we now let

$$\underline{\Psi}(t, t_o) = \begin{bmatrix} \underline{\Psi}_{11}(t, t_o) & \vdots & \underline{\Psi}_{12}(t, t_o) \\ \cdots & \vdots & \cdots \\ \underline{\Psi}_{21}(t, t_o) & \vdots & \underline{\Psi}_{22}(t, t_o) \end{bmatrix} \tag{2.23}$$

be the transition matrix of Eq. (2.19), it follows, due to the form of $\underline{Z}$ that

Lemma 2: $\underline{\Psi}(t, t_o)$ satisfies

$$\left.\begin{aligned} \underline{\Psi}_{11}(t, t_o) &= \underline{\Psi}'_{22}(t_o, t) \\ \underline{\Psi}_{12}(t, t_o) &= \underline{\Psi}'_{12}(t_o, t) \\ \underline{\Psi}_{21}(t, t_o) &= \underline{\Psi}'_{21}(t_o, t) \end{aligned}\right\} \tag{2.24}$$

Proof: Let $\underline{y} = \text{col}(\underline{x}, \underline{p})$, and let $\underline{\Psi}(t, t_o)$ be the transition matrix of $\dot{\underline{y}} = \underline{Z}\,\underline{y}$. Let $\underline{J}\underline{v} = \underline{y}$, so that the vector $\underline{v}$ satisfies the equation

$$\dot{\underline{v}} = \underline{J}^{-1}\underline{Z}\underline{J}\underline{v} = -\underline{J}\underline{Z}\underline{J}\underline{v} = -\underline{Z}'\underline{v}$$

Since $\underline{\Psi}'(t_o, t)$ is the transition matrix of $\dot{\underline{v}} = -\underline{Z}'\underline{v}$ and since $\underline{v} = \underline{J}^{-1}\underline{y} = \underline{J}'\underline{y}$, the transition matrix of $\dot{\underline{y}} = \underline{Z}\,\underline{y}$ is given by

$$\underline{\Psi}(t, t_o) = \underline{J}^{-1}\underline{\Psi}'(t_o, t)\underline{J} = \underline{J}'\underline{\Psi}'(t_o, t)\underline{J}$$

which yields the desired results. ‖

We can now show,

Theorem 2:

$$\underline{K}(t; T, \underline{F}) = \left[\underline{\Psi}_{22}(T, t) - \underline{F}\underline{\Psi}_{12}(T, t)\right]^{-1} \cdot \left[\underline{F}\underline{\Psi}_{11}(T, t) - \underline{\Psi}_{21}(T, t)\right] \qquad \text{(i)}$$

$$= \left[\underline{\Psi}_{21}(t, T) + \underline{\Psi}_{22}(t, T)\underline{F}\right] \cdot \left[\underline{\Psi}_{11}(t, T) + \underline{\Psi}_{12}(t, T)\underline{F}\right]^{-1} \qquad \text{(ii)}$$

Proof: Since

$$\begin{bmatrix} \underline{x}(T) \\ \cdots \\ \underline{p}(T) \end{bmatrix} = \underline{\Psi}(T, t) \begin{bmatrix} \underline{x}(t) \\ \cdots \\ \underline{p}(t) \end{bmatrix}$$

and since $\underline{p}(T) = \underline{F}\underline{x}(T)$ we have

$$\underline{x}(T) = \underline{\Psi}_{11}(T, t)\,\underline{x}(t) + \underline{\Psi}_{12}(T, t)\,\underline{p}(t)$$

$$\underline{p}(T) = \underline{F}\underline{x}(T) = \underline{\Psi}_{21}(T, t)\underline{x}(t) + \underline{\Psi}_{22}(T, t)\underline{p}(t)$$

But since $\underline{p}(t) = \underline{K}(t; T, \underline{F})\, \underline{x}(t)$, we find that

$$\underline{K}(t; T, \underline{F}) = \left[\underline{\Psi}_{22}(T, t) - \underline{F}\underline{\Psi}_{12}(T, t)\right]^{-1} \cdot \left[\underline{F}\underline{\Psi}_{11}(T, t) - \underline{\Psi}_{21}(T, t)\right]$$

The inverse term in the above expression exists provided $\underline{K}(t; T, \underline{F})$ exists. In particular, this inverse will exist for all $t \leq T$ if $\underline{F}$ is positive semi-definite, by Theorem 1. This condition rules out the so-called "conjugate points" of the Calculus of Variations.[19]

Now, since $\underline{K}(t; T, \underline{F})$ is symmetric, we have, taking the transpose of the above expression, that

$$\underline{K}(t; T, \underline{F}) = \left[-\underline{\Psi}_{21}(T, t) + \underline{\Psi}'_{11}(T, t)F\right] \cdot \left[\underline{\Psi}'_{22}(T, t) - \underline{\Psi}'_{12}(T, t)F\right]^{-1}$$

substitution of Eq. (2.24) of Lemma 2 yields the desired relation (ii). ‖

Notice that unless $\underline{A}(t)$, $\underline{B}(t)$ and $\underline{C}(t)$ are constant matrices (i.e., Σ is a time-invariant system), the result of Theorem 2 simply replaces the difficult problem of solving the Riccati equation (2.10) by another of similar difficulty, since only in the rarest cases can $\underline{\Psi}(t, T)$ be expressed in analytic form. However, we have shown that the solution of time-varying linear regulator problems involves the same analytic difficulties as the solution of linear differential equations with time-varying coefficients.

Besides being of interest from a theoretical point of view, the results of Theorem 2 present a foundation upon which to build an iterative scheme for the determination of $\underline{K}(t; T, \underline{F})$ on a digital computer. This in fact has been done[16] and the resulting iterative technique is commonly

referred to as the Automatic Synthesis Program (ASP) which we briefly discuss in Section C of Chapter III.

The results which we presented above are valid only when the terminal time T is finite. In order to extend these results to cover the case $T = \infty$, and thereby gain a firm understanding of the nature of the Riccati equation solution $\underline{K}(t; T, \underline{F})$ as regards the parameter T, it is first necessary to introduce the concepts of controllability and observability. This is the object of the next section. In the succeeding sections we present the appropriate results for the solution to the linear regulator problem as $T \rightarrow \infty$, and investigate the stability properties of the resulting optimal system.

E. CONTROLLABILITY AND OBSERVABILITY-DEFINITIONS[11, 12]

The fundamental concepts of controllability and observability occupy central positions if one wishes to investigate properties of the Riccati equation solution $\underline{K}(t; T, \underline{F})$ and, in turn, properties of the optimal solution itself, e.g., stability, speed of response, etc. Among the more useful results which these concepts provide us with, are upper and lower bounds to the optimal cost[13] -- bounds which can be pre-computed prior to actually determining $\underline{K}(t; T, \underline{F})$.

In this section we present the various definitions associated with these linear system concepts. From the definitions we also obtain a necessary and sufficient condition for the invertibility of $\underline{K}(t; T, \underline{F})$.

We consider the linear dynamical system Σ which is characterized by the equations

$$\dot{\underline{x}}(t) = \underline{A}(t)\underline{x}(t) + \underline{B}(t)\underline{u}(t)$$

$$\underline{y}(t) = \underline{C}(t)\underline{x}(t)$$

and we let $\underline{\Phi}(t, t_o)$ denote the transition matrix of Σ. We then have

Definition 2: Σ is completely controllable if and only if for every t there exists a time $t_1(t) > t$ such that the symmetric matrix

$$\underline{W}(t, t_1) = \int_t^{t_1} \underline{\Phi}(t, \tau)\underline{B}(\tau)\underline{B}'(\tau)\underline{\Phi}'(t, \tau)\, d\tau \qquad (2.25)$$

is positive definite.†

Definition 3: Σ is completely observable if and only if for every t there exists a time $t_2(t) > t$ such that the symmetric matrix

$$\underline{M}(t, t_2) = \int_t^{t_2} \underline{\Phi}'(\tau, t)\underline{C}'(\tau)\underline{C}(\tau)\underline{\Phi}(\tau, t)\, d\tau \qquad (2.25')$$

is positive definite.

In the special case when Σ is stationary (i.e., $\underline{A}, \underline{B}, \underline{C}$ are constant matrices) it has been shown (Ref. 11) that

Lemma 3: If Σ = constant then

(a) Σ is completely controllable if and only if

$$\text{rank } [\underline{B}, \underline{A}\underline{B}, \ldots, \underline{A}^{n-1}\underline{B}] = n \qquad (2.26)$$

(b) Σ is completely observable if and only if

$$\text{rank } [\underline{C}', \underline{A}'\underline{C}', \ldots (\underline{A}')^{n-1}\underline{C}'] = n \qquad (2.26')$$

(c) If Σ is completely controllable [observable] then $\underline{W}(t, t_1)[\underline{M}(t, t_2)]$ is invertible for all $t_1 > [t_2 > t]$.

Finally, we wish to make definitions which will remove the dependence of $\underline{W}(t, t_1)$ and $\underline{M}(t, t_2)$ upon t. If $\underline{A}$ and $\underline{B}$ are positive semi-definite

---

† Note that if $\underline{W}(t, t_1)$ is positive definite then so is $\underline{W}(t, t)$, $t > t_1$.

matrices, we use the notation $\underline{A} > \underline{B}$ or $\underline{A} \geq \underline{B}$ to indicate that $\underline{A}-\underline{B}$ is either positive definite or positive semi-definite respectively. Hence

Definition 4: Σ is (i) uniformly completely controllable and (ii) uniformly completely observable if there exists positive constants σ, $\alpha(\sigma)$, $\beta(\sigma)$ such that for all t

(i) $\underline{0} < \alpha(\sigma)\underline{I} < \underline{W}(t, t+\sigma) \leq \beta(\sigma)\underline{I}$

(ii) $\underline{0} < \alpha(\sigma)\underline{I} < \underline{M}(t, t+\sigma) \leq \beta(\sigma)\underline{I}$

Note that, in particular, Definition 4 implies that

$$0 < \alpha(\sigma) \leq \| \underline{W}(t, t+\sigma) \| \leq \beta(\sigma)$$ †

and $$n\alpha(\sigma) \leq \text{tr}\left[W(t, t+\sigma) \right] \leq n\beta(\sigma)$$

and similarly for $\underline{M}(t, t+\sigma)$.

Several cases for which Σ is uniformly completely controllable, in the event that $\underline{A}(t) = \underline{A} = \text{constant}$, are

1. $\underline{B}(t) = \underline{B} = \text{constant}$ and the pair $\{\underline{A}, \underline{B}\}$ is controllable (i.e., Eq. 2.26 is satisfied).
2. $\underline{B}(t) = b(t)\underline{B}$, where $\{\underline{A}, \underline{B}\}$ is controllable and the scalar function of time $b(t)$ satisfies $0 < a \leq b(t) \leq c$ for all time.
3. $\underline{B}_1\underline{B}_1' \leq \underline{B}(t)\underline{B}'(t) \leq \underline{B}_2\underline{B}_2'$ for all t, where $\{\underline{A}, \underline{B}_1\}$ and $\{\underline{A}, \underline{B}_2\}$ are completely controllable.

Similar results hold for Σ to be uniformly completely observable but with $\underline{C}'(t)$ replacing $\underline{B}(t)$.

An immediate relationship between these concepts and our optimal control problem is afforded by the following lemma.

---

† The matrix norm in this expression is the one induced by the inner product on X and is defined by Eqs. A.3 and A.4 of Appendix A.

Lemma 4: $\underline{K}^{-1}(t; T, \underline{0})$ exists for $t < T$ (i) if and (ii) only if Σ is completely observable with $t_2(t) \leq T$ †

Proof: (i) If $\underline{K}^{-1}(t; T, \underline{0})$ does not exist then there is a non-zero vector $\underline{x} \in E_n$ such that

$$\frac{1}{2} < \underline{x}, \underline{K}(t; T, \underline{0})\underline{x} > = 0 = \min_{\underline{u}(\cdot) \in \Omega} J(\underline{x}, t, T, \underline{u})) = J^*(\underline{x}, t, T)$$

But in order for $J^*(\underline{x}, t, T)$ to equal zero it is necessary that $\underline{u}^*(t) \equiv \underline{0}$. Consequently

$$0 = J^*(\underline{x}, t, T) = \frac{1}{2} \int_t^T < \underline{x}(\tau), \underline{C}'(\tau)\underline{C}(\tau)\, \underline{x}(\tau) > d\tau$$

Since the motion is free ($\underline{u} = \underline{0}$) we have $\underline{x}(\tau) = \underline{\Phi}(\tau, t)\underline{x}$ and so

$$0 = \frac{1}{2} < \underline{x}, \left[\int_t^T \underline{\Phi}'(\tau, t)\underline{C}'(\tau)\underline{C}(\tau)\, \underline{\Phi}(\tau, t)\, d\tau \right] \underline{x} >$$

which implies that $\underline{M}(t, T)$ is singular and so $\underline{M}(t, t_2)$ is singular for all $t_2 < T$.

(ii) If Σ is not completely observable with $t_2 \leq T$, then $\underline{M}(t, T)$ is singular and so there is a non-zero $x \in E_n$ such that $< \underline{x}, \underline{M}(t, T)\underline{x} > = 0$. Hence

$$J(\underline{x}, t, T, \underline{u}(\cdot))\Big|_{\underline{u}(\cdot) = \underline{0}} = \frac{1}{2} < \underline{x}, \underline{M}(t, T)\, \underline{x} > = 0$$

which implies that $\underline{K}(t; T, \underline{0})$ is singular. ‖

Note that if $\underline{K}(t)$ is invertible, then $\underline{K}(t; T, \underline{0})$ will be positive definite since $\underline{K}(t; T, \underline{0}) \geq \underline{0}$ by Lemma 1. On the basis of Lemma 4 we have the following corollaries whose proofs are immediate.

---

† $t_2(t)$ is the first time $t_2 > t$ for which $\det \underline{M}(t, t_2) \neq 0$.

Corollary 1: If Σ is completely observable and if $\underline{K}^{-1}(t; T, \underline{0})$ exists, then $\underline{K}^{-1}(\tau; T, \underline{0})$ will exist for all $\tau < t$.

Corollary 2: If $\underline{K}(t; T, \underline{0})$ is invertible then so is $\underline{K}(t; T, \underline{F})$ for all $\underline{F} \geq \underline{0}$.

Finally, we mention the fact that if $\underline{F} > \underline{0}$ then $\underline{K}(t; T, \underline{F})$ is positive definite for all $t \leq T$ irrespective of the observability of Σ.

F. THE OPTIMAL LINEAR REGULATOR SOLUTION FOR $T = \infty$

Having established the required preliminaries we are now in a position to investigate the case where we allow the terminal time $T \to \infty$. We shall show that, under suitable hypotheses, $\lim \underline{K}(t; T, \underline{0})$ exists as $T \to \infty$. We also investigate the stability of the optimal closed-loop system in this case, noting that, in general, optimality does not imply stability.

In what follows we assume $\underline{F} = \underline{0}$ to avoid the mathematical subtleties of "weighing" a terminal state $\underline{x}(T)$ as $T \to \infty$. Under such conditions it is possible to show

Theorem 3:[2] If Σ is completely controllable, then

$$\lim_{T \to \infty} \underline{K}(t; T, \underline{0}) = \bar{\underline{K}}(t)$$

(i) exists for all t, and (ii) $\bar{\underline{K}}(t)$ satisfies the Riccati equation

$$\dot{\underline{K}}(t) = -\underline{K}(t)\underline{A}(t) - \underline{A}'(t)\underline{K}(t) - \underline{C}'(t)\underline{C}(t) + \underline{K}(t)\underline{B}(t)\underline{B}'(t)\underline{K}(t)$$

With some abuse of notation, we shall henceforth call $\bar{\underline{K}}(t)$ the "equilibrium" solution of the Riccati equation.

Having established the fact that $\lim_{T \to \infty} \{\min_{\underline{u}} J(\underline{x}, t, T, \underline{u})\}$ exists and is given by $\frac{1}{2} < \underline{x}, \bar{\underline{K}}(t)\underline{x} >$, we would now like to show that it is equal to $\min_{\underline{u}} \{\lim_{T \to \infty} J(\underline{x}, t, T, \underline{u})\} = \min_{\underline{u}} J(\underline{x}, t, \infty, \underline{u})$, and that this minimum is achieved when

$$\underline{u}(\underline{x}, t) = -\underline{B}'(t)\,\bar{\underline{K}}(t)\,\underline{x}(t)$$

so that $\bar{\underline{K}}(t)$ is associated with a meaningful optimal linear regulator problem which is defined on the infinite interval. This is indeed the case and we can prove

Theorem 4:[2] Assuming $\underline{F} = \underline{0}$ and $T = \infty$, we have

$$\min_{\underline{u}(\cdot)} J(\underline{x}, t, \infty, \underline{u}(\cdot)) = \lim_{T \to \infty} J^*(\underline{x}, t, T)$$

$$= \frac{1}{2} < \underline{x}, \bar{\underline{K}}(t)\,\underline{x} > \qquad (2.27)$$

and the minimum is achieved at $\underline{u}(\cdot) = \underline{u}^*(\underline{x}(t), t)$ where

$$\underline{u}^*(\underline{x}(t), t) = -\underline{B}'(t)\,\bar{\underline{K}}(t)\,\underline{x}(t) \qquad (2.28)$$

Inasmuch as the solution to our control problem for $T = \infty$ is well-defined over the infinite interval $[t_o, \infty]$, it is meaningful to ask questions concerning the stability of the optimal closed-loop system with the control law 2.28. We wish to obtain conditions which will guarantee stability (relative to the equilibrium solution $\underline{x}(t) \equiv \underline{0}$), noting that, in general, a system's optimality does not preclude its stability.

Our main result is

Theorem 5:[2, 16] If Σ is uniformly completely controllable and uniformly completely observable, the controlled system

$$\dot{\underline{x}}(t) = [\underline{A}(t) - \underline{B}(t)\underline{B}'(t)\bar{\underline{K}}(t)]\,\underline{x}(t) \tag{2.29}$$

is uniformly asymptotically stable and

$$V(\underline{x}, t) = J^*(\underline{x}, t, \infty) = \frac{1}{2} < \underline{x}, \bar{\underline{K}}(t)\,\underline{x} > \tag{2.30}$$

is a suitable Lyapunov function.[20]

If Σ is not uniformly completely observable and controllable the optimal closed loop system may be unstable; hence mere observability and controllability are not sufficient to assure stability. For example, consider a first order system characterized by the equations, $\lambda > 0$,

$$\dot{x}(t) = ax(t) + e^{\lambda t} u(t)$$

Σ:

$$y(t) = ce^{-\lambda t} x(t)$$

Note that Σ is completely controllable and completely observable but is neither uniformly completely controllable nor observable. In fact

$$W(t, t+\sigma) = e^{2\lambda t}\left[\frac{e^{2(\lambda-a)\sigma} - 1}{2(\lambda-a)}\right]$$

and

$$M(t, t+\sigma) = c^2 e^{-2\lambda t}\left[\frac{e^{2(a-\lambda)\sigma} - 1}{2(a-\lambda)}\right]$$

so that there exists no constants $\alpha(\sigma)$, $\beta(\sigma)$ such that for all t, (i) $W(t, t+\sigma) \leq \beta(\sigma)$ and (ii) $M(t, t+\sigma) \geq \alpha(\sigma) > 0$. Consequently, Definition 4 is not satisfied for this system (unless $\lambda = 0$). Note, however, that for all t, $W(t, t+\sigma)$ and $M(t, t+\sigma)$ are always strictly positive.

The Riccati equation corresponding to the cost functional

$$J(x_o, 0, \infty, u(\cdot)) = \frac{1}{2} \int_0^\infty [y^2(t) + u^2(t)]\, dt$$

is $$\dot{k}(t) = -2ak(t) - e^{-2\lambda t} c^2 + e^{2\lambda t} k^2(t)$$

The equilibrium solution $\bar{k}(t)$ is given by

$$\bar{k}(t) = \lim_{T \to \infty} k(t; T, 0) = e^{-2\lambda t} \hat{k}$$

where $$\hat{k} = (a-\lambda) + \sqrt{(a-\lambda)^2 + c^2}$$

The optimal feedback control law is

$$u^*(x(t), t) = -e^{-\lambda t} \hat{k}\, x(t)$$

and the optimal closed-loop system is described by

$$\dot{x}(t) = \left[\lambda - \sqrt{(a-\lambda)^2 + c^2}\right] x(t) = \bar{\lambda} x(t)$$

which is a linear, time-invariant equation. Consequently, if

$$0 < a < 2\lambda$$

there always exists a $c^2 > 0$, such that $\bar{\lambda} > 0$. Hence, the solutions of the optimal system, being of the form $e^{\bar{\lambda} t} x_o$, will be unstable.

Note that in general the "equilibrium" solution $\bar{\underline{K}}(t)$ of the Riccati equation will be time-varying. There is, however, an important class of problems for which $\bar{\underline{K}}(t)$ is a constant matrix and for which the optimal control is simply a linear, time-invariant feedback control law. For this class of problems, Σ is time-invariant and is characterized by the equations

$$\Sigma: \quad \begin{aligned} \dot{\underline{x}}(t) &= \underline{A}\underline{x}(t) + \underline{B}\underline{u}(t) \\ \underline{y}(t) &= \underline{C}\underline{x}(t) \end{aligned}$$

We assume that Σ is completely controllable and we seek the control $\underline{u}^*(\cdot)$ which minimizes

$$J(\underline{x}, t, \infty, \underline{u}(\cdot)) = \frac{1}{2} \int_t^\infty \left[< \underline{y}(\tau), \underline{y}(\tau) > + < \underline{u}(\tau), \underline{u}(\tau) > \right] d\tau$$

By Theorem 4, the solution to this regulator problem is given by

$$\underline{u}^*(\underline{x}(t), t) = -\underline{B}'\bar{\underline{K}}(t)\,\underline{x}(t) \tag{2.31}$$

where $\bar{\underline{K}}(t)$ is the equilibrium solution of the Riccati equation

$$\frac{d}{dt}\underline{K}(t) = -\underline{K}(t)\underline{A} - \underline{A}'\underline{K}(t) - \underline{C}'\underline{C} + \underline{K}(t)\underline{B}\underline{B}'\underline{K}(t) \tag{2.32}$$

i.e.,

$$\bar{\underline{K}}(t) = \lim_{T \to \infty} \underline{K}(t; T, \underline{0})$$

Finally, the optimal cost is given by

$$J^*(\underline{x}, t, \infty) = \frac{1}{2} < \underline{x}, \bar{\underline{K}}(t)\, \underline{x} > \tag{2.33}$$

In order to determine $\bar{\underline{K}}(t)$, we first note that since Σ is constant, the matrix $\underline{\Psi}(t, T)$ appearing in Theorem 2 is given by $\underline{\Psi}(t-T)$, more specifically by

$$\underline{\Psi}(t-T) = e^{\underline{Z}(t-T)}$$

Therefore, by virtue of Theorem 2, we deduce that

$$\underline{K}(t; T, \underline{0}) = \underline{K}(t-T, \underline{0})$$

and consequently $\bar{\underline{K}}(t) = \lim_{T \to \infty} \underline{K}(t; T, \underline{0})$ will be a <u>constant</u>, since

$$\bar{\underline{K}} = \lim_{T \to \infty} \underline{K}(t_1 - T, \underline{0}) = \lim_{T \to \infty} \underline{K}(t_2 - T, \underline{0}) \quad \text{for all } t_1, t_2.$$

Now, since $\bar{\underline{K}}(t) = \bar{\underline{K}}$ is a constant (positive semi-definite) matrix which must satisfy the Riccati equation (2.32) we see that $\bar{\underline{K}}$ is a solution of the <u>algebraic</u> system of equations

$$\underline{0} = -\underline{K}\,\underline{A} - \underline{A}'\underline{K} - \underline{C}'\underline{C} + \underline{K}\,\underline{B}\,\underline{B}'\underline{K} \tag{2.34}$$

$\bar{\underline{K}}$ must be at least positive semi-definite. However, Eq. (2.34), being a system of quadratic equations, may possess more than one solution, in fact there may exist more than one positive semi-definite solution. At this point we would like conditions guaranteeing the existence of a <u>unique</u> positive semi-definite solution of Eq. (2.34). The requirement for this is the observability of Σ, (see Definition 3), and in Appendix B we prove

Theorem 6:[7] If the time-invariant system Σ is completely controllable and completely observable then

(a) The algebraic equation

$$\underline{0} = \underline{K}\underline{A} + \underline{A}'\underline{K} + \underline{C}'\underline{C} - \underline{K}\,\underline{B}\underline{B}'\underline{K} \tag{2.35}$$

cannot possess a positive semi-definite solution, but may possess a positive definite solution.

(b) $\bar{\underline{K}} = \lim_{T \to \infty} \underline{K}(t; T, \underline{0})$ is the unique positive definite solution of Eq. (2.35)

(c) The optimal closed-loop system

$$\dot{\underline{x}}(t) = (\underline{A} - \underline{B}\,\underline{B}'\bar{\underline{K}})\,\underline{x}(t) \tag{2.36}$$

is asymptotically stable, i.e., $\text{Re}\ \lambda_i(\underline{A} - \underline{B}\underline{B}'\bar{\underline{K}}) < 0$ and

$$V(\underline{x}) = \frac{1}{2} < \underline{x}, \bar{\underline{K}}\underline{x} > \tag{2.37}$$

is a suitable Lyapunov function.

Note that even under the assumption of complete observability, there may exist indefinite solutions to Eq. (2.35). (By Theorem 6 there will not exist any positive semi-definite solutions.) However, there will only be one positive definite solution, and only this one will be associated with our optimization problem--although isolating this solution may be extremely tedious. In Chapter III we shall discuss several iterative schemes for the determination of $\bar{\underline{K}}$.

This concludes our abridgment of the optimal linear regulator problem. We have explicitly defined this problem and we have investigated its solution in terms of the solution to a non-linear matrix differential equation (the Riccati equation). We then obtained some properties of the Riccati equation solution $\underline{K}(t; T, \underline{F})$, in particular its relationship to the optimal cost $J^*(\underline{x}, t, T)$, as well as a representation theorem (Theorem 2) giving a closed-form, explicit expression for $\underline{K}(t; T, \underline{F})$. Finally, we extended our results to include the case $T = \infty$ and examined the stability properties of the optimal system using the fundamental linear system concepts of controllability and observability. In the next chapter we investigate some further properties of the Riccati equation solution and focus our attention on computational schemes which one may use to determine $\underline{K}(t; T, \underline{F})$.

CHAPTER III

RICCATI EQUATION COMPUTATIONS

In the foregoing chapter we have defined the linear regulator problem and we have studied the properties of its solution. In particular, the key role played by the matrix Riccati equation has been delineated.

Having established a theoretical foundation for the study of linear regulator systems, we will now begin to investigate such associated problems as methods of implementing the optimal control, computational schemes for the solution of the Riccati equation, etc. These problems are slightly more of an engineering nature than of a theoretical one, and their analysis is undertaken in this chapter.

A. IMPLEMENTATION OF THE OPTIMAL FEEDBACK SYSTEM

In Chapter II we saw that the optimal control may be represented as a feedback control law in the form

$$\underline{u}^*(t) = -\underline{B}'(t)\,\underline{K}(t)\,\underline{x}(t) = -\underline{L}^*(t)\underline{x}(t) \qquad (3.1)$$

Thus, the system state at time t is operated on by the linear transformation $\underline{K}(t)$ (which is the Riccati equation solution) and then by the linear transformation $-\underline{B}'(t)$ to generate the optimal control. The optimal feedback system is therefore linear and time-varying, its behavior is governed by the matrix $\underline{K}(t)$, inasmuch as $\underline{B}(t)$ is known. Figure 3.1 shows the structure of the optimal feedback system.

The positive definite matrix $\underline{K}(t)$ is central to the implementation of the optimal control as a feedback law. In order to construct this optimal feedback controller it is necessary to compute and/or implement $\underline{K}(t)$

in some manner. We recall that $\underline{K}(t)$, for $t \in [t_o, T]$ is the (unique) solution of the matrix Riccati equation

$$\dot{\underline{K}}(t) = -\underline{K}(t)\,\underline{A}(t) - \underline{A}'(t)\underline{K}(t) - \underline{C}'(t)\underline{C}(t) + \underline{K}(t)\underline{B}(t)\underline{B}'(t)\underline{K}(t) \quad (3.2)$$

with the boundary condition

$$\underline{K}(T) = \underline{F} \quad (3.3)$$

Equation 3.2 is nonlinear and for this reason it seldom admits closed-form solutions. Therefore, we must compute $\underline{K}(t)$ using a digital computer. There are numerous computational schemes for the solution of ordinary differential equations[23] and theoretically any desired degree of computational accuracy can be obtained. By using such a scheme it is possible to compute $\underline{K}(t)$ for $t < T$ by solving the Riccati equation 3.2 backwards in time, starting from the boundary condition (3.3) at the terminal time. This computation can be done "off-line", before the control system is placed into actual operation. Once $\underline{K}(t)$ for $t \in [t_o, T]$ is computed, the gain matrix $\underline{L}^*(t) = \underline{B}'(t)\underline{K}(t)$ is stored on tape in the system's feedback controller and is played back upon command in forward time to generate the optimal control according to Eq. (3.1).

All is not as simple as it sounds, however. For multi-input, multi-output systems we may be required to store a large number (depending upon the dimension of $\underline{L}^*(t)$) of time-varying signals on tape. Each signal requires its own playback head and associated playback circuitry. In addition, all signals must be played back in time-synchronization with each other. The demands which these tasks place upon the design engineer can therefore become formidable if not overpowering. We shall have more to say on the problem of tape storage in Chapter IV.

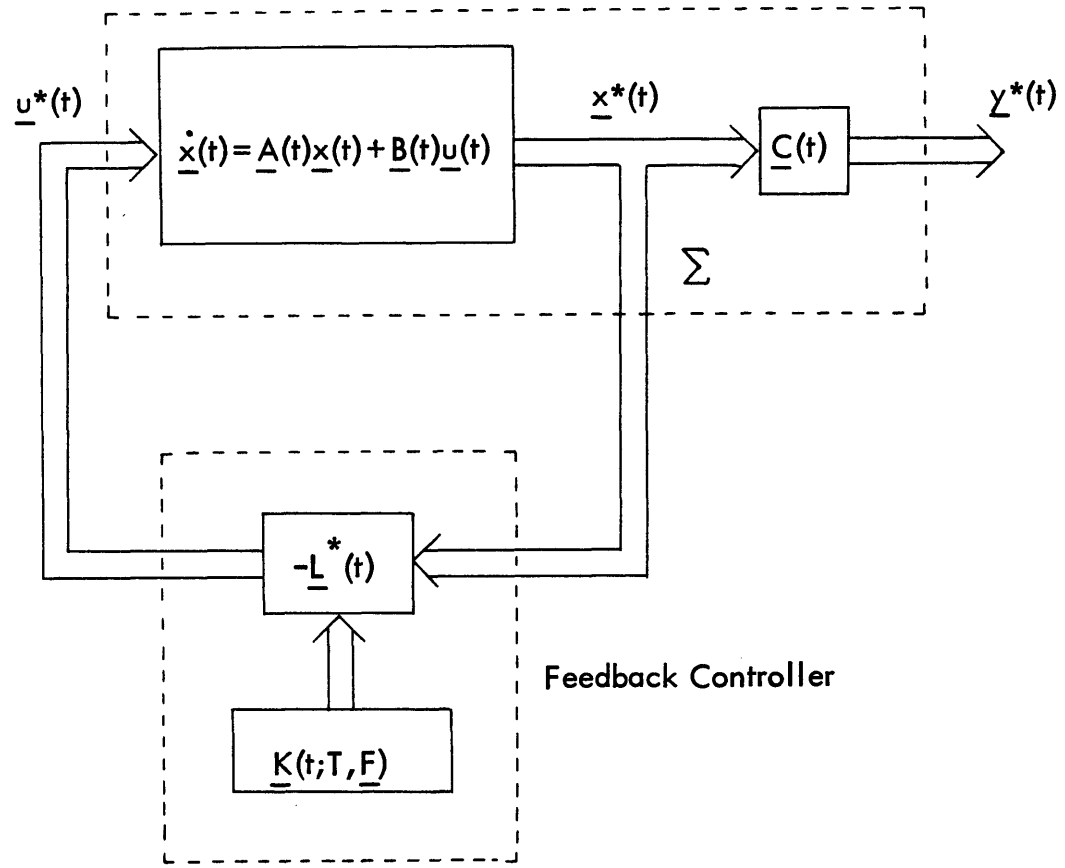


Fig.3.1 Structure of the Optimal Regulator System

$$u^*(t) = -L^*(t)x(t) = -B'(t)K(t;T,F)x^*(t)$$

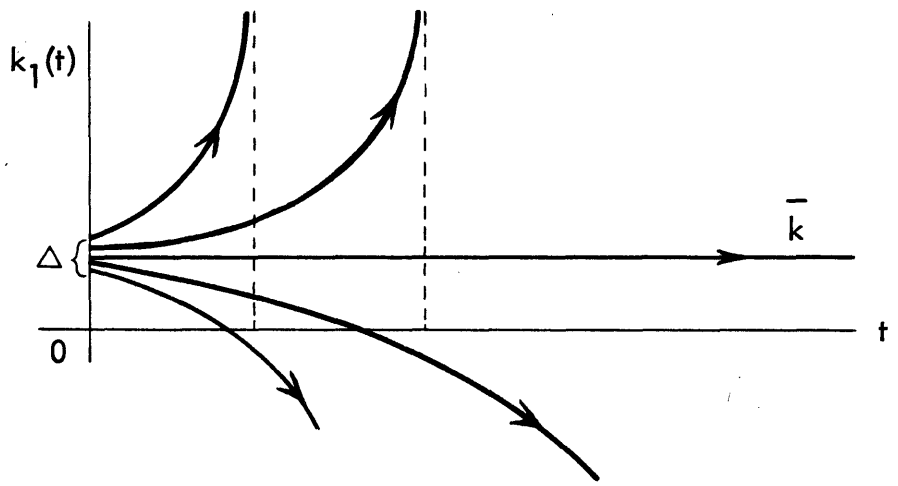


Fig. 3.2 Instability of the Scalar Riccati Equation

As an alternate scheme to computing $\underline{K}(t)$ off-line and storing its elements on tape, we may generate $\underline{K}(t)$ on-line while the system is in operation as follows. We first use a computer to solve the Riccati equation backwards in time to determine the matrix $\underline{K}(t_o)$. We need then store only the elements of the matrix $\underline{K}(t_o)$ in our control system and compute $\underline{K}(t)$ for $t > t_o$ in real time by including a small, specific purpose digital computer into our feedback loop. This computer would have to integrate Eq. 3.2 forward in time starting from the initial condition $\underline{K}(t_o)$. The important thing to remember is that since $\underline{K}(t)$ is independent of the system state, the matrix $\underline{K}(t_o)$ may be precomputed (once $\underline{A}(t)$, $\underline{B}(t)$ and $\underline{C}(t)$ have been specified). In the sequel we shall show that this method is unsatisfactory and that it is not possible to generate $\underline{K}(t)$ accurately in an on-line manner.

In the following sections we will examine the computational properties of the Riccati equation and discuss various computational schemes which are applicable for its solution.

B. STABILITY PROPERTIES OF THE RICCATI EQUATION

Any study of numerical schemes for solving a differential equation should be preceded by an investigation of the stability properties of the equation itself. This is necessary because often the result of such an investigation will favor one computational scheme over another. Consequently, we shall first examine some of the stability properties of the matrix Riccati equation.

Let us suppose that our system Σ is uniformly completely controllable and uniformly completely observable. Let $\underline{F}_1$ and $\underline{F}_2$ be arbitrary positive semi-definite matrices, and let $\underline{K}_1(t)$ and $\underline{K}_2(t)$ be the (unique) solutions of the Riccati equation 3.2 with the respective boundary conditions $\underline{K}_1(T) = \underline{F}_1$ and $\underline{K}_2(T) = \underline{F}_2$. We wish to examine the difference $\delta\underline{K}(t) = \underline{K}_2(t) - \underline{K}_1(t)$ in order to determine whether the two Riccati solutions $\underline{K}_1(t)$ and $\underline{K}_2(t)$ diverge or converge for $t < T$.

Since $\underline{K}_1(t)$ and $\underline{K}_2(t)$ are both solutions of equation 3.2 (but having different values at T), $\delta\underline{K}(t)$ obeys the differential equation, where $\underline{A}_1(t) \triangleq \underline{A}(t) - \underline{B}(t)\underline{B}'(t)\underline{K}_1(t)$,

$$\frac{d}{dt}\delta\underline{K}(t) = -\delta\underline{K}(t)\underline{A}_1(t) - \underline{A}_1'(t)\delta\underline{K}(t) + \delta\underline{K}(t)\underline{B}(t)\underline{B}'(t)\delta\underline{K}(t) \qquad (3.4)$$

with the boundary conditions

$$\delta\underline{K}(T) = \underline{K}_2(T) - \underline{K}_1(T) = \underline{F}_2 - \underline{F}_1 \qquad (3.5)$$

Equation (3.4) is derived by writing

$$\begin{aligned}\dot{\underline{K}}_2(t) &= -\underline{K}_2(t)\underline{A}(t) - \underline{A}'(t)\underline{K}_2(t) - \underline{C}'(t)\underline{C}(t) + \underline{K}_2(t)\underline{B}(t)\underline{B}'(t)\underline{K}_2(t) \\ &= -\underline{K}_2(t)\underline{A}_1(t) - \underline{A}_1'(t)\underline{K}_2(t) - \underline{C}'(t)\underline{C}(t) + \underline{K}_2(t)\underline{B}(t)\underline{B}'(t)\underline{K}_2(t) \\ &\quad -\underline{K}_2(t)\underline{B}(t)\underline{B}'(t)\underline{K}_1(t) - \underline{K}_1(t)\underline{B}(t)\underline{B}'(t)\underline{K}_2(t)\end{aligned}$$

and subtracting

$$\dot{\underline{K}}_1(t) = -\underline{K}_1(t)\underline{A}_1(t) - \underline{A}_1'(t)\underline{K}_1(t) - \underline{C}'(t)\underline{C}(t) - \underline{K}_1(t)\underline{B}(t)\underline{B}'(t)\underline{K}_1(t)$$

By virtue of Theorem 1 and Lemma 4, both $\underline{K}_1(t)$ and $\underline{K}_2(t)$ exist for all $t < T$ and are invertible for $t < T-\sigma$. Consequently it is possible to investigate the stability properties of Eq. (3.4) as $t \to -\infty$ by use of the Lyapunov function

$$V(\delta\underline{K}, t) = \frac{1}{2}\,\mathrm{tr}(\delta\underline{K}\cdot\underline{K}_1^{-1})^2 \qquad (3.6)$$

In References 2 and 13 it is shown that

Theorem 7: If Σ is uniformly completely controllable and uniformly completely observable, then any two solutions $\underline{K}(t;T,\underline{F}_1)$ and $\underline{K}(t;T,\underline{F}_2)$ of the Riccati equation are uniformly asymptotically stable to each other as $t\to-\infty$ and

$$V(\delta\underline{K}, t) = \frac{1}{2}\,\mathrm{tr}\,(\delta\underline{K}\cdot\underline{K}_1^{-1})^2 \qquad (3.6)$$

is a suitable Lyapunov function.

This theorem states that the difference $\delta\underline{K}(t)$ between any two solutions tends to zero as $t\to-\infty$. Consequently, all solutions of the Riccati equation approach each other as $t\to-\infty$. Recalling that the equilibrium solution of the Riccati equation, $\bar{\underline{K}}(t)$, is well-defined for all t, we can summarize this convergence property as a lemma.

Lemma 5: If Σ is uniformly c.c. and uniformly c.o. then for any T and any positive semi-definite matrix $\underline{F}$, the solution $\underline{K}(t;T,\underline{F})$ converges asymptotically to $\bar{\underline{K}}(t)$ as $t\to-\infty$, where $\bar{\underline{K}}(t)$ is the equilibrium solution of the Riccati equation.

In other words, Lemma 5 states that for any Riccati equation solution, the effect of the terminal condition $\underline{K}(T) = \underline{F}$ is gradually "forgotten" as $t\to-\infty$.

Since asymptotic stability in negative time implies instability in positive time, any solution $\underline{K}(t;T,\underline{F})$ is unstable as $t\to+\infty$. That

is to say, the difference between any two Riccati solutions $\underline{K}_1(t)$ and $\underline{K}_2(t)$ tending asymptotically to zero as $t \to -\infty$ implies that $\delta\underline{K}(t)$ will increase without limit as $t \to +\infty$. Consequently, one cannot compute $\underline{K}(t; T, \underline{F})$ on the interval $[t_o, T]$ by integrating Eq. (3.2) in the forward time direction, starting with the initial condition $\underline{K}(t_o)$. This procedure is computationally unstable; a small numerical error in $\underline{K}(t_o)$ will manifest itself in a large error in $\underline{K}(t)$ for $t > t_o$. To make this notion more precise we consider the following first order example.

Let Σ be characterized by the equations

$$\Sigma: \quad \begin{aligned} \dot{x}(t) &= a\,x(t) + u(t) \\ y(t) &= c\,x(t) \end{aligned}$$

and let the cost functional $J(x, 0, T, u)$ be given by

$$J(x, 0, T, u) = \frac{1}{2} f\, x^2(T) + \frac{1}{2} \int_0^T [y^2(\tau) + u^2(\tau)]\, d\tau$$

The Riccati equation associated with the solution of the optimization problem is

$$\dot{k}(t) = -2a\,k(t) - c^2 + k^2(t)\ ;\ k(T) = f \geq 0$$

The solution of this equation is given by[1]

$$k(t; T, f) = (\beta - a)\, \frac{\frac{(\beta+a)}{(\beta-a)} + \frac{f-a-\beta}{f-a+\beta}\, e^{2\beta(t-T)}}{1 - \frac{f-a-\beta}{f-a+\beta}\, e^{2\beta(t-T)}}$$

where $\beta = (a^2 + c^2)^{1/2}$.

The equilibrium solution $\bar{k}(t)$ is given by

$$\bar{k}(t) = \lim_{T \to \infty} k(t; T, 0) = \beta + a = \bar{k}$$

We now wish to verify (by example) the instability of the equilibrium solution $\bar{k}$ as $t \to \infty$. For this purpose, then, let $k_1(t)$ be a solution of the Riccati equation which differs from $\bar{k}$ by a small amount Δ at $t = 0$, i.e.,

$$k_1(0) = (\beta+a) + \Delta$$

The expression for $k_1(t)$, $t \geq 0$ is given by

$$k_1(t) = (\beta-a)\frac{\frac{(\beta+a)}{(\beta-a)} + \frac{\Delta}{2\beta+\Delta} e^{2\beta t}}{1 - \frac{\Delta}{2\beta+\Delta} e^{2\beta t}}$$

The expression for $\delta k(t) = k_1(t) - \bar{k}$ is

$$\delta k(t) = \frac{2\beta e^{2\beta t}}{\frac{(2\beta+\Delta)}{\Delta} - e^{2\beta t}}$$

From either the expression for $k_1(t)$ or for $\delta k(t)$ we see that for any value of Δ, the solution $k_1(t)$ diverges from $\bar{k}$. In fact, if Δ is a small positive number, the solution $k_1(t)$ fails to exist for

$$t \geq \hat{t} = \frac{1}{2\beta} \ln \left(\frac{2\beta+\Delta}{\Delta}\right)$$

The approximate shape of the solutions $k_1(t)$ for different values of Δ are shown in Fig. 3.2.

This simple example demonstrates that one cannot compute $\bar{k}$ for $t > 0$ from knowledge of $\bar{k}(0)$. A small error at any step of the computation will manifest itself as time increases in the positive direction.

It should be remarked that although $\bar{k}$ and $k_1(t)$ diverge in the forward time direction, this does not necessarily imply that the state trajectories corresponding to the feedback gains diverge. In fact it

is quite conceivable that as long as $k_1(t)$ exists and is positive, the state trajectory resulting from application of $k_1(t)$ will be "close" in some sense to the trajectory resulting from use of $\bar{k}$.

If such is the case, then in the multi-dimensional problem, the value of $\underline{K}(t_o;T,\underline{F})$ may provide a feasible means of generating state trajectories which "approximate" the optimal state trajectory over the interval $[t_o, T]$, even though computing $\underline{K}(t;T,\underline{F})$ for $t > t_o$ by integrating the Riccati equation forward in time is subject to large numerical errors. In References 13 and 21 this concept is investigated in more detail and additional results are presented concerning the behavior of the difference between two Riccati equation solutions.

C. OFF-LINE, NUMERICAL SOLUTION OF THE RICCATI EQUATION

In the foregoing section we showed that due to the computational instability of the Riccati equation solutions it is unfeasible to store $\underline{K}(t_o)$ in our feedback controller and then compute $\underline{K}(t)$ for $t > t_o$ in an on-line manner. Computation errors and computer round-off errors at any step in the forward time integration of the Riccati equation will become magnified as time increases, thereby making it virtually impossible to generate $\underline{K}(t)$ accurately in real time. This phenomenon has been illustrated in the first order case by Figure 3.2.

Consequently we must abandon the hope of computing $\underline{K}(t)$ on-line and revert to the other alternative of precomputing $\underline{K}(t)$ for $t \in [t_o, T]$, storing $\underline{L}^*(t) = \underline{B}'(t)\underline{K}(t)$ on tape, and playing the tape back upon command in real-time. The computation of $\underline{K}(t)$ must therefore be done off-line,

before the control system is placed into actual operation. This fact leads us to consider numerical techniques for the off-line computation of $\underline{K}(t; T, \underline{F})$.

There are numerous schemes which are available for the numerical solution of ordinary differential equations by use of a digital computer. Each one offers its own advantages and disadvantages. Basically, these schemes may be classified as either "one-step" or "multi-step" methods.[23] In this section we shall outline several one-step methods which are applicable to the off-line solution of the Riccati equation, and discuss the relative merits of each scheme.

We seek a numerical solution of the Riccati equation† over the interval $t_o \leq t \leq T$. To do this we first sub-divide the interval of interest into N subintervals of length δ, where δ is small. We therefore obtain a sequence of times $t_o, t_1, t_2, \ldots t_N = T$ such that

$$t_{i-1} = t_i - \delta \qquad \text{for } i = 1, 2, \ldots, N \tag{3.7}$$

We wish to generate a sequence of nxn matrices $\underline{K}_1, \underline{K}_2, \ldots, \underline{K}_i, \ldots, \underline{K}_N = \underline{F}$, such that $\underline{K}_i$ approximates the true solution $\underline{K}(t; T, \underline{F})$ at $t = t_i$, i.e.,

$$\underline{K}_i \approx \underline{K}(t_i; T, \underline{F}) \tag{3.8}$$

and such that the matrices $\underline{K}_i$ are generated by a recursion relation of the form

$$\underline{K}_{i-1} = \underline{g}_i(\underline{K}_i ; \delta) \tag{3.9}$$

---

† Naturally we will obtain only an approximation to $\underline{K}(t)$.

where $\underline{g}_i$ for $i = 1, \ldots, N$ is a mapping from the space of nxn matrices into itself. Some examples of such schemes are:

1. Runge-Kutta Scheme This is a popular computational scheme for digital computers because of its programming efficiency. Its application to the solution of the Riccati equation has been investigated in Reference 25. The basic iterative scheme is as follows.

We first define

$$\underline{f}(t, \underline{K}) = -\underline{K}\underline{A}(t) - \underline{A}'(t)\underline{K} - \underline{C}'(t)\underline{C}(t) + \underline{K}\underline{B}(t)\underline{B}'(t)\underline{K} \qquad (3.10)$$

The Runge-Kutta method then uses the following recursion formula to compute $\underline{K}_{i-1}$ given $\underline{K}_i$:

$$\underline{K}_{i-1} = \underline{g}_i(\underline{K}_i; \delta) = \underline{K}_i - \frac{\delta}{6}\left[\underline{G}_i^{(1)} + 2\underline{G}_i^{(2)} + 2\underline{G}_i^{(3)} + \underline{G}_i^{(4)}\right] \qquad (3.11)$$

with

$$\underline{K}_N = \underline{F} \qquad (3.12)$$

The nxn matrices $\underline{G}_i^{(j)}$ are given by

$$\underline{G}_i^{(1)} = \underline{f}(t_i, \underline{K}_i) \qquad (3.13a)$$

$$\underline{G}_i^{(2)} = \underline{f}(t_i - \frac{\delta}{2}, \underline{K}_i + \frac{1}{2}\underline{G}_i^{(2)}) \qquad (3.13b)$$

$$\underline{G}_i^{(3)} = \underline{f}(t_i - \frac{\delta}{2}, \underline{K}_i + \frac{1}{2}\underline{G}_i^{(2)}) \qquad (3.13c)$$

$$\underline{G}_i^{(4)} = \underline{f}(t_i - \delta, \underline{K}_i + G_i^{(3)}) \qquad (3.13d)$$

Using this method we can theoretically achieve any degree of approximation to the true solution $\underline{K}(t; T, \underline{F})$ by choosing δ sufficiently small.

2. Automatic Synthesis Program (ASP)[16] Another difference method for determining $\underline{K}(t)$ is based on Theorem 2. From this theorem we have, if $\underline{K}(T) = \underline{F}$, that $\underline{K}(T-\delta) = \underline{K}(t_{N-1})$ is exactly given by

$$\underline{K}(t_{N-1}; t_N, \underline{F}) = [\underline{\Psi}_{21}(t_{N-1}, t_N)+\underline{\Psi}_{22}(t_{N-1}, t_N)\underline{F}] \cdot [\underline{\Psi}_{11}(t_{N-1}, t_N)+\underline{\Psi}_{12}(t_{N-1}, t_N)\underline{F}]^{-1} \tag{3.14}$$

where the 2nx2n transition matrix $\underline{\Psi}(t, \tau)$ has been defined in Eq. (2.23).

Hence we have recursively from Eq. 3.14 that

$$\underline{K}_{i-1} = \underline{g}_i(\underline{K}_i; \delta) = [\underline{\Psi}_{21}(t_i-\delta, t_i)+\underline{\Psi}_{22}(t_i-\delta, t_i)\underline{K}_i] \cdot [\underline{\Psi}_{11}(t_i-\delta, t_i)+\underline{\Psi}_{12}(t_i-\delta, t_i)\underline{K}_i]^{-1} \tag{3.15}$$

where in this case the sequence of generated matrices $\{\underline{K}_i;\ i = 0, \ldots, N-1\}$ are exactly equal to $\underline{K}(t_i; T, \underline{F})$ i.e.,

$$\underline{K}_i = \underline{K}(t_i; t_N, \underline{F}) \qquad \text{for } i = 0, 1, \ldots, N \tag{3.16}$$

Although the scheme gives exact results for $\underline{K}(t_i; T, \underline{F})$, (for any value of δ, incidentally) this fact is overshadowed by the complexity of the method. The computation of $\underline{\Psi}(t_i-\delta, t_i)$ is generally difficult, and at each step in the iteration (3.15) we must invert an nxn matrix. However, if Σ were a constant system, then

$$\underline{\Psi}(t_i-\delta, t_i) = e^{-\delta \underline{Z}} \qquad \text{for all } i$$

where

$$\underline{Z} = \begin{bmatrix} \underline{A} & \vdots & -\underline{B}\underline{B}' \\ \cdots\cdots & \vdots & \cdots\cdots \\ -\underline{C}'\underline{C} & \vdots & -\underline{A}' \end{bmatrix}$$

and the recursive scheme of Eq. (3.15) could be applied with relative ease.

3. Discrete Optimization Method Another method which is useful to determine an approximation to $\underline{K}(t; T, \underline{F})$ is based on the application of discrete optimization techniques.[26] After subdividing our interval $[t_o, T]$ into N equal segments we replace our continuous time system Σ by a discrete time system, Σ_d, over $[t_o, T]$. Σ_d is described by the difference equations

$$\underline{x}_{i+1} = (\underline{I} + \delta \underline{A}_i)\underline{x}_i + \delta \underline{B}_i \underline{u}_i \; ; \; \underline{x}_o = \underline{x}(t_o) \tag{3.17}$$

Σ_d:

$$\underline{y}_i = \underline{C}_i \underline{x}_i \tag{3.18}$$

where $\underline{A}_i$, $\underline{B}_i$ and $\underline{C}_i$ are equal to $\underline{A}(t_i)$, $\underline{B}(t_i)$ and $\underline{C}(t_i)$ respectively.

We also replace the cost functional (2.3) by the summation

$$J_d(\underline{x}_o, t_o, \{\underline{u}_i\}) = \frac{1}{2} <\underline{x}_N, \underline{F}\underline{x}_N> + \frac{\delta}{2} \sum_{i=0}^{N-1} [<\underline{y}_i, \underline{y}_i> + <\underline{u}_i, \underline{u}_i>] \tag{3.19}$$

We wish to determine the control sequence $\{\underline{u}_i^*, i = 0, 1, \ldots, N-1\}$ and the corresponding state sequence $\{\underline{x}_i^*, i = 0, 1, \ldots, N\}$ such that J_d is absolutely minimized. This problem is the so-called "discrete linear regulator problem" and is investigated in Ref. 26, with the result that the optimal control sequence is given by

$$\underline{u}_i^* = -(\underline{I} + \delta \underline{B}_i' \underline{K}_{i+1} \underline{B}_i)^{-1} \underline{B}_i' \underline{K}_{i+1} (\underline{I} + \delta \underline{A}_i) \underline{x}_i^* \tag{3.20}$$

The symmetric matrix $\underline{K}_i$ satisfies the difference equation

$$\begin{aligned} \underline{K}_i &= \underline{g}_i(\underline{K}_{i+1}; \delta) \\ &= (\underline{I} + \delta \underline{A}_i)' [\underline{K}_{i+1} - \delta \underline{K}_{i+1} \underline{B}_i (\underline{I} + \delta \underline{B}_i' \underline{K}_{i+1} \underline{B}_i)^{-1} \underline{B}_i' \underline{K}_{i+1}] (\underline{I} + \delta \underline{A}_i) + \delta \underline{C}_i' \underline{C}_i \end{aligned} \tag{3.21}$$

with the boundary condition $\underline{K}_N = \underline{F}$.

It can also be shown[26] that the matrix $\underline{K}_i$ for $i = 0, 1, \ldots, N$ has the property that

$$J_d^*(\underline{x}_j^*, t_j) \triangleq \frac{1}{2}<\underline{x}_N^*, \underline{F}\underline{x}_N^*> + \frac{\delta}{2} \sum_{i=j}^{N-1} [<\underline{y}_i^*, \underline{y}_i^*> + <\underline{u}_i^*, \underline{u}_i^*>]$$

$$= \frac{1}{2}<\underline{x}_j^*, \underline{K}_j\underline{x}_j^*> \qquad (3.22)$$

This result is notably similar to the result of Lemma 1 for the continuous time regulator problem, and with this in mind we may regard Eq. (3.21) as a discrete analog the the matrix Riccati (differential) equation. In particular (3.22) guarantees that $\underline{K}_i$ will be positive semi-definite for all i.

Examining Eq. (3.21) we note that as $\delta \to 0$ the solution to this equation approaches $\underline{K}(t; T, \underline{F})$, since the difference equation in the limit approaches the Riccati differential equation. Consequently, for small δ it is reasonable to use Eq. (3.21) to generate a sequence of matrices $\underline{K}_i$, $i = 0, \ldots, N$ such that

$$\underline{K}_i \approx \underline{K}(t_i, T, \underline{F})$$

This scheme is simple to implement on a digital computer, the computation time is small and the method is not adversely affected by round-off errors.[27] A computer program for the generation of $\underline{K}_i$ exists[27] and is written in Fortran II for the case when $\underline{A}, \underline{B}$ and $\underline{C}$ are constant matrices. Further research on this discrete approximation scheme is currently being pursued.

4. Comparison of Methods 1-3

Each of the aforementioned algorithms for the off-line computation of the Riccati equation solution offers its own advantages and disadvantages. From a strictly computational point of view, the Runge-Kutta scheme is the simplest to use. No matrix inversion is necessary at any step in the iteration, whereas the other methods require a matrix inversion (which is a time-consuming process for a digital computer, especially if the order of the matrix is large). The Runge-Kutta scheme requires only matrix multiplications and additions thereby making it computationally efficient. On the other hand, this scheme can lead to wildly erroneous results as follows. In using the Runge-Kutta technique we cannot guarantee that every matrix in the sequence $\underline{K}_{N-1}, \underline{K}_{N-2}, \ldots$ generated by Eq. (3.11) will be positive semi-definite because of the discretation error which this scheme introduces at each step.[23] Suppose for example that $\underline{K}_a$ (which is our approximation to the Riccati equation solution at $t = t_a$) is an indefinite matrix for some integer $a < N$. However, the solution of the Riccati equation satisfying $\underline{K}(t_a) = \underline{K}_a$ may fail to exist for $t < t_a$ since the conditions of Theorem 1 are not satisfied for $\underline{K}_a$. Consequently, the Runge-Kutta scheme can "pick-up", and begin integrating along, a Riccati equation solution which has a finite escape time for some $t < t_a$. As experience has shown, it is this very dilemma which makes Euler's method unsuitable for Riccati equation computations, even for extremely small time-increments δ.

The second method which we discussed has the property of yielding exact results (subject to computer round-off errors) for the Riccati equation solution at the times t_i, $i = 0, 1, \ldots, N$. However, this accuracy is achieved at the expense of difficult and time consuming calculations. At each step in the ASP program we must evaluate the transition matrix $\underline{\Psi}(t_i - \delta, t_i)$ and invert the nxn matrix $[\underline{\Psi}_{11}(t_i - \delta, t_i) + \underline{\Psi}_{12}(t_i - \delta, t_i)\underline{K}_i]$. For sufficiently small, δ, however, we can approximate the matrix $\underline{\Psi}(t_i - \delta, t_i)$ by

$$\underline{\Psi}(t_i - \delta, t_i) \approx \begin{bmatrix} \underline{I} - \int_{t_i-\delta}^{t_i} \underline{A}(\tau)d\tau & \vdots & \int_{t_i-\delta}^{t_i} \underline{B}(\tau)\underline{B}'(\tau)d\tau \\ \cdots\cdots\cdots\cdots\cdots & \vdots & \cdots\cdots\cdots\cdots\cdots \\ \int_{t_i-\delta}^{t_i} \underline{C}'(\tau)\underline{C}(\tau)d\tau & \vdots & \underline{I} + \int_{t_i-\delta}^{t_i} \underline{A}'(\tau)\,d\tau \end{bmatrix}$$

In such a case we are sacrificing numerical accuracy for computational simplicity, although we are still left with performing the inversion of an nxn matrix.

The discrete optimization method also requires the inversion of a matrix at each step, however the order of the matrix to be inverted is rxr where r is the number of control inputs. In most control applications, the number of control variables is less than the number (n) of state variables, so that the computational difficulty of the discrete optimization method is generally less than that of the ASP method although greater than that of the Runge-Kutta scheme. Naturally, the results of this method will only be an approximation to the true Riccati equation solution. However, there is an important property which the discrete optimization

method possesses. By Eq. (3.22) we are guaranteed that each matrix in the generated sequence $\underline{K}_{N-1}, \underline{K}_{N-2}, \ldots$ is at least positive semi-definite. Therefore, this sequence of matrices will not diverge from the true Riccati equation solution $\underline{K}(t; T, \underline{F})$ as may be the case for the Runge-Kutta scheme. Hence, for small δ, the discrete optimization scheme will <u>always</u> give a reasonable approximation to $\underline{K}(t; T, \underline{F})$, at the expense of inverting an r x r (positive definite) matrix at each step in the iteration.

In the above we have discussed three methods which are practical for the numerical, off-line, solution of the Riccati equation. There are other schemes which may also be satisfactory and their exploration remains a subject for a considerable amount of further research. We re-emphasize that all Riccati equation computations are done <u>off-line</u>, so that computing time is not <u>the</u> essential factor in our iterative scheme but is subordinate to computational accuracy.

Finally we need mention that in the event Σ is time-invariant and $T = \infty$, any of the above techniques can be used to compute $\bar{\underline{K}}$. By Lemma 5 we may obtain $\bar{\underline{K}}$ as

$$\bar{\underline{K}} = \lim_{i \to -\infty} \underline{K}_i \tag{3.23}$$

where $\underline{K}_{i-1} = \underline{g}_i(\underline{K}_i; \delta)$ and $\underline{K}_N = \underline{F}$

D. COMPUTATION OF $\underline{K}(t; T, \underline{F})$ BY SUCCESSIVE APPROXIMATIONS

In the previous section we discussed schemes which are applicable to the numerical solution of the Riccati equation. These methods share a common basis in their approximation of the nonlinear (Riccati) differential equation by a nonlinear difference equation.

Another method for the off-line solution of nonlinear differential equations is to introduce an associated sequence of linear differential equations whose successive solutions approach the solution of the original nonlinear equation. This is the strategy behind such numerical schemes as Newton's method[18] and the method of successive approximations as advanced by Kalaba[17] (often referred to as "quasilinearization").

In Reference 17, the method of successive approximations is applied to the solution of a first-order Riccati equation. In this section we shall extend Kalaba's results to the matrix case and generate a sequence of approximations to $\underline{K}(t; T, \underline{F})$ which possess certain monotone convergence properties.

In the sequel we shall again use the notation $\underline{A} \geq \underline{B}$ to mean that the matrix $\underline{A} - \underline{B}$ is positive semi-definite, and $\underline{A} > \underline{B}$ to denote that $\underline{A} - \underline{B}$ is positive definite, where both $\underline{A}$ and $\underline{B}$ are arbitrary positive semi-definite matrices. We first prove a result of general interest concerning the matrix $\underline{K}(t; T, \underline{F})$. By using the fact that $\underline{K}(t; T, \underline{F})$ is associated with the optimal control we show

Lemma 6: Let $\underline{V}_L(t)$ denote the (unique symmetric positive semi-definite) solution of the linear matrix differential equation

$$\dot{\underline{V}}(t) = -\underline{V}(t)\left[\underline{A}(t)-\underline{B}(t)\underline{L}(t)\right] - \left[\underline{A}(t) - \underline{B}(t)\underline{L}(t)\right]'\underline{V}(t) - \underline{C}'(t)\underline{C}(t) - \underline{L}'(t)\underline{L}(t) \tag{3.24}$$

satisfying $\underline{V}(T) = \underline{F}$. Then

$$\underline{K}(t; T, \underline{F}) \leq \underline{V}_L(t) \qquad \text{for all } \underline{L}(t) \tag{3.25}$$

Proof: Consider the system Σ with the linear feedback control law $\underline{u}_L(t, \underline{x}(t)) = -\underline{L}(t)\,\underline{x}(t)$, so that the closed-loop system satisfies

$$\dot{\underline{x}}(t) = \left[\underline{A}(t) - \underline{B}(t)\,\underline{L}(t)\right]\,\underline{x}(t)$$

If we let $\underline{\Phi}_L(t, t_o)$ be the transition matrix corresponding to $\underline{A}(t) - \underline{B}(t)\underline{L}(t)$ i.e., $\underline{\Phi}_L(t, t_o)$ satisfies

$$\frac{d}{dt}\underline{\Phi}_L(t, t_o) = \left[\underline{A}(t) - \underline{B}(t)\underline{L}(t)\right]\underline{\Phi}_L(t, t_o); \quad \underline{\Phi}_L(t_o, t_o) = \underline{I}$$

and if $t \in (t_o, T)$ and $\underline{x} \in E_n$, the cost associated with using the control $\underline{u}_L(\cdot)$ is

$$J(\underline{x}, t, T, \underline{u}(\cdot))\Big|_{\underline{u}=\underline{u}_L} = \frac{1}{2}\langle \underline{x}, \underline{V}(t)\underline{x}\rangle \tag{3.26}$$

where

$$\underline{V}(t) = \underline{\Phi}'_L(T, t)\underline{F}\,\underline{\Phi}_L(T, t) + \int_t^T \underline{\Phi}'_L(\tau, t)\left[\underline{C}'(\tau)\underline{C}(\tau) + \underline{L}'(\tau)\underline{L}(\tau)\right]\underline{\Phi}_L(\tau, t)d\tau \tag{3.27}$$

Upon differentiating both sides of the above expression we find that

$$\dot{\underline{V}}(t) = \underline{V}(t)[\underline{A}(t) - \underline{B}(t)\underline{L}(t)] - [\underline{A}(t) - \underline{B}(t)\underline{L}(t)]'\underline{V}(t) - \underline{C}'(t)\underline{C}(t) - \underline{L}'(t)\underline{L}(t) \tag{3.24}$$

with $\underline{V}(T) = \underline{F}$. Expression 3.27 is the unique solution to Eq.(3.24) since Eq.(3.24) is linear is $\underline{V}(t)$. Note that $\underline{V}(t) \geq \underline{0}$. Let $\underline{V}_L(t)$ denote this solution, emphasizing its dependence upon $\underline{L}(t)$. But now, since $\underline{u}_L(\cdot)$ is not the optimal control, we have

$$\frac{1}{2} < \underline{x}, \underline{K}(t; T, \underline{F})\underline{x} > = J(\underline{x}, t, T, \underline{u}(\cdot))\Big|_{\underline{u} = \underline{u}^*}$$

$$\leq J(\underline{x}, t, T, \underline{u}(\cdot))\Big|_{\underline{u} = \underline{u}_L} = \frac{1}{2} < \underline{x}, \underline{V}_L(t)\underline{x} >$$

for any rxn time-varying matrix $\underline{L}(t)$. Hence, since $\underline{x}$ is arbitrary,

$$\underline{K}(t; T, \underline{F}) \leq \underline{V}_L(t)$$

for all $\underline{L}(t)$ as claimed, with equality holding if and only if $\underline{L}(t) = \underline{B}'(t)\underline{K}(t; T, \underline{F})$. ‖

For arbitrary $\underline{L}(t)$, Lemma 6 furnishes us with upper bounds for the solution to the Riccati equation, which may prove helpful in any a priori investigation of the regulator problem by providing upper bounds to the optimal cost $J^*(\underline{x}_o, t_o, T)$.

Henceforth, we shall call the matrix $\underline{V}_L(t)$ given by Eq.(3.27) (and satisfying Eq. 3.24) the cost matrix associated with $\underline{L}(t)$, inasmuch as

$$J(\underline{x}, t, T, \underline{u}(\cdot))\Big|_{\underline{u}(t) = -\underline{L}(t)\underline{x}(t)} = \frac{1}{2} < \underline{x}, \underline{V}_L(t)\,\underline{x} > \tag{3.26}$$

Definition 5: For all $t \in [t_o, T]$, the cost matrix $\underline{V}_L(t)$ associated with the feedback gain matrix $\underline{L}(t)$ is given by

$$\underline{V}_L(t) = \underline{\Phi}'_L(T,t)\underline{F}\underline{\Phi}_L(T,t) + \int_t^T \underline{\Phi}'_L(\tau,t)\left[\underline{C}'(\tau)\underline{C}(\tau)+\underline{L}'(\tau)\underline{L}(\tau)\right]\underline{\Phi}_L(\tau,t)d\tau$$

where $\underline{\Phi}_L(\tau,t)$ is the transition matrix corresponding to $\underline{A}(\tau)-\underline{B}(\tau)\underline{L}(\tau)$.

In Appendix C we discuss this concept of a cost matrix further and we derive various expressions for the difference between cost matrices associated with different $\underline{L}(t)$'s.† These expressions are quite useful, and in fact application of Eq. (C.17) yields a simple expression for $\underline{V}_L(t) - \underline{K}(t;T,\underline{F})$, as follows.

If $\underline{V}_1(t)$ and $\underline{V}_2(t)$ denote the cost matrices corresponding to $\underline{L}_1(t)$ and $\underline{L}_2(t)$, respectively, then from Eq. (C.17) we find that

$$\underline{V}_1(t) - \underline{V}_2(t) = \int_t^T \underline{\Phi}_1(\tau,t)\left[(\underline{L}_1-\underline{L}_2)'(\underline{L}_1-\underline{L}_2) - (\underline{L}_1-\underline{L}_2)'(\underline{B}'\underline{V}_2-\underline{L}_2)\right.$$

$$\left.-(\underline{B}'\underline{V}_2-\underline{L}_2)'(\underline{L}_1-\underline{L}_2)\right]\underline{\Phi}_1(\tau,t)d\tau \qquad (C.17)$$

where $\underline{\Phi}_1(\tau,t)$ satisfies

$$\frac{d}{d\tau}\underline{\Phi}_1(\tau,t) = \left[\underline{A}(\tau)-\underline{B}(\tau)\underline{L}_1(\tau)\right]\underline{\Phi}_1(\tau,t); \quad \underline{\Phi}_1(t,t) = \underline{I}$$

We now let $\underline{L}_1(t) = \underline{L}(t)$ and $\underline{L}_2(t) = \underline{B}'(t)\underline{K}(t;T,\underline{F})$, so that $\underline{V}_2(t) = \underline{K}(t;T,\underline{F})$ and therefore

$$\underline{V}_L(t) - \underline{K}(t;T,\underline{F}) = \int_t^T \underline{\Phi}'_L(\tau,t)\left[\underline{L}(\tau)-\underline{K}(\tau)\right]'\left[\underline{L}(\tau)-\underline{K}(\tau)\right]\underline{\Phi}_L(\tau,t)d\tau$$

(3.28)

† Note that all cost matrices have the same boundary condition $\underline{V}(T) = \underline{F}$.

which again shows that $\underline{V}_L(t) \geq \underline{K}(t)$ for all $\underline{L}(t)$.

Lemma 6 suggests a method of generating a series of approximations to the solution $\underline{K}(t; T, \underline{F})$ of the Riccati equation. It is logical to expect that if $\underline{L}_n(t)$ is a reasonable approximation to $\underline{B}'(t)\underline{K}(t; T, \underline{F})$, then the cost matrix $\underline{V}_n(t)$ corresponding to $\underline{L}_n(t)$ will be "close" to $\underline{K}(t; T, \underline{F})$. We then suspect that $\underline{L}_{n+1}(t) = \underline{B}'(t)\underline{V}_n(t)$ will be even a better approximation to $\underline{B}'(t)\underline{K}(t; T, \underline{F})$ than was $\underline{L}_n(t)$, and so on. It is the purpose of the following theorem to make this notion precise. The proof may be found in Appendix D. (See Reference 17 for the one-dimensional case).

Theorem 8: (Method of Successive Approximations)

Let $\underline{V}_{n+1}(t)$, $n = 0, 1, \ldots,$ be the cost matrix associated with $\underline{L}_{n+1}(t)$ where $\underline{L}_{n+1}$ is recursively determined by

$$\underline{L}_{n+1}(t) = \underline{B}'(t)\underline{V}_n(t); \quad n = 0, 1, \ldots \tag{3.29}$$

and where $\underline{L}_o(t)$ is arbitrary with associated cost matrix $\underline{V}_o(t)$. Then

(a) $\underline{K}(t; T, \underline{F}) \leq \underline{V}_{n+1}(t) \leq \underline{V}_n(t)$ for $n = 0, 1, \ldots$

(b) $\lim_{n \to \infty} \underline{V}_n(t) = \underline{V}_\infty(t)$ exists

(c) $\underline{V}_\infty(t) = \underline{K}(t; T, \underline{F})$

Before discussing the implications of Theorem 8 from a computational aspect, we shall examine some of its mathematical ramifications. We first note that for any $\underline{x} \in E_n$, the theorem assures that $\langle \underline{x}, \underline{V}_n(t)\underline{x} \rangle$ converges monotonically to $\langle \underline{x}, \underline{K}(t; T, \underline{F})\underline{x} \rangle$. Since $\underline{V}_n(t)$ and $\underline{K}(t; T, \underline{F})$

are continuous, we conclude by 7.2.2 of Ref. 15, that $<\underline{x}, \underline{V}_n(t)\underline{x}>$ converges uniformly in n to $<\underline{x}, \underline{K}(t)\underline{x}>$ over any interval $[t_o, T]$ in which $\underline{K}(t; T, \underline{F})$ exists. Since this result is valid for all $\underline{x}$, we can assert that $\underline{V}_n(t)$ converges uniformly to $\underline{K}(t; T, \underline{F})$ over any interval $[t_o, T]$, i.e., given an $\epsilon > 0$ there exists an N such that for all $n \geq N$

$$\sup_{t_o \leq t \leq T} \| \underline{V}_n(t) - \underline{K}(t; T, \underline{F}) \| < \epsilon$$

Secondly, the iterative scheme suggested by Theorem 8 is precisely that which is obtained when one applies Newton's method to recursively determine the solution $\underline{K}(t; T, \underline{F})$ of the Riccati equation. (See Reference 18, Chapter 18 for the application of Newton's method in function spaces). In Appendix E we examine this equivalence further by way of a short, non-rigorous exposition. We hasten to add that the successive approximation scheme of Kalaba[17] (also referred to as "quasi-linearization") and Newton's method are not always equivalent, although in this application they are, and hence we have indirectly shown monotone convergence for Newton's method. For further relationships that exist between these two iterative schemes see Reference 17.

The method of successive approximations as discussed above is interesting from a mathematical viewpoint. However, the actual use of such a scheme to compute $\underline{K}(t; T, \underline{F})$ is slightly handicapped from a computational point of view. At the n-th iteration we must compute $\underline{V}_n(t)$ for all $t \epsilon [t_o, T]$. This is accomplished by integrating the linear equation

$$\dot{\underline{V}}_n(t) = -\underline{A}_n'(t)\underline{V}_n(t) - \underline{V}_n(t)\underline{A}_n(t) - \underline{L}_n'(t)\underline{L}_n(t) - \underline{C}'(t)\underline{C}(t) \qquad (3.30)$$

backwards in time starting from the boundary condition $\underline{V}_n(T) = \underline{F}$. For $n = 1, 2, \ldots$, $\underline{L}_n(t)$ is given by $\underline{L}_n(t) = \underline{B}'(t)\underline{V}_{n-1}(t)$ and $\underline{A}_n(t) = \underline{A}(t) - \underline{B}(t)\underline{L}_n(t)$. Therefore, to integrate Eq. (3.30) on a computer we must store the value of $\underline{L}_n(t)$ for all $t \in [t_o, T]$, as well as the matrices $\underline{A}(t)$, $\underline{B}(t)$, $\underline{C}(t)$. Once $\underline{V}_n(t)$ is determined, we set $\underline{L}_{n+1}(t) = \underline{B}'(t)\underline{V}_n(t)$ and compute $\underline{V}_{n+1}(t)$, etc., until $\underline{V}_k(t)$ is sufficiently close to $\underline{K}(t; T, \underline{F})$ for large enough k. However, in order to obtain a reasonable approximation to $\underline{K}(t; T, \underline{F})$ we may require a large number of iterations,† each one entailing an integration of Eq. (3.30).

Therefore, when this scheme for determining the Riccati equation solution is compared with those of Section C we see that the successive approximation scheme requires more off-line computation. Instead of integrating a non-linear equation (the Riccati equation) once, we must integrate a linear equation several times. (Note that no matrix inversions are required in this integration.) It is difficult to say which approach is more reliable, as far as accuracy is concerned, without a comparatory numerical investigation and error analysis. Nor can we definitely state a priori which scheme is more efficient from a computational viewpoint, as to programming simplicity, although the fact that Eq. (3.30) is linear may permit us to use a simpler numerical integration technique (i.e., Euler's method) than was allowed in Section C. In any case, this topic remains a subject for further research.

---

† For a given system this depends entirely on the initial choice of $\underline{L}_o(t)$. If $\underline{L}_o(t)$ is a good approximation to $\underline{B}'(t)\underline{K}(t; T, \underline{F})$ then only a few iterations of Eq. (3.30) will be required to yield a thoroughly adequate approximation to $\underline{K}(t; T, \underline{F})$. This is because of the extremely rapid convergence of Newton's method once the iterations begin to approach the desired solution[18] (quadratic convergence property).

In the previous discussion we showed that the method of successive approximations yielded a monotonic sequence of matrices which converged to $\underline{K}(t; T, \underline{F})$. At each step in the method the solution of a linear time-varying differential equation was required.

In the special case when Σ is time-invariant and $T = \infty$ we know that $\underline{K}(t) = \bar{\underline{K}} =$ constant. For this situation it then seems reasonable to expect that if the method of successive approximations were employed to find $\bar{\underline{K}}$ it should only be necessary to solve a linear, time-invariant algebraic equation at each iteration inasmuch as $\bar{\underline{K}}$ itself satisfies an algebraic equation, namely

$$\underline{0} = \bar{\underline{K}}\underline{A} + \underline{A}'\bar{\underline{K}} + \underline{C}'\underline{C} - \underline{K}\underline{B}\underline{B}'\bar{\underline{K}} \tag{3.31}$$

This is indeed the case and in Appendix D we extend Theorem 8 to cover this situation. We list the result as a corollary to the main theorem.

Corollary 1: Let the time-invariant system Σ be completely observable and controllable and let $\underline{V}_n$, $n = 0, 1, \ldots$ be the (unique) positive definite solution of the linear algebraic equation

$$\underline{0} = \underline{V}_n\underline{A}_n + \underline{A}_n'\underline{V}_n + \underline{C}'\underline{C} + \underline{L}_n'\underline{L}_n \tag{3.32}$$

where, recursively

$$\underline{L}_n = \underline{B}'\underline{V}_{n-1} \quad \text{for } n = 1, 2, \ldots$$

$$\underline{A}_n = \underline{A} - \underline{B}\underline{L}_n$$

and where $\underline{L}_o$ is chosen such that the matrix $\underline{A}_o = \underline{A} - \underline{B}\underline{L}_o$ has eigenvalues with negative real parts. Then,

(a) $\bar{\underline{K}} \leq \underline{V}_{n+1} < \underline{V}_n \leq \ldots$ for $n = 0, 1, \ldots$

(b) $\lim_{n \to \infty} \underline{V}_n = \bar{\underline{K}}$

Corollary 1 provides us with yet another method of determining $\bar{\underline{K}}$ on a digital computer. Linear matrix equations such as Eq.(3.32) (which arise constantly when seeking Lyapunov functions for linear time-invariant systems[20]) may be solved by use of Kroneker products (see Ref. 22). If we do this, $\underline{V}_k$ can be determined at any step by inverting an $\frac{n(n+1)}{2} \times \frac{n(n+1)}{2}$ matrix. However, if the number of state variables, n, is large we will be required to invert a very high order matrix at each step in the iterative scheme. Thus, the order of the system places a severe limitation on the usefulness of this method for determining $\bar{\underline{K}}$, since matrix inversion is a very time consuming process for a digital computer, especially when the order of the matrix to be inverted is large. In such a case it would then seem more efficient to compute $\bar{\underline{K}}$ by Eq. 3.23, using the methods of Section C.

In using the successive approximation scheme of Corollary 1 we must choose $\underline{L}_o$ such that the resulting closed-loop system

$$\dot{\underline{x}}(t) = (\underline{A} - \underline{B}\underline{L}_o)\underline{x}(t) = \underline{A}_o\underline{x}(t)$$

is asymptotically stable. By virtue of Theorem 6 there exists at least one such $\underline{L}_o$, namely $\underline{L}_o = \underline{B}'\bar{\underline{K}}$. However, it is possible to show that if Σ is completely controllable then there exists an $\underline{L}_o$ such that the poles of the closed loop system $\dot{\underline{x}} = \underline{A}_o\underline{x}$, i.e., the eigenvalues of $\underline{A}_o$, can achieve <u>any</u> desired configuration consistent with the dimension of

the system.[24] It is necessary for $\mathrm{Re}\,\lambda_i(\underline{A}_o) < 0$ to insure the boundedness of $\underline{V}_o$. Otherwise the corollary is meaningless.

Finally, we mention that Eq. (3.32) is precisely that which is obtained by applying Newton's method to solve Eq. (3.31). However, Newton's method alone will not provide conditions which will insure monotonic convergence such as we have done.

In summary, we have shown that the implementation of the optimal control in a linear feedback loop must be accomplished by introducing a tape record system. This is necessary due to our inability to compute $\underline{K}(t)$ accurately in an on-line manner because of the instability of the Riccati equation solution, $\underline{K}(t; T, \underline{F})$, in forward time. Realizing these facts we then discussed several iterative schemes which one might use to precompute $\underline{K}(t; T, \underline{F})$ in an off-line manner. Once $\underline{K}(t; T, \underline{F})$, or a reasonable facsimile, is computed, the gain matrix $\underline{L}^*(t) = \underline{B}'(t)\underline{K}(t; T, \underline{F})$ is stored on tape. The tape is then placed into our feedback controller to be played back in real time once the system begins operating. Besides being a tedious task, the storage of a large number of time-varying signals on magnetic tapes can become impractical from a hardware point of view, especially when we require all signals to be played back in time-synchronization. In the following chapter we shall look at this problem in greater detail and suggest a means of designing the linear feedback loop in a "sub-optimal" fashion. In effect we shall accomplish a sub-optimal design by simplifying the hardware requirements at the expense of optimal performance.

CHAPTER IV

"SUBOPTIMAL" DESIGN TECHNIQUES

A. INTRODUCTION

In Section III.D we obtained a sequence of linear control laws $\underline{u}_n(\underline{x}, t) = -\underline{L}_n(t)\underline{x}$ which for $n \to \infty$, $\underline{u}_n(\underline{x}, t)$ approached the optimal control. For a fixed n, therefore, $\underline{u}_n(\underline{x}, t)$ can be regarded as a "suboptimal" control; its use in a feedback loop will necessarily result in a cost which is greater than the optimal cost. However, by taking n sufficiently large, the performance of the "suboptimal" system can be made arbitrarily close to that of the optimal system as shown in Theorem 8.

Yet one major difficulty remains. The implementation of $\underline{u}_n(\underline{x}, t)$ in an actual control system must overcome the same hurdles as those in the path of implementing the optimal control, $\underline{u}^*(\underline{x}, t) = -\underline{B}'(t)\underline{K}(t; T, \underline{F})\underline{x}$. In both cases it is necessary to pre-compute time-varying gain matrices, store them on tape, and play the tape back upon command in an on-line manner.

Briefly, let us reflect upon the inherent problem associated with this method of implementing the optimal control. Suppose for example that we deal with a 10-th order system, i.e., $n = 10$, having three control variables, i.e., $r = 3$, so that the matrix $\underline{L}^*(t) = \underline{B}'(t)\underline{K}(t; T, \underline{F})$ has a total of $10 \times 3 = 30$ time-varying elements. Once determined, each of these components must be stored on a separate tape track requiring 30 separate tape heads.

At time t_o these signals must not only be played back by the tape recorder, but also be played back in time synchronization with one another, and in synchronization with real time. The circuitry required for simultaneous playback and synchronization of 30 signals can therefore become quite unwieldy, forcing one to consider more practical schemes for control implementation.

Our major obstacle to the realizable implementation of the optimal gain matrix $\underline{L}^*(t) = \underline{B}'(t)\underline{K}(t; T, \underline{F})$ is that in general we can say nothing a priori concerning its time-varying structure. (Except in the very special case when the system Σ to be controlled is time invariant and the terminal time $T = \infty$, for which $\underline{B}'(t)\overline{\underline{K}} = \text{constant}$, as described in Section II.F.) Even if Σ is stationary, the optimal feedback gains will be time-varying if T is finite. If Σ is time-varying it is virtually hoepless to expect any qualitative results concerning the time-varying structure of $\underline{B}'(t)\underline{K}(t; T, \underline{F})$. The main theoretical tool for such an investigation is Theorem 2, however its application to an analytical study is severely limited since only in the rarest cases can we specify the time structure of the $2n \times 2n$ matrix $\underline{\Psi}(T, t)$ which appears in Theorem 2 knowing the time structure of the matrices $\underline{A}(t)$, $\underline{B}(t)$ and $\underline{C}(t)$.

Let us digress and suppose for a moment that it is possible to ascertain information as to the nature of the time variation of $\underline{L}^*(t)$. For example, suppose $\underline{L}^*(t)$ for $t \in [t_o, T]$ is known to be of the form

$$\underline{L}^*(t) = a(t)\,\underline{L}^* \tag{4.1}$$

where $a(t)$ is a (known) scalar function of time and $\underline{L}^*$ is a constant r x n matrix. In order to implement the feedback control law $\underline{u}^*(\underline{x}, t) = -a(t)\,\underline{L}^*\underline{x}$ it is only necessary to set r.n feedback gains at fixed values in the control loop and to multiply the r signals $\underline{L}^*\underline{x}(t)$ by $a(t)$. The scalar function $a(t)$ may either be stored on tape or else generated in real time by a digital or analog computer. The implementation of this closed-loop system is shown in Fig. 4.1.

Expanding on this view, let us suppose we know that the optimal gain matrix $\underline{L}^*(t)$ for $t \in [t_o, T]$ has a structure given by

$$\underline{L}^*(t) = \sum_{j=1}^{M} a_j(t)\,\underline{L}_j^* \tag{4.2}$$

where $a_j(t)$ $j=1, \ldots, M$ are scalar time functions and $\underline{L}_j^*$ are constant r x n matrices. The implementation of the optimal feedback law for this case is shown in Fig. 4.2. Once again, the M functions $a_j(t)$ may be stored on tape and played back in synchronization upon command. Notice how this a priori knowledge of the time structure of $\underline{L}^*(t)$ enables us to transform the problem of storing the r.n elements of $\underline{L}^*(t)$ on tape into one of setting gain amplifiers and storing (or generating somehow) only the functions $a_j(t)$, $j = 1, \ldots, M$. (Hopefully $M < r.n$).

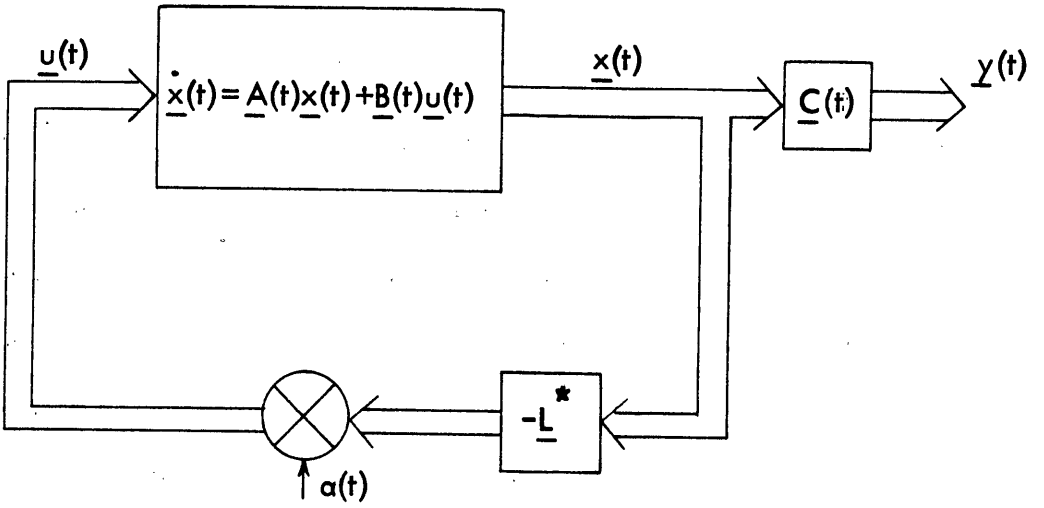


Fig. 4.1 Implementation of $\underline{u}(\underline{x},t) = -\alpha(t)\underline{L}^{*}\underline{x}(t)$

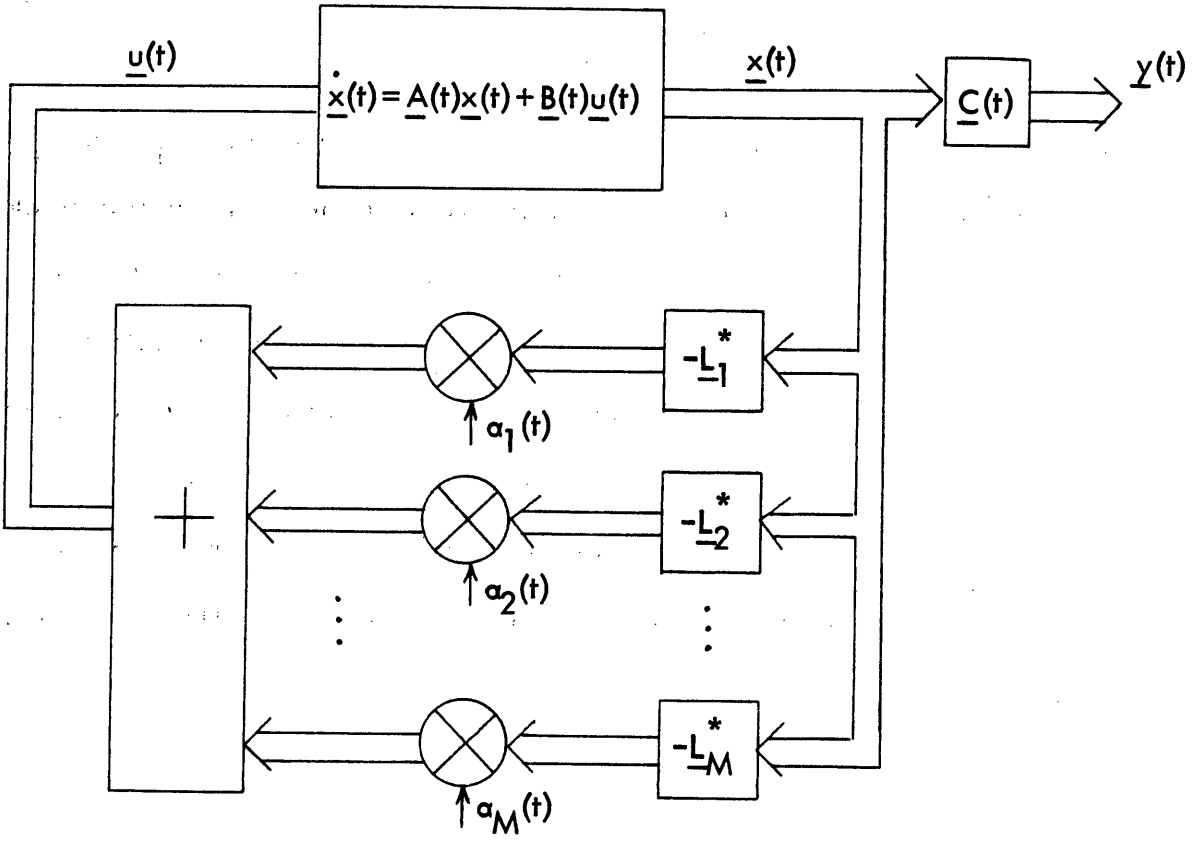


Fig. 4.2 Implementation of $\underline{u}(\underline{x},t) = -\sum_{i=1}^{M} \alpha_i(t)\underline{L}_i^{*}\underline{x}(t)$

Returning now to reality, the prospect of having an optimal gain matrix of the form suggested by Eq. (4.2) is rather slim to say the least. Even if $\underline{L}^*(t)$ were of this form it would not be an easy task to determine this fact a priori. This is indeed unfortunate for it would have presented us with a method of circumventing the engineering difficulties inherent in storing a large number of signals on tape. All is not lost, however, for it seems reasonable to expect that we can sacrifice a small piece of optimal performance for the sake of structural simplicity.

To be more specific, we assume that we have at our disposal a set of scalar (linearly independent) time functions $a_j(t)$ $j = 1, \ldots, M$ for $t \in [t_o, T]$. We then restrict the control input to our system Σ to be of the form

$$\underline{u}(\underline{x}, t) = - \sum_{j=1}^{M} a_j(t) \underline{L}_j \underline{x}(t) \tag{4.3}$$

where $\underline{L}_j$ for $j = 1, \ldots, M$ are arbitrary, constant r x n matrices. We are free to choose these matrices in such a manner as to make the control law (4.3) "close" to the optimal control law $\underline{u}^*(\underline{x}, t) = -\underline{L}^*(t)\underline{x}(t)$ in some reasonable fashion. In such a case (4.3) may be regarded as "suboptimal"--for a given set of a_j's it is unreasonable to expect that there will exist matrices $\underline{L}_j$ such that (4.3) will in fact be the optimal control. On the other hand, as discussed

above, a linear control law of the proposed form (4.3) is relatively easy to implement. By its use, therefore, we are attempting to trade mathematical optimality in return for engineering simplicity and practical usefulness. In this chapter we shall make this notion more precise from a mathematical viewpoint and present a theory for determining the matrices $\underline{L}_j$, by defining a "suboptimal linear regulator problem."

B. STRUCTURE OF THE SUBOPTIMAL CONTROL

In proposing a structure for a suboptimal regulator control two things should be considered. First that it be of a form which readily lends itself to actual implementation, and second that it be of sufficient generality so that the optimal control can be approximated to any degree of accuracy. With these thoughts in mind we will consider in detail a proposed suboptimal control structure.

Let $\underline{u}^*(\underline{x}, t) = -\underline{L}^*(t)\underline{x}$ denote the optimal linear regulator control. The suboptimal design scheme to be investigated will consist of approximating the control matrix $\underline{L}^*(t)$ over the interval of interest $[t_o, T]$. In this manner the suboptimal control will be a linear feedback control law, so that it becomes unnecessary to introduce nonlinear function generators into the feedback system. This is only reasonable since the system Σ is linear and the optimal control itself is a linear feedback law. One technique for approximating $\underline{L}^*(t)$ is briefly described in Section A via Eq. (4.3).

In that scheme $\underline{L}^*(t)$ is approximated over the entire interval $[t_o, T]$ by a gain matrix of the form

$$\underline{L}(t) = \sum_{j=1}^{M} a_j(t)\,\underline{L}_j$$

The implementation of such a gain matrix requires M synchronized tape tracks on which the a_j's are stored and $(n \cdot r) \cdot M$ constant gain amplifiers whose fixed settings correspond to the values of the components of $\underline{L}_j$, $j = 1, \ldots, M$. As M increases so does the complexity of the feedback controller. But on the other hand we certainly expect that as M increases it should be possible to choose the set of matrices $\underline{L}_j$, $j = 1, \ldots, M$ so that $\underline{L}(t)$ becomes a finer and finer approximation to $\underline{L}^*(t)$. In the sequel this question will be analyzed further.

The basic concept behind the above procedure is that it enables us to specify a time-varying structure for $\underline{L}(t)$ which is amenable to an engineering implementation. Once we choose the M matrices $\underline{L}_j$ (given the $a_j(t)$'s) the gain matrix $\underline{L}(t)$ is completely specified over the entire interval $[t_o, T]$. This can also be a liability, however. For example, a particular set of matrices $\underline{L}_j$, $j = 1, \ldots, M$ may result in a gain matrix $\underline{L}(t)$ which is a good approximation to $\underline{L}^*(t)$ over one subinterval $(t_1, t_2) \subset [t_o, T]$ but which is a poor approximation to $\underline{L}^*(t)$ over a different subinterval $(t_3, t_4) \subset [t_o, T]$.

Alternatively, another set of matrices $\underline{L}_j$ may correspond to an $\underline{L}(t)$ which well approximates $\underline{L}^*(t)$ over (t_3, t_4) but yields a poor approximation to $\underline{L}^*(t)$ over (t_1, t_2).

Dilemmas of this sort may be eliminated, if instead of specifying the structure of $\underline{L}(t)$ over the entire interval $[t_o, T]$ we specify its structure over various subintervals. To formulate this notion in more precise terms we assume that we may choose an integer $N \geq 1$ and a set of times $t_1, t_2, \ldots, t_N$ such that

$$t_o < t_1 < t_2 < \ldots < t_{N-1} < t_N = T \tag{4.4}$$

Thus, the intervals

$$I_i = (t_i, t_{i+1}] \quad i = 0, 1, \ldots, N-1$$

are disjoint with

$$\bigcup_{i=0}^{N-1} I_i = (t_o, T]$$

We can now independently specify the structure of $\underline{L}(t)$ over each of the intervals I_i. For any fixed value of i between $i = 0$ and $i = N-1$ let $a_{ij}(t)$ $j=1, \ldots, M$ be a set of M continuous, real-valued scalar time functions which are linearly independent over the interval $I_i = (t_i, t_{i+1}]$. In addition, for a fixed integer i, $0 < i \leq N-1$, let $\underline{L}_{ij}$ $j = 1, \ldots, M$ be a set of $r \times n$ constant matrices which we are free to arbitrarily choose.

We now constrain the gain matrix $\underline{L}(t)$, for $t \in (t_i, t_{i+1}]$ to be of the form

$$\underline{L}(t) = \sum_{j=1}^{M} a_{ij}(t)\underline{L}_{ij} \quad \text{for } t_i < t \leq t_{i+1};\ i = 0, 1, \ldots, N-1 \qquad (4.5)$$

(Notice that if $N = 1$, this case reduces to that described in Section A since $I_o = (t_o, T]$, indicating the generality expressed by Eq.(4.5).)

When $\underline{L}(t)$ is of this form, the choice of the $N \cdot M$ time functions $a_{ij}(t)$, and the $N \cdot M$ matrices $\underline{L}_{ij}$ completely specifies the gain matrix over the entire interval $[t_o, T]$. However, unlike the scheme discussed previously, this method specifies $\underline{L}(t)$ for $t \in [t_o, T]$ by specifying $\underline{L}(t)$ over disjoint subintervals whose union is the entire interval of interest. The problem we are now faced with remains basically the same. Once we are given (or choose) the integers M and N and the functions $a_{ij}(t)$ $i = 0, \ldots, N-1$, $j = 1, \ldots, M$ we wish to determine the $N \cdot M$ matrices $\underline{L}_{ij}$ so that $\underline{L}(t)$ as given by Eq. (4.5) well-approximates $\underline{L}^*(t)$ over $[t_o, T]$. We shall discuss this point further in the next section.

The implementation of a gain matrix of the form (4.5) becomes more and more difficult as N increases. We have already discussed the case $N = 1$. For $N > 1$ the implementation of $\underline{L}(t)$ still requires only M separate, but synchronized, tape tracks; the time function, $a_j(t)$, to be stored on the j-th track is given by

$$a_j(t) = a_{ij}(t) \quad \text{for} \quad t \in (t_i, t_{i+1}]; \; i = 1, \ldots, N \tag{4.6}$$

In addition, we still require $(r \cdot n)M$ simple gain amplifiers in the feedback loop, but unlike the case $N = 1$ these amplifiers must be provided with circuitry to increase or decrease their gains at the times t_i, $i = 0, 1, \ldots, N-1$. For example, consider the $k\ell$-th element of $\underline{L}(t)$. Corresponding to this element we require an amplifier whose gain $g_{k\ell}(t)$ is adjusted according to

$$g_{k\ell}(t) = \underline{L}_{ij}^{(k\ell)} \quad \text{for} \quad t \in (t_i, t_{i+1}] \quad i = 0, \ldots, N-1$$

We shall have more to say on the discrete time adjustment of amplifier gains in a feedback loop when we consider piecewise constant gain matrices in Chapter V.

C. THE SUBOPTIMAL LINEAR REGULATOR PROBLEM

As described in the foregoing sections we wish to constrain the gain matrix $\underline{L}(t)$ to be of the specific form (4.5). The purpose of such a constraint is to circumvent the implementation difficulties associated with storing a large number $(n \cdot r)$ of time-varying quantities on tape. The proposed scheme requires only M tape tracks (where we are free to choose M) and $(n \cdot r)M$ simple gain amplifiers whose gains must be readjusted at the times t_i. These gains correspond to the elements of the matrices $\underline{L}_{ij}$, for $i = 0, \ldots, N-1$; $j = 1, \ldots, M$.

Once M, N and the time functions $a_{ij}(t)$ are chosen, the matrices $\underline{L}_{ij}$ completely specify the gain matrix $\underline{L}(t)$. We shall determine the matrices $\underline{L}_{ij}$ in such a manner that the associated control law

$$\underline{u}(\underline{x}, t) = -\underline{L}(t)\,\underline{x}(t)$$

minimizes $J(\underline{x}_o, t_o, T, \underline{u}(\cdot))$ subject to the structural constraints (4.5) placed on $\underline{L}(t)$. For convenience we define

Definition 6: Let M and N be fixed positive integers, and let t_i, $i = 0, 1, \ldots, N$ be a given set of times such that

$$t_o < t_1 < \ldots < t_{N-1} < t_N = T$$

For every i, $i = 0, 1, \ldots, N-1$, let $a_{ij}(t)$, $j = 1, 2, \ldots M$ be a given, linearly independent, set of M scalar time functions which are defined and uniformly bounded over the time interval $(t_i, t_{i+1}]$.

We then say that the function $\underline{L}(\cdot)$ is of class Λ_{NM} if

$$\underline{L}(t) = \sum_{j=1}^{M} a_{ij}(t)\,\underline{L}_{ij} \quad \text{for } t \in (t_i, t_{i+1}]\ i = 0, \ldots, N-1 \tag{4.7}$$

where $\underline{L}_{ij}$; $i = 0, \ldots, N-1$, $j = 1, \ldots, M$ are arbitrary $r \times n$ constant matrices.

More succintly we have

$$\Lambda_{NM} = \{\underline{L}(\cdot): \underline{L}(t) = \sum_{j=1}^{M} a_{ij}(t)\underline{L}_{ij} \text{ for } t\epsilon(t_i, t_{i+1}]; \; i = 0, 1, \ldots, N-1\}$$

Note that Λ_{NM} as defined is a linear function space; if $\underline{L}_1(\cdot)$ and $\underline{L}_2(\cdot) \; \epsilon\Lambda_{NM}$, then $a\underline{L}_1(\cdot) + b\underline{L}_2(\cdot) \; \epsilon\Lambda_{NM}$ for all scalars a, b. We shall make Λ_{NM} a <u>normed</u> linear space by introducing a suitable norm. If $\underline{L}(\cdot)\epsilon\Lambda_{NM}$ we define

$$\|\underline{L}(\cdot)\|_2 = [\int_{t_o}^{T} \|\underline{L}(t)\|^2 \, dt]^{1/2} \tag{4.8a}$$

where the norm on the matrix $\underline{L}(t)$ is the induced matrix norm defined earlier in Chapter II, viz.,

$$\|\underline{L}(t)\|^2 = \sup_{\|\underline{x}\| = 1} \|\underline{L}(t)\underline{x}\|^2 = \lambda_{max}[\underline{L}'(t)\underline{L}(t)] \tag{4.8b}$$

We shall make use of Eqs. (4.8a) and (4.8b) in Section D.

In the definition of Λ_{NM}, the scalar time functions $a_{ij}(t)$ are assumed given. Thus $\underline{L}(t)$ is determined solely by the choice of the matrices $\underline{L}_{ij}$. We shall now discuss the manner in which these matrices are to be determined. Let $\underline{L}(\cdot)\epsilon\Lambda_{NM}$ and let $\underline{V}_L(t)$ be the cost matrix associated with $\underline{L}(t)$ as defined in Appendix C. $\underline{V}_L(t)$ is therefore given by

$$\underline{V}_L(t) = \underline{\Phi}_L'(T,t)\underline{F}\underline{\Phi}_L(T,t) + \int_t^T \underline{\Phi}_L'(\tau,t)[\underline{C}'(\tau)\underline{C}(\tau) + \underline{L}'(\tau)\underline{L}(\tau)]\underline{\Phi}_L(\tau,t)d\tau \tag{4.9}$$

If, now, $\underline{x}_o$ is the initial state of our system Σ at time t_o, the cost associated with the control $\underline{u}(\underline{x},t) = -\underline{L}(t)\,\underline{x}(t)$ is given by

$$J(\underline{x}_o, t_o, T, \underline{u}(\cdot))\Big|_{\underline{u} = -\underline{L}(t)\underline{x}(t)} = \frac{1}{2}<\underline{x}_o, \underline{V}_L(t_o)\underline{x}_o> \tag{4.10}$$

It would then seem reasonable, inasmuch as our original control objective was to minimize the cost $J(\underline{x}_o, t_o, T, \underline{u}(\cdot))$, that we should attempt to choose the matrices $\underline{L}_{ij}$ in such a manner that the resulting gain matrix $\underline{L}(t)$ minimizes (4.10). Consequently our problem is to choose $\underline{L}(\cdot)\epsilon\Lambda_{NM}$ (or equivalently the constant matrices $\underline{L}_{ij}$) such that 4.10 is minimized.

On the surface this proposition for choosing $\underline{L}_{ij}$ is quite reasonable. There is, however, one difficulty--the optimal choice of $\underline{L}_{ij}$ will in general depend upon the initial state $\underline{x}_o$. In an actual control system the initial state is not known a priori and must be measured in real time. If we demand that $\underline{L}(\cdot)$ minimize (4.10) the resulting dependence of $\underline{L}_{ij}$ upon $\underline{x}_o$ precipitates on-line computation. This defeats our entire purpose of trying to simplify the implementation requirements of our feedback controller. We wish to have all computations done off-line.

Our only alternative in such a case is to choose $\underline{L}(\cdot)$ to minimize a functional independent of $\underline{x}_o$. One such functional is

$$\nu(\underline{L}) = \lambda_{max}[\underline{V}_L(t_o)] = \sup_{\|\underline{x}_o\|=1} <\underline{x}_o, \underline{V}_L(t_o)\underline{x}_o> \tag{4.11}$$

which is the maximum value that (4.10) can attain when $\underline{x}_o$ ranges over the unit hyper-sphere. $\nu(\underline{L})$ is always nonnegative since $\underline{V}_L(t_o)$ is positive semidefinite, yet this particular choice of ν is mathematically unpleasant since $\lambda_{max}(\cdot)$ is a nonlinear functional of its argument.

The functional $\nu(\underline{L})$ is the maximum eigenvalue of $\underline{V}_L(t_o)$. But since all eigenvalues of $\underline{V}_L(t_o)$ are positive, a useful, and mathematically tractible substitute for $\nu(\underline{L})$ is simply the trace of the matrix $\underline{V}_L(t_o)$, i.e., the sum of all the eigenvalues of $\underline{V}_L(t_o)$. Hence, we write

$$\mu(\underline{L}) = \text{tr } \underline{V}_L(t_o) \tag{4.12}$$

The trace of a matrix is a linear functional of its argument, which is an extremely useful property as we shall see in later sections. Consequently, in the sequel, we shall seek the control matrix $L(\cdot)\epsilon\Lambda_{NM}$ which minimizes (4.12). We denote the optimal choice of $\underline{L}(\cdot)$ by $\underline{L}^o(\cdot)$ i.e., $\underline{L}^o(\cdot)$ is the argument of $\mu(\underline{L})$ for which $\mu(\underline{L})$ attains its minimum value. Mathematically this is expressed as

$$\underline{L}^{o}(\cdot) = \arg \min_{\underline{L}(\cdot)\epsilon\Lambda_{NM}} [\mathrm{tr}\underline{V}_{L}(t_o)] \tag{4.13}$$

Besides being reasonable from a mathematical standpoint, the choice of the functional $\mu(\underline{L})$ also has a physical interpretation as follows. Suppose that the initial state $\underline{x}_o$ is a random variable which is uniformly distributed over the surface of the unit sphere in E_n. Under these conditions it is most reasonable to seek the control matrix $\underline{L}(\cdot)\epsilon\Lambda_{NM}$ which minimizes the expected value (over $\underline{x}_o$) of the cost $J(\underline{x}_o, t_o, T, \underline{u}(\cdot))$, i.e.,

$$\underline{E}_{\underline{x}_o}\{J(\underline{x}_o, t_o, T, \underline{u}(\cdot)) \Big|_{\underline{u} = -\underline{L}(t)\underline{x}}\} = E_{\underline{x}_o}\{\tfrac{1}{2} <\underline{x}_o, \underline{V}_L(t_o)\underline{x}_o> \} \tag{4.14}$$

but $$E_{\underline{x}_o}\{\tfrac{1}{2} <\underline{x}_o, \underline{V}_L(t_o)\underline{x}_o>\} = E_{\underline{x}_o}\{\tfrac{1}{2}\,\mathrm{tr}\underline{x}_o'\underline{V}_L(t_o)\underline{x}_o\}$$

$$= E_{\underline{x}_o}\{\tfrac{1}{2}\,\mathrm{tr}\,\underline{V}_L(t_o)\underline{x}_o\underline{x}_o'\}$$

$$= \tfrac{1}{2}\,\mathrm{tr}\cdot\{\underline{V}_L(t_o)\cdot E(\underline{x}_o\underline{x}_o')\}$$

where the last step follows from the fact that the trace and expectation operations commute. But now since $\underline{x}_o$ is uniformly distributed over the surface of the unit sphere,

$$E(\underline{x}_o\underline{x}_o') = \underline{I}\,, \text{ (the identity matrix)}$$

so that $$E\{\tfrac{1}{2} < \underline{x}_o, \underline{V}_L(t_o)\underline{x}_o > \} = \tfrac{1}{2} \operatorname{tr} \underline{V}_L(t_o) = \tfrac{1}{2} \mu(\underline{L}) \qquad (4.15)$$

Therefore, the gain matrix which minimizes (4.14) is that which minimizes (4.12), and is characterized by Eq. (4.13). In Section E we shall give necessary conditions for $\underline{L}^o(\cdot)$ to minimize $\mu(\cdot)$.

There is yet another interpretation of μ. The functional $\nu(\underline{L})$ is the <u>maximum</u> value attainable by $< \underline{x}_o, \underline{V}_L(t_o)\underline{x}_o >$ as $\underline{x}_o$ ranges over the surface of the unit sphere, ∂S, in E_n. It is then possibly to show that μ/n is the <u>average</u> value of the cost $< \underline{x}_o, \underline{V}_L(t_o)\underline{x}_o >$ as $\underline{x}_o$ varies over ∂S. To prove this statement let ξ denote this average, then by definition, ξ is given mathematically by

$$\xi = \int_{\partial S} < \underline{x}, \underline{V}_L(t_o)\underline{x} > d\underline{s} \Big/ \int_{\partial S} d\underline{s} \qquad (4.16)$$

where $d\underline{s}$ denotes an element of surface area in E_n. But on the unit sphere, $\|\underline{x}\|^2 = < \underline{x}, \underline{x} > = 1$ so that we can write (4.16) in the form

$$\xi = \int_{\partial S} < \underline{x}, \underline{V}_L(t_o)\underline{x} > d\underline{s} \Big/ \int_{\partial S} < \underline{x}, \underline{x} > d\underline{s} \qquad (4.17)$$

But $\underline{x}$ is the unit normal to the surface ∂S. We therefore make use of the divergence theorem[29] to replace both surface integrals over ∂S in (4.17) by volume integrals over the unit sphere S. Since div $\underline{A}\underline{x}$ = trace $\underline{A}$ for any square matrix $\underline{A}$, we obtain

$$\xi = \frac{[\,t_r \underline{V}_L(t_o)] \cdot \int_S d\underline{v}}{(\text{tr } \underline{I}) \cdot \int_S d\underline{v}} = \frac{1}{n} \text{ tr } \underline{V}_L(t_o) \tag{4.18}$$

which shows that the average value of $< \underline{x}_o, \underline{V}_L(t_o)\underline{x}_o >$ over the unit hypersphere $\|\underline{x}_o\| = 1$ is the average of the eigenvalues of $\underline{V}_L(t_o)$.

We can readily obtain a lower bound for $\mu(\underline{L})$. In Lemma 6 we showed that for any gain matrix $\underline{L}(t)$, the associated cost matrix $\underline{V}_L(t)$ satisfied

$$\underline{K}(t; T, \underline{F}) \leq \underline{V}_L(t) \quad \text{for all } t \in [t_o, T] \tag{4.19}$$

In particular, for $t = t_o$ and $\underline{L}(\cdot) \in \Lambda_{NM}$ we obtain

$$\underline{K}(t_o; T, \underline{F}) \leq \underline{V}_L(t_o) \quad \text{for all } \underline{L}(\cdot) \in \Lambda_{NM} \tag{4.20}$$

Then, since $\underline{K}(t_o)$ and $\underline{V}_L(t_o)$ are both positive semidefinite, taking the trace of both sides of Eq. (4.20) yields

$$\text{tr } \underline{K}(t_o) \leq \text{tr } \underline{V}_L(t_o) = \mu(\underline{L}) \quad \text{for all } \underline{L}(\cdot) \in \Lambda_{NM} \tag{4.21}$$

We may carry Eq. (4.21) one step further by asserting that equality can only hold <u>if and only if</u> $\underline{L}(t) = \underline{L}^*(t) = \underline{B}'(t)\,\underline{K}(t; T, \underline{F})$ for all $t \in [t_o, T]$. This fact is immediately obtainable from Eq. (3.28) since

$$\text{tr } \underline{V}_L(t_o) - \text{tr } \underline{K}(t_o) = \text{tr } [\underline{V}_L(t_o) - \underline{K}(t_o)]$$

$$= \text{tr} \int_{t_o}^{T} \underline{\Phi}'_L(t, t_o)[\underline{L}(t) - \underline{L}^*(t)]'[\underline{L}(t) - \underline{L}^*(t)] \cdot \underline{\Phi}_L(t, t_o)dt$$

$$> 0 \quad \text{if} \quad \underline{L}(t) \neq \underline{L}^*(t)$$

In summary, the suboptimization problem which we shall analyze in the sequel is

The suboptimal linear regulator problem

Given the set of times $\{t_i, i=0, 1, \ldots, N\}$ and the set of functions $\{a_{ij}(t), i=0, 1, \ldots, N-1; j=1, 2, \ldots, M\}$. Determine the gain matrix $\underline{L}^o(\cdot) \epsilon \Lambda_{NM}$, where

$$\Lambda_{NM} = \{\underline{L}(\cdot) : \underline{L}(t) = \sum_{j=1}^{M} a_{ij}(t) \underline{L}_{ij} \quad \text{for} \quad t \epsilon (t_i, t_{i+1}], \; i=0, 1, \ldots, N-1\}$$

which minimizes

$$\mu(\underline{L}) = \text{tr } \underline{V}_L(t_o)$$

We shall call $\underline{L}^o(\cdot)$ the suboptimal gain matrix and $\mu^o = \mu(\underline{L}^o)$ the suboptimal cost.

D. CONVERGENCE OF THE SUBOPTIMAL SOLUTION AS $M \to \infty$

In the foregoing development we constrained the time structure of gain matrix $\underline{L}(t)$ to be of the form (4.7), or in other words we

required $\underline{L}(\cdot)\epsilon\Lambda_{NM}$. The argument in favor of such a constraint is based on the fact that it is easier to implement a gain matrix of this form (by using M tape channels and associated piecewise constant feedback gains) than it is to implement the gain matrix $\underline{L}^*(t) = \underline{B}'(t)\,\underline{K}(t;T,\underline{F})$. As M is allowed to increase, N remaining constant, the circuitry demands of the linear feedback loop also increase and it becomes more and more difficult to implement $\underline{L}(t)$. On the other hand, as M increases, the set Λ_{NM} encompasses a more inclusive class of gain matrices, i.e.,

$$\Lambda_{NM_1} \subset \Lambda_{NM_2} \quad \text{for} \quad M_1 < M_2 \tag{4.22}$$

since any element $\underline{L}(\cdot)\epsilon\Lambda_{NM_1}$ is automatically an element of Λ_{NM_2} for $M_2 > M_1$. In other words, for a prespecified set of scalar time functions $\{a_{ij}(t)\}$, the sets Λ_{N1}, Λ_{N2}, ..., form a nested sequence:

$$\Lambda_{N1} \subset \Lambda_{N2} \subset \cdots$$

Consequently, if $\underline{L}_M^o(\cdot)$ denotes the element of Λ_{NM} which minimizes $\mu(\underline{L})$, we suspect that as M increases, $\underline{L}_M^o(t)$ will become a finer and finer approximation to $\underline{L}^*(t) = \underline{B}'(t)\underline{K}(t;T,\underline{F})$. In addition, if μ_M^o denotes the minimum value of $\mu(\underline{L})$ over the class Λ_{NM}, we then also expect $\mu_M^o \to \text{tr}\ \underline{K}(t_o)$ as $M \to \infty$. Prior to actually determining necessary conditions on $\underline{L}_M^o(\cdot)$ for the minimization of $\mu(\underline{L})$ we shall investigate the above notions in a more precise mathematical framework.

For convenience in what follows, we assume that the integer N is specified and that the times t_i are equally spaced on the interval $[t_o, T]$, i.e.,

$$t_{i+1} = t_i + \Delta \qquad i = 0, 1, \ldots, N-1$$

where $\Delta = (T-t_o)/N$.

We then let $\phi_1(t), \phi_2(t), \ldots$ be an infinite sequence of real-valued scalar time functions which are defined and square integrable on the interval $[0, \Delta]$,† and which are <u>complete</u> in this class $\mathcal{L}^2[0, \Delta]$ where we define

<u>Definition 7:</u> The sequence $\{\phi_j(t)\}$ is said to be complete (in $\mathcal{L}^2[0, \Delta]$) if given any $\beta(t) \in \mathcal{L}^2[0, \Delta]$ and any $\epsilon > 0$ there is a linear combination $\beta_k(t)$ of the form

$$\beta_k(t) = \sum_{j=1}^{k} a_j \phi_j(t) \tag{4.23}$$

where $a_1, \ldots, a_k$ are real numbers such that

$$\| \beta_k(\cdot) - \beta(\cdot) \|_2 \triangleq \left[\int_0^{\Delta} |\beta_k(t) - \beta(t)|^2 dt \right]^{1/2} \leq \epsilon \tag{4.24}$$

where k depends on ϵ

---

† This class of functions is denoted by $\mathcal{L}^2[0, \Delta]$.

Examples of such sequences which are complete in $\mathcal{L}^2[0,\Delta]$ are

(i) $\phi_j(t) = t^{(j-1)}$, $j = 1, 2, \ldots$

(ii) $\phi_j(t) = e^{(j-1)t}$ $j = 1, 2, \ldots$

We now turn to our suboptimization problem. For any fixed M the class Λ_{NM} is described by

$$\Lambda_{NM} = \{\underline{L}(\cdot): \underline{L}(t) = \sum_{j=1}^{M} a_{ij}(t)\, \underline{L}_{ij} \text{ for } t\epsilon[t_i, t_i + \Delta]\ i = 0, 1, \ldots N-1\}$$

We wish to choose the functions $a_{ij}(t)$ akin to the functions $\phi_j(t)$ introduced above. We therefore choose

$$a_{ij}(t) = \phi_j(t-i\Delta) \text{ for } t\epsilon(t_i, t_i+\Delta]\ \ i=0, 1, \ldots, N-1$$

$$\text{for } j = 1, 2, \ldots, M \tag{4.25}$$

Thus $\underline{L}(\cdot)\epsilon\Lambda_{NM}$ is of the form

$$\underline{L}(t) = \sum_{j=1}^{M} \phi_j(t-i\Delta)\underline{L}_{ij} \text{ for } t\epsilon(t_i, t_i+\Delta]\ i = 0, 1, \ldots N-1 \tag{4.26}$$

We let $\underline{L}_M^o(t)$ be the gain matrix of the form (4.26) which minimizes

$$\mu(\underline{L}) = \mathrm{tr}\underline{V}_{\underline{L}}(t_o)$$

so that $\underline{L}_M^o(\cdot)$ is the suboptimal gain matrix; and we let

$$\mu_M^o = \mu(\underline{L})\Big|_{\underline{L}(\cdot) = \underline{L}_M^o(\cdot)} \tag{4.27}$$

denote the associated suboptimal cost, indicating the dependence of these quantities upon M. Clearly μ_M^o cannot increase with M, i.e.,

$$\mu_1^o \geq \mu_2^o \geq \cdots \tag{4.28}$$

since any linear combination of $\phi_1, \ldots, \phi_M$ of the form (4.26) is automatically a linear combination of $\phi_1, \ldots, \phi_M, \phi_{M+1}$. Correspondingly,

$$\Lambda_{N1} \subset \Lambda_{N2} \subset \cdots \tag{4.29}$$

We then show in Appendix G that

Theorem 9:

(i) $$\lim_{M\to\infty} \mu_M^o = \text{tr } \underline{K}(t_o) = \mu(\underline{L}^*)$$

and (ii) $$\lim_{M\to\infty} \|\underline{L}_M^o(\cdot) - \underline{L}^*(\cdot)\|_2 = \lim_{M\to\infty} \left[\int_{t_o}^{T} \|\underline{L}_M^o(t) - \underline{L}^*(t)\|^2 dt\right]^{1/2} = 0$$

where $$\underline{L}^*(t) = \underline{B}'(t)\,\underline{K}(t; T, \underline{F})$$

Hence, the intuitively expected results which we discussed earlier are indeed true. For a specific set of complete functions $\{\phi_j\}$, the minimum value of $\mu(\underline{L})$ over the class Λ_{NM} (denoted by μ_M^o) converges to $\text{tr}\underline{K}(t_o)$ as $M \to \infty$, where we recall from Eq. (4.21) that

tr $\underline{K}(t_o)$ is absolutely the smallest value that $\mu(\underline{L})$ can ever attain. In addition we have shown that the sequence $\underline{L}_1^o(\cdot)$, $\underline{L}_2^o(\cdot), \ldots,$ where $\underline{L}_M^o(\cdot)$ is the suboptimal gain matrix belonging to the set Λ_{NM}, converges to a limit in a mean-square sense and that the limit is $\underline{L}^*(\cdot)$. This fact is often written as $\underline{L}_M^o(\cdot) \Rightarrow \underline{L}^*(\cdot)$. Therefore, by allowing greater complexity in the form of the gain matrix $\underline{L}(\cdot)$ we can achieve a suboptimal system performance approaching optimality. The price we must pay is reflected in the increased hardware requirements necessary to implement a gain matrix of the form (4.26) as M increases.

The above theorem gives results which are notably similar to those which appear in the study of the Ritz method in the calculus of variations. (See Ref. 19, Chapter 8). Borrowing the terminology indigenous to this method, we have shown that the sequence $\underline{L}_1^o(\cdot)$, $\underline{L}_2^o(\cdot), \ldots$ is a minimizing sequence for the functional $\mu(\underline{L})$ since $\mu_M^o \rightarrow \mu(\underline{L}^*)$. In addition, we have shown that the minimizing sequence has a <u>limit</u>. Generally, this is an extremely difficult task in most applications of the Ritz method, and depends on the detailed structural form of the functional to be minimized (in our case $\mu(\underline{L})$). At the present time the Ritz method is quite familiar to physicists and is a standard direct method in the calculus of variations. However, it does not seem to have been applied to the solution of optimal control problems; Theorem 9 suggests that it may be fruitful to do so.

The speed of convergence of μ_M^o to $\mu(L^*)$ and of $\underline{L}_M^o(\cdot)$ to $\underline{L}^*(\cdot)$ obviously depends both upon the original optimization problem itself and on the choice of the functions $\phi_j(t)$. In any particular situation we would like to choose the set $\{\phi_j\}$ so that linear combinations of the form (4.26) involving only a very small number of functions ϕ_j will result in quite satisfactory approximations to $\underline{L}^*(\cdot)$ and $\mu(\underline{L}^*)$. This is indeed a difficult problem and remains a subject for further research.

E. NECESSARY CONDITIONS FOR THE SUBOPTIMALITY OF $\underline{L}^o(\cdot)$

In this section we shall obtain necessary conditions for $\underline{L}^o(\cdot)\epsilon\Lambda_{NM}$ to minimize $\mu(\underline{L}) = \mathrm{tr}\underline{V}_L(t_o)$. We assume that N and M are fixed and that the functions $a_{ij}(t)$ i = 0, . . . , N-1, j=1, . . . , M and the times t_i, i = 0, 1, . . . , N are given. Under these circumstances, the specification of $\underline{L}(\cdot)\epsilon\Lambda_{NM}$ is equivalent to the specification of the N.M constant matrices $\underline{L}_{ij}$ appearing in the expression for $\underline{L}(t)$. Therefore, the task of determining $\underline{L}^o(\cdot)$ reduces to determining the matrices $\underline{L}_{ij}$ such that $\mu(\underline{L})$ is minimized. Hence we may regard $\mu(\underline{L})$ as a function of $\underline{L}_{ij}$. We denote the set of matrices $\underline{L}_{ij}$ at which $\mu(\underline{L})$ attains its minimum by $\{\underline{L}_{ij}^o, i = 0, 1, \ldots, N-1, j = 1, \ldots, M\}$. Finally, we define $\underline{V}^o(t)$ as the cost matrix associated with $\underline{L}^o(t)$, i.e.,

$$\underline{V}^o(t) \triangleq \underline{V}_L(t) \Big|_{\underline{L}(t) = \underline{L}^o(t)} \tag{4.30}$$

where

$$\underline{L}^{o}(t) = \sum_{j=1}^{M} a_{ij}(t)\,\underline{L}_{ij}^{o} \quad \text{for} \quad t\epsilon(t_i, t_{i+1}] \quad i = 0, 1, \ldots, N-1$$

We now determine conditions which $\underline{L}_{ij}^{o}$ must satisfy in order to minimize $\mu(\underline{L})$.

For fixed values of i and j, the functional $\mathrm{tr}\,\underline{V}_L(t_o)$ is continuously differentiable with respect to the elements of the matrix $\underline{L}_{ij}$. Therefore, $\mathrm{tr}\,\underline{V}_L(t_o)$ is a trace function of the r x n matrix $\underline{L}_{ij}$ as defined in Appendix F. In addition, since $\mathrm{tr}\,\underline{V}_L(t_o)$ is continuously differentiable in the elements of $\underline{L}_{ij}$ we know, appealing to the concepts of basic calculus, that in order for the set of matrices $\{\underline{L}_{ij}^{o}\}$ to minimize $\mathrm{tr}\,\underline{V}_L(t_o)$ it is necessary that

$$\frac{\partial\,\mathrm{tr}\,\underline{V}_L(t_o)}{\partial(\underline{L}_{ij}^{o})^{k\ell}} = 0 \quad \text{for } i = 0, 1, \ldots, N-1;\ j=1, \ldots, M \qquad k = 1, \ldots, r;\ \ell = 1, \ldots, n \tag{4.31}$$

where $(\underline{L}_{ij})^{k\ell}$ denotes the $k\ell$-th element of the matrix $\underline{L}_{ij}$. Equation (4.31) simply expresses the requirement that the partial derivatives of a functional evaluated at its minimum must be zero. Introducing the concept of a "gradient matrix" as defined in Appendix F. Equation (4.31) may be written in a more compact form as

$$\frac{\partial\,\mathrm{tr}\,\underline{V}_L(t_o)}{\partial\,\underline{L}_{ij}^{o}} = \underline{0} \quad \text{for} \quad i = 0, 1, \ldots, N-1;\ j = 1, \ldots, M \tag{4.31a}$$

We now wish to evaluate the above gradient matrix in order to obtain necessary conditions on the matrices $\underline{L}^o_{ij}$. We employ the technique outlined in Appendix F to calculate the gradient matrix of the trace function $\underline{V}_L(t_o)$, and show

Theorem 10: (Necessary conditions for suboptimality)

If $\underline{L}^o(\cdot)\epsilon\Lambda_{NM}$ minimizes $tr\underline{V}_L(t_o)$ then for all $i = 0, 1, \ldots, N-1$

$$\int_{t_i}^{t_{i+1}} a_{ij}(t)[\underline{L}^o(t)-\underline{B}'(t)\underline{V}_o(t)]\underline{\Phi}_o(t,t_o)\underline{\Phi}'_o(t,t_o)dt = \underline{0}; j=1,\ldots,M \tag{4.32}$$

where $\underline{\Phi}_o(t,t_o)$ denotes the transition matrix corresponding to $\underline{L}^o(t)$, i.e.,

$$\frac{d}{dt}\underline{\Phi}_o(t,t_o) = [\underline{A}(t)-\underline{B}(t)\underline{L}^o(t)]\underline{\Phi}_o(t,t_o);\ \underline{\Phi}_o(t_o,t_o) = \underline{I}$$

Proof: To compute $\dfrac{\partial\, tr\, \underline{V}_L(t_o)}{\partial \underline{L}^o_{ij}}$ we define, for fixed i and j

$$\underline{L}^\epsilon(t) = \begin{cases} \underline{L}^o(t) & \text{for } t_o \le t \le t_i;\ t_{i+1} < t \le T \\ \underline{L}^o(t) + \epsilon a_{ij}(t)\Delta\underline{L}_{ij} & \text{for } t_i < t \le t_{i+1} \end{cases} \tag{4.33}$$

where $\epsilon\Delta\underline{L}_{ij}$ denotes a small deviation from $\underline{L}^o_{ij}$. Then if $\underline{V}_\epsilon(t)$ is the cost matrix associated with $\underline{L}^\epsilon(t)$ we have, by linearity of the

of the trace functional

$$\text{tr } \underline{V}_\epsilon(t_o) - \text{tr } \underline{V}_o(t_o) = \text{tr}[\underline{V}_\epsilon(t_o) - \underline{V}_o(t_o)] \tag{4.34}$$

We now use Eq. (C.17) of Appendix C to write

$$\underline{V}_\epsilon(t_o)-\underline{V}_o(t_o) = \int_{t_o}^{T} \underline{\Phi}'_\epsilon(t, t_o)[(\underline{L}^\epsilon-\underline{L}^o)'(\underline{L}^\epsilon-\underline{L}^o) - (\underline{L}^\epsilon-\underline{L}^o)'(\underline{B}'\underline{V}_o-\underline{L}^o)$$

$$-(\underline{B}'\underline{V}_o-\underline{L}^o)'(\underline{L}^\epsilon-\underline{L}^o)]\underline{\Phi}_\epsilon(t, t_o)dt \tag{4.35}$$

where $\underline{\Phi}_\epsilon(t, t_o)$ satisfies

$$\frac{d}{dt}\underline{\Phi}_\epsilon(t, t_o) = [\underline{A}(t) - \underline{B}(t)\underline{L}^\epsilon(t)]\underline{\Phi}_\epsilon(t, t_o);\ \underline{\Phi}_\epsilon(t_o, t_o) = \underline{I}$$

Substituting Eq. (4.35) into Eq. (4.34) and using the definition of $\underline{L}^\epsilon(t)$ yields

$$\text{tr}[\underline{V}_\epsilon(t_o)-\underline{V}_o(t_o)] =$$

$$\text{tr}\int_{t_i}^{t_{i+1}} \underline{\Phi}'_\epsilon(t, t_o)[\epsilon^2 a_{ij}^2(t)(\Delta\underline{L}_{ij})'(\Delta\underline{L}_{ij})-2\epsilon a_{ij}(t)(\underline{B}'\underline{V}_o-\underline{L}^o)'(\Delta\underline{L}_{ij})]\underline{\Phi}_\epsilon(t, t_o)dt$$

But to first order in ϵ for $t\epsilon(t_i, t_{i+1}]$ we have

$$\underline{\Phi}_\epsilon(t, t_o) = \underline{\Phi}_o(t, t_o) + \epsilon\underline{\Phi}_o(t, t_o)\int_{t_i}^{t} a_{ij}(\tau)\underline{\Phi}_o(t_o, \tau)\underline{B}(\tau)(\Delta\underline{L}_{ij})\underline{\Phi}_o(\tau, t_o)d\tau$$

Hence, to <u>first order</u> in ϵ, $tr[\underline{V}_\epsilon(t_o) - \underline{V}_o(t_o)]$ becomes

$$tr[\underline{V}_\epsilon(t_o)-\underline{V}_o(t_o)] = -2\epsilon tr\int_{t_i}^{t_{i+1}} a_{ij}(t)\underline{\Phi}_o'(t,t_o)[\underline{B}'(t)\underline{V}_o(t)-\underline{L}^o(t)]'(\Delta\underline{L}_{ij})\underline{\Phi}_o(t,t_o)dt$$

$$=-2\epsilon\, tr\int_{t_i}^{t_{i+1}} a_{ij}(t)\underline{\Phi}_o(t,t_o)\underline{\Phi}_o'(t,t_o)[\underline{B}'(t)\underline{V}_o(t)-\underline{L}^o(t)]'(\Delta\underline{L}_{ij})dt$$

Finally, as shown in Appendix F, this implies

$$\left.\frac{\partial tr\underline{V}_L(t_o)}{\partial \underline{L}_{ij}}\right|_{\underline{L}_{ij}=\underline{L}^o_{ij}} = -2\int_{t_i}^{t_{i+1}} a_{ij}(t)[\underline{B}'(t)\underline{V}_o(t)-\underline{L}^o(t)]\underline{\Phi}_o(t,t_o)\underline{\Phi}_o'(t,t_o)dt$$

Since this matrix must equal zero if $\underline{L}^o(\cdot)$ is to minimize $tr\,\underline{V}_L(t_o)$ we obtain the desired result. For every fixed integer i, the above gradient matrix must equal zero for all $j = 1, \ldots, M$. ‖

In the very special (and unusual) situation when the initial state $\underline{x}_o$ of the system Σ is known a priori it is then possible to choose $\underline{L}(\cdot)\epsilon\Lambda_{NM}$ which minimizes

$$2J(\underline{x}_o, t_o, T, \underline{u}(\cdot))\Big|_{\underline{u}=-\underline{L}\underline{x}} = <\underline{x}_o, \underline{V}_L(t_o)\underline{x}_o> = tr[\underline{V}_L(t_o)\underline{x}_o\underline{x}_o'] \qquad (4.36)$$

If we let $\underline{L}^\ell(\cdot)$ be that element of Λ_{NM} which minimizes Eq.(4.36) and if $\underline{V}_\ell(t)$ denotes the cost matrix associated with $\underline{L}^\ell(t)$ we can

obtain, via the same technique used in the proof of Theorem 10, that

Corollary 1: If $\underline{L}^{\ell}(\cdot)\epsilon\Lambda_{NM}$ minimizes $\langle x_o, \underline{V}_L(t_o)\underline{x}_o\rangle$ then for all $i = 0, 1, \ldots, N-1$; $j=1, \ldots, M$

$$\int_{t_i}^{t_{i+1}} a_{ij}(t)[\underline{L}^{\ell}(t) - \underline{B}'(t)\underline{V}_{\ell}(t)]\underline{\Phi}_{\ell}(t, t_o)\underline{x}_o\underline{x}_o'\underline{\Phi}_{\ell}'(t, t_o)dt = \underline{0}$$

The results of Theorem 10 may be extended to cover yet another situation. In Section IV.C we showed that if the initial state $\underline{x}_o$ is a random variable which is uniformly distributed over the surface of the unit hypersphere, $\langle \underline{x}_o, \underline{x}_o\rangle = 1$, then the gain matrix $\underline{L}^o(\cdot)$ which minimizes $\mu(\underline{L})$ also minimizes the expected value of

$$J(\underline{x}_o, t_o, T, \underline{u}(\cdot))\Big|_{\underline{u}=-\underline{L}(t)\underline{x}} = \langle \underline{x}_o, \underline{V}_L(t_o)\underline{x}_o\rangle \qquad (4.37)$$

as $\underline{x}_o$ varies over the unit hypersphere. Suppose now, that the initial state $\underline{x}_o$ is a random variable which is uniformly distributed over the surface of a p-dimensional ($p\leq n$) ellipsoid, described mathematically by $\langle \underline{x}_o, \underline{P}\underline{x}_o\rangle = 1$, where $\underline{P}$ is a positive semidefinite matrix of rank p. We now wish to choose $\underline{L}(\cdot)\epsilon\Lambda_{NM}$ which minimizes the expected value of Eq.(4.37) as $\underline{x}_o$ varies over the given ellipsoid. In this case we can easily show, by making use of the fact $E(\underline{x}_o\underline{x}_o') = \underline{P}$, that

Corollary 2: If $\underline{x}_o$ is uniformly distributed over the surface of the ellipsoid $\langle \underline{x}_o, \underline{P}\underline{x}_o\rangle = 1$, and if $\hat{\underline{L}}^{\ell}(\cdot)\epsilon\Lambda_{NM}$

minimizes the expected value (over $\underline{x}_o$) of $\langle \underline{x}_o, \underline{V}_L(t_o)\underline{x}_o \rangle$ then for all $i = 0, 1, \ldots, N-1$; $j=1, \ldots, M$

$$\int_{t_i}^{t_{i+1}} \alpha_{ij}(t)[\hat{\underline{L}}^{\ell}(t) - \underline{B}'(t)\hat{\underline{V}}_{\ell}(t)]\,\hat{\underline{\Phi}}_{\ell}(t, t_o)\underline{P}\hat{\underline{\Phi}}_{\ell}'(t, t_o)dt = \underline{0}$$

Note that Corollary 1 becomes a special case of Corollary 2 by taking $p = 1$.

Theorem 10 gives only necessary conditions for $\underline{L}^o(\cdot) \epsilon \Lambda_{NM}$ to locally minimize $\mathrm{tr}\, \underline{V}_L(t_o)$. We have not shown sufficiency of this condition nor have we shown that for our given set of functions $\alpha_{ij}(t)$ there exists a unique $\underline{L}^o(\cdot) \epsilon \Lambda_{NM}$ satisfying Eq. (4.32). These are matters for future research.

A useful property of the function $\underline{L}^o(\cdot)$ which may be of help in any further investigation of our suboptimal problem is

Lemma 7: If $\underline{L}^o(\cdot) \epsilon \Lambda_{NM}$ minimizes $\mathrm{tr}\, \underline{V}_L(t_o)$ then for any $\underline{L}(\cdot) \epsilon \Lambda_{NM}$, the n x n matrix

$$\underline{S} = \int_{t_o}^{T} \underline{L}'(t)[\underline{B}'(t)\underline{V}_o(t) - \underline{L}^o(t)]\,\underline{\Phi}_o(t, t_o)\underline{\Phi}_o'(t, t_o)dt = \underline{0} \qquad (4.38)$$

Proof: We write the integral (4.38) as a sum of integrals, viz

$$\underline{S} = \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} \underline{L}'(t)[\underline{B}'(t)\underline{V}_o(t) - \underline{L}^o(t)]\,\underline{\Phi}_o(t, t_o)\underline{\Phi}_o'(t, t_o)dt$$

but on the interval $(t_i, t_{i+1}]$,

$$\underline{L}'(t) = \sum_{j=1}^{M} a_{ij}(t)\underline{L}'_{ij}$$

so that

$$\underline{S} = \sum_{i=0}^{N-1} \sum_{j=1}^{M} \underline{L}'_{ij} \int_{t_i}^{t_{i+1}} a_{ij}(t)[\underline{B}'(t)\underline{V}_o(t) - \underline{L}^o(t)]\underline{\Phi}_o(t, t_o)\underline{\Phi}'_o(t, t_o)dt$$

Employing Theorem 10 we find that each integral in the above summation equals zero, establishing the required result. ||

As an immediate application of Lemma 7 we can obtain the following result which may also be of use in further investigations.

Lemma 8: If $\underline{L}^o(\cdot) \epsilon \Lambda_{NM}$ minimizes $\mu(\underline{L})$, then for any $\underline{L}(\cdot) \epsilon \Lambda_{NM}$ with associated cost matrix $\underline{V}_L(t)$,

$$\mu(\underline{L}) - \mu(\underline{L}^o) = tr \int_{t_o}^{T} [(\underline{L}-\underline{L}^o)'(\underline{L}-\underline{L}^o) - 2(\underline{L}-\underline{L}^o)'(\underline{B}'\underline{V}_L - \underline{B}'\underline{V}_o)]\underline{\Phi}_o(t, t_o)\underline{\Phi}'_o(t, t_o)dt \tag{4.39}$$

Proof: Taking the trace of Eq. (C.15) with $\underline{L}_1 = \underline{L}$, $\underline{L}_2 = \underline{L}^o$ yields

$$\mu(\underline{L}) - \mu(\underline{L}^o) = tr \int_{t_o}^{T} [(\underline{L}-\underline{L}^o)'(\underline{L}-\underline{L}^o) - 2(\underline{L}-\underline{L}^o)'(\underline{B}'\underline{V}_L - \underline{L}^o)]\underline{\Phi}_o(t, t_o)\Phi'_o(t, t_o)dt \tag{4.40}$$

The second term in the above expression may be rewritten as

$$\operatorname{tr}\int_{t_o}^{T}(\underline{L}-\underline{L}^o)'(\underline{B}'\underline{V}_L-\underline{L}^o)\underline{\Phi}_o(t,t_o)\underline{\Phi}_o'(t,t_o)dt$$

$$=\operatorname{tr}\int_{t_o}^{T}(\underline{L}-\underline{L}^o)'[(\underline{B}'\underline{V}_L-\underline{B}'\underline{V}_o)+(\underline{B}'\underline{V}_o-\underline{L}^o)]\underline{\Phi}_o(t,t_o)\underline{\Phi}_o'(t,t_o)dt \tag{4.41}$$

Now since $\underline{L}$ and $\underline{L}_o$ are elements of Λ_{NM} and since Λ_{NM} is a linear space, $(\underline{L}-\underline{L}^o)\epsilon\Lambda_{NM}$, and by Lemma 7

$$\operatorname{tr}\int_{t_o}^{T}(\underline{L}-\underline{L}^o)'[\underline{B}'\underline{V}_o-\underline{L}^o]\underline{\Phi}_o(t,t_o)\underline{\Phi}_o'(t,t_o)dt=0 \tag{4.42}$$

Therefore, substituting Eqs.(4.42) and (4.41) into Eq.(4.40) yields the desired result. ||

In this chapter we developed and investigated the new concept of a suboptimal linear regulator problem as defined in Section C. In Section D we showed that under certain completeness assumptions, the solution of the <u>suboptimal</u> linear regulator problem approached the solution of the <u>optimal</u> linear regulator problem as $M\rightarrow\infty$. The mathematical and engineering implications of this result were discussed. In the final section we developed necessary conditions for $\underline{L}^o(\cdot)\epsilon\Lambda_{NM}$ to minimize $\mu(\underline{L})$, and derived some simple properties of the suboptimal solution.

In the following chapter we shall apply the results of Section IV.E to investigate a specific, yet important, special case. The case in question will be that in which the gain matrix $\underline{L}(t)$ is constrained to be constant over each interval $(t_i, t_{i+1}]$, i.e., $\underline{L}(t)$ is piecewise constant on $[t_o, T]$.

CHAPTER V

SUBOPTIMAL PIECEWISE CONSTANT GAIN MATRICES

A. INTRODUCTION

In general, the implementation of a gain matrix of the class Λ_{NM} in a linear feedback loop precipitates the use of M, time-synchronized tape channels. As we have seen, the signals which must be recorded on these tapes correspond to the functions $a_{ij}(t)$ as indicated in Eq. (4.6). There is, however, one special case for which no tape recorders are needed, and it is because of this property that the case in question assumes major practical importance. The situation to which we are alluding, and which will be investigated in this chapter, is that in which the gain matrices are constrained to be piecewise constant over the control interval $[t_o, T]$.

A piecewise constant constraint, such as the one envisioned here, arises naturally in engineering practice. Suppose for example, that the system Σ to be controlled is time invariant and the terminal time T is finite. Then, even in this case, the control which minimizes $J(\underline{x}_o, t_o, T, \underline{u}(\cdot))$, and which is given by

$$\underline{u}^*(\underline{x}, t) = -\underline{B}'(t)\,\underline{K}(t; T, \underline{F})\underline{x} = -\underline{L}^*(t)\underline{x}$$

represents a linear time-varying feedback law. This fact presents a slight delemma--an engineer might be unwilling to instrument time-varying quantities in an otherwise stationary system. On the other hand, it seems reasonable to assume that he would settle for piecewise

constant control gains (chosen in such a manner as to keep the resulting cost "close" to the optimal cost) in lieu of purely time varying ones.

To incorporate a piecewise-constant constraint into the suboptimal framework introduced in Chapter IV we simply take $M = 1$ and

$$a_{ij}(t) = 1 \quad \text{for } i = 0, 1, \ldots, N-1 \tag{5.1}$$

Consequently, the set Λ_{N1} is given by

$$\Lambda_{N1} = \{\underline{L}(\cdot): \underline{L}(t) = \underline{L}_{ij} \text{ for } t\epsilon(t_i, t_{i+1}] \; i = 0, 1, \ldots, N-1\} \tag{5.2}$$

and the suboptimization problem is to choose the matrices $\{\underline{L}_{ij}\}$ which minimize $\mu(\underline{L}) = \text{tr}\, \underline{V}_{\underline{L}}(t_o)$. For ease of notation we shall, in this chapter, drop the double subscripts on Λ_{N1}, $\underline{L}_{ij}$ and instead write Λ_N and $\underline{L}_i$ respectively. The suboptimization problem we shall subsequently investigate is repeated for convenience.

Piecewise-Constant Suboptimal Linear Regulator Problem

Let $N \geq 1$ be a fixed integer and let $t_i, i = 1, \ldots, N$ be a given set of times such that

$$t_o < t_1 < \ldots < t_{N-1} < t_N = T \tag{5.3}$$

Determine the element $\underline{L}^o(\cdot)\epsilon\Lambda_N$, where

$$\Lambda_N = \{\underline{L}(\cdot): \underline{L}(t) = \underline{L}_i \text{ for } t\epsilon(t_i, t_{i+1}] \; i = 0, 1, \ldots, N-1\} \tag{5.4}$$

which minimizes

$$\mu(\underline{L}) = \text{tr } \underline{V}_{\underline{L}}(t_o) \tag{5.5}$$

Hence, $\underline{L}^o(\cdot)$ is characterized by

$$\underline{L}^o(\cdot) = \arg \min_{\underline{L}(\cdot)\epsilon\Lambda_N} \mu(\underline{L}) \tag{5.6}$$

and we let $\underline{L}_i^o$, $i = 0, 1, \ldots, N-1$ be the set of constant matrices which characterize $\underline{L}^o(\cdot)$.†

B. PROPERTIES OF THE SUBOPTIMAL SOLUTION AS $N \to \infty$

Before presenting a theoretical exposé of the piecewise constant suboptimization problem, let us briefly examine the implications of constraining $\underline{L}(\cdot)\epsilon\Lambda_N$. Since there is no explicit time variation in the structure of $\underline{L}(\cdot)$, the necessity of having playback tapes in the feedback loop is alleviated. The implementation of a piecewise constant gain matrix therefore requires only $r \cdot n$ gain apmplifiers whose gains must be readjusted at the times t_i to correspond with the elements of $\underline{L}_i$. For instance if $g_{k\ell}(t)$ is the gain of the $k\ell$-th amplifier, where k and ℓ run through the integers $1 \to r$ and $1 \to n$ respectively, then

$$g_{k\ell}(t) = \underline{L}_i^{(k\ell)} \quad \text{for } t\epsilon(t_i, t_{i+1}] \; i = 0, 1, \ldots, N-1 \tag{5.7}$$

---

† Note that if $N = 1$, $\underline{L}^o(\cdot)$ is simply constant over $[t_o, T]$.

The gain adjustments indicated by Eq. (5.7) can be easily effected by a small, special purpose digital computer. The computer must store numbers which correspond to the various gain increments (or decrements) of the feedback amplifiers at times t_i, $i = 1, \ldots, N-1$. To be more precise, we must store the gain increment matrices

$$\delta \underline{\underline{L}}_i \triangleq \underline{\underline{L}}_i - \underline{\underline{L}}_{i-1}$$

for $i = 1, \ldots, N-1$. The $k\ell$-th element of $\delta \underline{\underline{L}}_i$ is the amount by which the gain of the $k\ell$-th amplifier is to be adjusted at time t_i.

Therefore, the implementation of a gain matrix $\underline{\underline{L}}(\cdot) \epsilon \Lambda_N$ requires the storage of $(r \cdot n)(N-1)$ numbers in the memory banks of an on-line digital computer. However, the storage capacity of a computer is limited, notably that of a small, special purpose machine. This places a high premium on storage space, especially if several memory banks have been set aside for a purpose other than storing the matrices $\delta \underline{\underline{L}}_i$. These storage factors will generally suggest an a priori choice of N. The larger we wish N to be, the more we must be willing to spend for increased storage requirements.

However, there is a trade-off to be sought here. As N increases it is reasonable to expect that we can choose the times $t_1, t_2, \ldots, t_{N-1}$ so that $\underline{\underline{L}}^o(\cdot)$ becomes a better and better approximation to $\underline{\underline{L}}^*(\cdot)$ over the interval $[t_o, T]$. This is only natural inasmuch as we are allowing ourselves a finer subdivision of the control interval.

Therefore, we wish to investigate the limiting behavior of $\underline{L}_N^o(\cdot)$ and μ_N^o as $N \to \infty$, where we write

$$\underline{L}_N^o(\cdot) = \arg\min_{\underline{L}(\cdot) \in \Lambda_N} [\operatorname{tr} \underline{V}_{\underline{L}}(t_o)] \tag{5.8}$$

and

$$\mu_N^o = \operatorname{tr} \underline{V}_{\underline{L}}(t_o) \Big|_{\underline{L}(\cdot) = \underline{L}_N^o(\cdot)} \tag{5.9}$$

to indicate the dependence of these quantities upon N. We can then show

Theorem 11: For each N, $N = 1, 2, \ldots$ let t_i $i = 0, 1, \ldots, N$ be a prescribed set of times such that

$$t_o < t_1 < \ldots < t_{N-1} < t_N = T$$

and such that as $N \to \infty$

$$|t_{i+1} - t_i| \to 0 \quad \text{for all } i = 0, 1, \ldots, N-1 \tag{5.10}$$

then

(i) $$\lim_{N \to \infty} \mu_N^o = \operatorname{tr} \underline{K}(t_o) = \mu(\underline{L}^*)$$

(ii) $$\lim_{N \to \infty} \| \underline{L}_N^o(\cdot) - \underline{L}^*(\cdot) \|_2 = 0$$

Proof: (i) We define the gain matrix $\underline{L}_N^a(\cdot) \in \Lambda_N$ by

$$\underline{L}_N^a(t) = \left(\frac{1}{t_{i+1} - t_i}\right) \int_{t_i}^{t_{i+1}} \underline{B}'(\tau) \underline{K}(\tau; T, \underline{F}) d\tau, \quad \text{for } t \in (t_i, t_{i+1}) \tag{5.11}$$

so that $\underline{L}_N^a(t)$ for $t\epsilon(t_i, t_{i+1}]$ is simply the average value of $\underline{L}^*(t)$ over the same interval. Now, since $|t_{i+1} - t_i| \to 0$ as $N \to \infty$ we have

$$\lim_{N\to\infty} \underline{L}_N^a(t) = \underline{L}^*(t) \tag{5.12}$$

But on the other hand if we write

$$\mu_N^a = \mu(\underline{L}) \Big|_{\underline{L}(\cdot) = \underline{L}_N^a(\cdot)}$$

then, since $\underline{L}_N^o(\cdot)$ minimizes $\mu(\underline{L})$ over Λ_N,

$$\mu(\underline{L}^*) \leq \mu_N^o \leq \mu_N^a \tag{5.13}$$

By virtue of Eq. (5.12), $\mu_N^a \to \mu(\underline{L}^*)$ as $N \to \infty$, since $\mu(\underline{L})$ is continuous in $\underline{L}(\cdot)$. Therefore, taking the limit of 5.13 we obtain

$$\lim_{N\to\infty} \mu_N^o = \mu(\underline{L}^*)$$

(ii) The proof of this result follows the same reasoning as in the proof of Theorem 9 which may be found in Appendix G. ‖

There is a great deal of similarity between Theorem 11 and Theorem 9. The latter theorem obtains the same conditions (i) and (ii) for fixed N as $M \to \infty$. For the case in question, $M = 1$ with $N \to \infty$; and we may regard $\underline{L}_N^o(\cdot)$ as a convergent minimizing sequence[19] for $\mu(\underline{L})$.

The results of Theorem 11 suggest a further analytical study regarding the convergence properties of μ_N^o and $L_N^o(\cdot)$ as $N \to \infty$. In particular, a study of this sort is extremely useful from a practical point of view, because as N increases the more we begin to tax the storage requirements of our digital computer. We shall have more to say on this and other unsolved problems of this nature in Chapter VI.

Having formulated the piecewise constant suboptimization problem and having obtained some properties of its solution as a function of N we now turn our attention to the problem of computing $\underline{L}^o(\cdot)$ for a fixed value of N. This is the object of the next section.

C. A COMPUTATIONAL SCHEME FOR DETERMINING $\underline{L}^o(\cdot)$

The arguments advanced in the proof of Theorem 11 suggest, that for a fixed value of N, the matrices

$$\underline{L}_i^a = \left(\frac{1}{t_{i+1}-t_i}\right) \int_{t_i}^{t_{i+1}} \underline{L}^*(t)\, dt \tag{5.14}$$

may serve as a good approximation to $\underline{L}_i^o$ for $i = 0, 1, \ldots, N-1$. If such is indeed the case it would seem that the calculation of the matrices $\underline{L}_i^o$ is superfluous. While for large N, Theorem 11 leads us to expect conclusions of this sort, there is no basis to expect such results when the time intervals $(t_i, t_{i+1}]$ are of the same order of magnitude as $(t_o, T]$. And it is this latter case which is of greatest interest from a practical point of view. Since the matrices $\underline{L}_i^o$

always yield a more satisfactory system performance (in the sense of minimizing $\mu(\underline{L})$), than do the matrices $\underline{L}_i^a$, the determination of $\underline{L}_i^o$ becomes a matter of considerable interest.

The most important thing to realize in the above situation is that the matrices $\underline{L}_i^o$ are computed off-line, before the control system is actually placed into operation. Hence, the determination of $\underline{L}_i^o$ is done at leisure, which presents yet another argument for the implementation of $\underline{L}_i^o$ as opposed to $\underline{L}_i^a$. In this section we therefore develop a computational algorithm for determining the matrices $\underline{L}_i^o$, and investigate the convergence properties of the proposed scheme. In the following section we shall illustrate its use by way of a numerical example.

We assume that the integer N is given and that the times $t_1, t_2, \ldots, t_N$ are also given. Therefore, the necessary conditions which must be satisfied if the sequence $\underline{L}_i^o$, $i = 0, 1, \ldots, N-1$ is to minimize $\mu(\underline{L})$ over the class Λ_N are readily obtained from Theorem 10 with $M = 1$, and $a_{ij}(t) = 1$. They are

$$\int_{t_i}^{t_{i+1}} [\underline{L}_i^o - \underline{B}'(t)\underline{V}_o(t)]\underline{\Phi}_o(t, t_o)\underline{\Phi}_o'(t, t_o)dt = \underline{0}; \quad i=0, 1, \ldots, N-1 \tag{5.15}$$

where $\underline{V}_o(t)$ is the cost matrix associated with $\underline{L}^o(t)$ and $\underline{\Phi}_o(t, t_o)$ is the transition matrix corresponding to $\underline{L}^o(t)$ i.e.,

$$\frac{d}{dt}\underline{\Phi}_o(t, t_o) = [\underline{A}(t) - \underline{B}(t)\underline{L}^o(t)]\underline{\Phi}_o(t, t_o); \underline{\Phi}_o(t_o, t_o) = \underline{I} \tag{5.16}$$

Note that we can also write for any $i = 0, 1, \ldots, N-1$

$$\underline{\Phi}_o(t, t_o) = \underline{\Phi}_o(t, t_i)\underline{\Phi}_o(t_i, t_o) \quad \text{for} \quad t\epsilon(t_i, t_{i+1}] \tag{5.17}$$

The constant matrix $\underline{\Phi}_o(t_i, t_o)$ depends only on $\underline{L}_j^o$ for $j=0, 1, \ldots, i-1$ and $\underline{\Phi}_o(t, t_i)$ depends only on $\underline{L}_i^o$, viz

$$\frac{d}{dt}\underline{\Phi}_o(t, t_i) = [\underline{A}(t)-\underline{B}(t)\underline{L}_i^o]\underline{\Phi}_o(t, t_i); \underline{\Phi}_o(t_i, t_i) = \underline{I}, t\epsilon(t_i t_{i+1}] \tag{5.18}$$

so that

$$\underline{\Phi}_o(t_i, t_o) = \prod_{j=0}^{i-1} \underline{\Phi}_o(t_{j+1}, t_j) \tag{5.19}$$

Expressions 5.17 through 5.19 can be quite useful in any numerical investigations, especially when $\underline{A}(t)$ and $\underline{B}(t)$ are constant matrices, for in this case

$$\underline{\Phi}_o(t, t_i) = e^{(\underline{A}-\underline{B}\underline{L}_i^o)(t-t_i)} \tag{5.20}$$

Returning to Eq. (5.15) we see that it may be written in another form by noting that since $\underline{\Phi}_o(t, t_o)$ is invertible for all t, $\underline{\Phi}_o(t, t_o)\underline{\Phi}_o'(t, t_o)$ is positive-definite. Hence

$$\underline{L}_i^o = \int_{t_i}^{t_{i+1}} \underline{B}'(t)\underline{V}_o(t)\underline{\Phi}_o(t, t_o)\underline{\Phi}_o'(t, t_o)dt - \left[\int_{t_i}^{t_{i+1}} \underline{\Phi}_o(t, t_o)\underline{\Phi}_o'(t, t_o)dt\right]^{-1} \tag{5.21}$$

It is virtually impossible to obtain an explicit analytic expression for $\underline{L}_i^o$ and so the need arises for developing an iterative scheme to solve Eq. (5.21). To achieve this end we let the sequence $\underline{L}_i^n$, $i = 0, 1, \ldots, N-1$ be the n-th iterate to $\underline{L}_i^o$, $i = 0, 1, \ldots, N-1$. The first iterate $\{\underline{L}_i^1\}$ is arbitrarily chosen; we shall have more to say on this point

in the sequel. In addition, we let $\underline{V}_n(t)$ be the cost matrix associated with the n-th iterate gain matrix $\underline{L}_i^n(\cdot)$, with $\underline{\Phi}_n(t,t_o)$ denoting the transition matrix corresponding to $\underline{L}_i^n(\cdot)$. Therefore,

$$\underline{\Phi}_n(t,t_o) = \underline{\Phi}_n(t,t_i)\prod_{j=0}^{i-1}\underline{\Phi}_n(t_{j+1},t_j) = \underline{\Phi}_n(t,t_i)\underline{\Phi}(t_i,t_o) \quad \text{for} \quad t\epsilon(t_i,t_{i+1}]$$

where

$$\frac{d}{dt}\underline{\Phi}_n(t,t_j) = [\underline{A}(t)-\underline{B}(t)\underline{L}_j^n]\underline{\Phi}_n(t,t_j); \quad t\epsilon(t_j,t_{j+1}]$$

$$\underline{\Phi}_n(t_j,t_j) = \underline{I}$$

and

$$\underline{V}_n(t) = \int_t^T \underline{\Phi}_n'(\tau,t)[\underline{C}'(\tau)\underline{C}(\tau) + \underline{L}'^n(\tau)\underline{L}^n(\tau)]\underline{\Phi}_n(\tau,t)d\tau$$

Given the set of matrices $\{\underline{L}_i^n\}$, the matrices $\underline{V}_n(t)$ and $\underline{\Phi}_n(t,t_o)$ are uniquely determined for all t. Finally we define

$$\mu^n = \mu(\underline{L})\Big|_{\underline{L}(\cdot)=\underline{L}^n(\cdot)} \tag{5.22}$$

We want to develop a sequence of iterates $\{\underline{L}^n(\cdot)\}=\{\underline{L}^1(\cdot),\underline{L}^2(\cdot),\ldots\}$ such that as $n\to\infty$

$$\underline{L}^n(\cdot) \to \underline{L}^o(\cdot)$$

$$\mu^n \to \mu^o = \mu(\underline{L})\Big|_{\underline{L}(\cdot) = \underline{L}^o(\cdot)} \tag{5.23}$$

In such a case we will have obtained a solution to Eq. (5.21) which yields a (local) minimum for $\mu(\underline{L})$. If we could show, by some means, that Eq. (5.21) has a unique solution then indeed, we would also obtain a global minimum of $\mu(\underline{L})$.

The form of Eq. (5.21) suggests an iterative scheme based on the method of successive approximations.† If $\underline{L}_i^n$, $i = 0, 1, \ldots, N-1$ is our n-th iterate, then the n+1-st iterate is obtained as

$$\underline{L}_i^{n+1} = \int_{t_i}^{t_{i+1}} \underline{B}'(t)\underline{V}_n(t)\underline{\Phi}_n(t, t_o)\underline{\Phi}_n'(t, t_o)dt \cdot \left[\int_{t_i}^{t_{i+1}} \underline{\Phi}_n(t, t_o)\underline{\Phi}_n'(t, t_o)dt\right]^{-1} \tag{5.24}$$

for $i = 0, 1, \ldots, N-1$. This iterative scheme is conceptually very simple. It is based on the fact that $\underline{L}_i^o$ is a fixed-point of Eq. (5.24). However, as $n \to \infty$ we have not as yet been able to show that convergence is obtained, although heuristically such a conclusion might seem reasonable. The difficulty is that we cannot guarantee $\mu^{n+1} \leq \mu^n$.

---

† Often referred to as Picard's method.

We therefore change our tack and approach the problem from a slightly different direction. Given the n-th iterate $\{\underline{L}_i^n\}$, we define for $i = 0, 1, \ldots, N-1$,

$$\hat{\underline{L}}_i^n = \int_{t_i}^{t_{i+1}} \underline{B}'(t)\underline{V}_n(t)\underline{\Phi}_n(t, t_o)\underline{\Phi}_n'(t, t_o)dt \cdot \left[\int_{t_i}^{t_{i+1}} \underline{\Phi}_n(t, t_o)\underline{\Phi}_n'(t, t_o)dt\right]^{-1} \tag{5.25}$$

so that $\hat{\underline{L}}_i^n$ is associated with $\underline{L}_i^n$, and as our (n+1)-st iterate we write

$$\underline{L}_i^{n+1} = \underline{L}_i^n + \epsilon_n[\hat{\underline{L}}_i^n - \underline{L}_i^n] \tag{5.26}$$

where ϵ_n is a positive parameter between zero and one which is to be chosen to insure that

$$\mu^{n+1} < \mu^n$$

Hence $\underline{L}_i^{n+1}$ may be regarded as a "better" approximation to $\underline{L}_i^o$ than was $\underline{L}_i^n$.†

We shall investigate the convergence properties of the iterative scheme suggested by Eqs. (5.25) and (5.26) as a function of the parameter ϵ_n. We note that if $\epsilon_n = 1$, $\underline{L}_i^{n+1} = \hat{\underline{L}}_i^n$, which is just the method of successive approximations, embodied in Eq. (5.24). If

---

† This approach is basically a "gradient" technique[33,34] for determining the minimum of the functional $\mu(\underline{L})$. We shall have more to say concerning this similarity in the sequel.

$\epsilon_n \approx 0$ then $\underline{L}_i^{n+1} \approx \underline{L}_i^n$ and the above iterative scheme, if it converges at all, will converge rather slowly. We may therefore regard ϵ_n as a convergence parameter; we would like to choose ϵ_n as close to unity as possible while still being able to assure convergence of our iterations. The main result along this line is

Theorem 12: Given any set of N matrices $\underline{L}_i^n$, i=0, 1, . . . , N-1. Let $\underline{L}_i^{n+1}$ be determined by Eq. (5.26). Then for sufficiently small ϵ_n

$$\mu^{n+1} \leq \mu^n \tag{5.27}$$

with equality holding if and only if $\underline{L}_i^n = \underline{L}_i^o$ for all i.

Proof: We use Eq. (C.15) to write

$$\mathrm{tr}\underline{V}_n(t_o) - \mathrm{tr}\underline{V}_{n+1}(t_o) = \mu^n - \mu^{n+1}$$

$$= \sum_{i=0}^{N-1} \mathrm{tr}\{(\underline{L}_i^n - \underline{L}_i^{n+1})' \int_{t_i}^{t_{i+1}} [(\underline{L}_i^n - \underline{L}_i^{n+1}) - 2(\underline{B}'\underline{V}_n(t) - \underline{L}_i^{n+1})]\underline{\Phi}_{n+1}(t, t_o)\underline{\Phi}'_{n+1}(t, t_o)\,dt\} \tag{5.28}$$

But to first-order in ϵ_n, since

$$\underline{L}_i^n - \underline{L}_i^{n+1} = -\epsilon_n(\hat{\underline{L}}_i^n - \underline{L}_i^n)$$

we obtain

$$\mu^n - \mu^{n+1} = 2\epsilon_n \sum_{i=0}^{N-1} \mathrm{tr}\{(\hat{\underline{L}}_i^n - \underline{L}_i^n)' \int_{t_i}^{t_{i+1}} [\underline{B}'(t)\underline{V}_n(t) - \underline{L}_i^{n+1}]\underline{\Phi}_n(t, t_o)\underline{\Phi}'_n(t, t_o)\,dt\}$$

Introducing Eq. (5.25) for $\hat{\underline{L}}_i^n$ yields

$$\mu^n - \mu^{n+1} = 2\epsilon_n \sum_{i=0}^{N-1} \mathrm{tr}\{(\hat{\underline{L}}_i^n - \underline{L}_i^n)'(\hat{\underline{L}}_i^n - \underline{L}_i^{n+1}) \int_{t_i}^{t_{i+1}} \underline{\Phi}_n(t, t_o)\underline{\Phi}_n'(t, t_o)dt\}$$

which, again, to first order in ϵ_n becomes, upon substituting Eq. (5.26)

$$\mu^n - \mu^{n+1} = 2\epsilon_n \,\mathrm{tr} \sum_{i=0}^{N-1} (\hat{\underline{L}}_i^n - \underline{L}_i^n)'(\hat{\underline{L}}_i^n - \underline{L}_i^n) \int_{t_i}^{t_{i+1}} \underline{\Phi}_n(t, t_o)\underline{\Phi}_n'(t, t_o)dt \qquad (5.28a)$$

But each matrix in the summation has a <u>strictly</u> positive trace, and so to first order in ϵ_n we have

$$\mu^n > \mu^{n+1} \qquad (5.27)$$

Equality may hold in Eq. (5.27) if and only if $\underline{L}_i^n = \hat{\underline{L}}_i^n$ for all i. This implies that $\underline{L}_i^n$ is a fixed-point of Eq. (5.25) and therefore also a fixed-point of Eq. (5.24). Hence $\underline{L}_i^n = \underline{L}_i^o$ for all i. ‖

Stated in different terms, Theorem 12 may be recast as follows

<u>Corollary 1</u>: Given any set of N matrices $\underline{L}_i^n$, i=0, 1, . . . , N-1. If $\underline{L}_i^n \neq \underline{L}_i^o$ for all i, then there always exists a number $\epsilon_n > 0$ such that for $\underline{L}_i^{n+1}$ as defined by Eq. (5.26)

$$\mu^{n+1} < \mu^n$$

In order to develop an iterative scheme for determining $\underline{L}_i^o$ which is based on Theorem 12 and its corollary it is necessary to be able to choose an ϵ_n, given the n-th iteration $\underline{L}_i^n$, which will guarantee $\mu^{n+1} < \mu^n$. We know that for ϵ_n sufficiently small we can guarantee this situation. The only question is "how small must ϵ_n be for $\mu^{n+1} < \mu^n$?" In other words we would like to obtain a bound M_ϵ such that for $\epsilon_n < M_\epsilon$, $\mu^{n+1} < \mu^n$.

In order to answer this question we are faced with a difficult mathematical problem. We must make an a priori choice of ϵ_n such that the right hand side of Eq. (5.28) is positive. One approach to the problem is to examine the second order terms in ϵ_n of Eq. (5.28), assuming that ϵ_n^3 is negligible with respect to ϵ_n. We first write

$$\underline{\Phi}_{n+1}(t, t_o) = \underline{\Phi}_n(t, t_o) + \int_{t_o}^{t} \underline{\Phi}_n(t, \tau)\underline{B}(\tau)[\underline{L}^n(\tau) - \underline{L}^{n+1}(\tau)]\underline{\Phi}_{n+1}(\tau, t_o)d\tau$$

which to first order in ϵ_n is

$$\underline{\Phi}_{n+1}(t, t_o) = \underline{\Phi}_n(t, t_o) - \epsilon_n \underline{S}_n(t, t_o) \qquad (5.29)$$

where, for $t \epsilon (t_i, t_{i+1}]$,

$$\underline{S}_n(t,t_o) = \sum_{j=0}^{i-1} \int_{t_j}^{t_{j+1}} \underline{\Phi}_n(t,\tau)\underline{B}(\tau)[\hat{\underline{L}}_j^n - \underline{L}_j^n]\underline{\Phi}_n(\tau,t_o)d\tau$$

$$+ \int_{t_i}^{t} \underline{\Phi}_n(t,\tau)\underline{B}(\tau)[\hat{\underline{L}}_i^n - \underline{L}_i^n]\underline{\Phi}_n(\tau,t_o)d\tau$$

and where we have written

$$\underline{L}_j^n - \underline{L}_j^{n+1} = -\epsilon_n(\hat{\underline{L}}_j^n - \underline{L}_j^n) \quad \text{for} \quad j=0,1,\ldots i$$

If we now substitute Eq. (5.29) into Eq. (5.28) we obtain, to second order in ϵ_n,

$$\mu^n - \mu^{n+1} = \epsilon_n \cdot \alpha - \epsilon_n^2 \cdot \beta \tag{5.30}$$

where

$$\alpha = 2\,\mathrm{tr} \sum_{i=0}^{N-1} (\hat{\underline{L}}_i^n - \underline{L}_i^n)'(\hat{\underline{L}}_i^n - \underline{L}_i^n) \int_{t_i}^{t_{i+1}} \underline{\Phi}_n(t,t_o)\underline{\Phi}_n'(t,t_o)dt$$

and

$$\beta = \frac{\alpha}{2} + 2\,\mathrm{tr} \sum_{i=0}^{N-1} (\hat{\underline{L}}_i^n - \underline{L}_i^n)' \int_{t_i}^{t_{i+1}} [\underline{B}'\underline{V}_n(t) - \underline{L}_i^{n+1}][\underline{S}_n(t,t_o)\underline{\Phi}_n'(t,t_o) + \underline{\Phi}_n(t,t_o)\underline{S}_n'(t,t_o)]\,dt$$

Both β and α depend only on $\underline{L}_i^n$, i=0, 1, . . . , N-1 and therefore may be computed before choosing ϵ_n. In order to choose ϵ_n which will assure $\mu^{n+1} < \mu^n$ we therefore require, by Eq. (5.30),

$$0 < \epsilon_n < \frac{\alpha}{|\beta|} \dagger \tag{5.31}$$

Hence, if ϵ_n^2 is negligible compared with unity and if Eq. (5.31) is satisfied, we can guarantee $\mu^{n+1} < \mu^n$.

The above argument is not entirely convincing, it gives us no feel for choosing ϵ_n if ϵ_n^2 is comparable to unity. While μ^{n+1} may be smaller than μ^n under these latter circumstances, an investigation for this case is exceedingly involved and would entail a study of the detailed form of Eq. (5.28).

One way out of this dilemma is to simply propose an ad hoc rule for picking ϵ_n which will assure $\mu^n < \mu^{n+1}$. One such method is as follows. We first pick $\epsilon_n = 1$ and check whether $\mu^{n+1} < \mu^n$. If not, we set $\epsilon_n = 1/2$ and again see if $\mu^{n+1} < \mu^n$. By successively dividing ϵ_n by 2, we will eventually reach a value of ϵ_n for which $\mu^{n+1} < \mu^n$, as guaranteed by Theorem 12.

Motivated by the foregoing remarks, we propose the following iterative scheme for determining the matrices $\underline{L}_i^o$, i=0, 1, . . . , N-1.

---

† A convenient choice of ϵ_n is $\frac{\alpha}{2|\beta|}$ for which (5.30) is maximized. Note also that if $\beta < 0$, $\mu^{n+1} < \mu^n$ for all ϵ_n.

Iterative Scheme for Computing the Suboptimal Gain Matrices $\{\underline{L}_i^o\}$

1. Guess an initial iterate $\{\underline{L}_i^1, i=0,1,\ldots,N-1\}$
2. Calculate $\underline{\Phi}_1(t,t_o)$, $\underline{V}_1(t)$ and $\{\hat{\underline{L}}_i^1,\ i=0,1,\ldots,N-1\}$
3. Set $\underline{L}_i^2 = \underline{L}_i^1 + \epsilon_1[\hat{\underline{L}}_i^1 - \underline{L}_i^1]$ for $i=0,1,\ldots,N-1$
4. Take $\epsilon_1=1$ and check to see if $\mu^2 < \mu^1$
5. If $\mu^2 > \mu^1$, set $\epsilon_1=1/2$ and again check if $\mu^2 < \mu^1$
6. If μ^2 is still not less than μ^1, keep dividing ϵ_1 by 2 until $\mu^2 < \mu^1$.
7. With $\{\underline{L}_i^2\}$ chosen via step 3, calculate $\underline{\Phi}_2(t,t_o)$, $\underline{V}_2(t)$ and $\{\hat{\underline{L}}_i^2\}$.
8. Repeat steps 3 through 7 until μ^{n+1} is sufficiently close to μ^n for some n.

The choice of the initial iterate $\{\underline{L}_i^1\}$ at step 1 in the proposed iterative scheme will naturally effect the rates of convergence of $\underline{L}_i^n$ to $\underline{L}_i^o$ and of μ^n to μ^o. Therefore, as an initial guess, we would like to be able to choose $\underline{L}_i^1 \approx \underline{L}_i^o$. Motivated by the introductory remarks of Section V.A, a suitable choice for $\underline{L}_i^1$ is

$$\underline{L}_i^1 = \underline{L}_i^a = \left(\frac{1}{t_{i+1}-t_i}\right)\int_{t_i}^{t_{i+1}} \underline{B}'(t)\underline{K}(t;T,\underline{F})dt; \quad i=0,1,\ldots,N-1$$

Besides being a reasonable initial iteration, this choice enables us to easily see the improvement in system performance which arise

from using the gain matrix $\underline{L}^o(\cdot)$ as opposed to $\underline{L}^a(\cdot)$. The matrices $\underline{L}_i^a$ can be calculated by any one of the iterative schemes suggested in Section III.C. In the following section we shall make these ideas clearer by way of a numerical example.

In the foregoing iterative scheme we note that if $\hat{\underline{L}}_i^n$ is close to $\underline{L}_i^n$ (in norm) for all i, then the parameter ϵ_n may be chosen close to one while still guaranteeing $\mu^{n+1} < \mu^n$. This is because the approximations of $(\mu^n - \mu^{n+1})$ which were made in Theorem 12 are valid as long as $\|\epsilon_n(\hat{\underline{L}}_i^n - \underline{L}_i^n)\|$ is small compared with unity. Consequently, if $\|\hat{\underline{L}}_i^n - \underline{L}_i^n\|$ is small we may choose $\epsilon_n \approx 1$ while still being able to write the first order expansion to $\mu^n - \mu^{n+1}$ as in Eq. (5.28a). We have therefore shown that as our iterative technique converges in the limit for large n, we can take values for ϵ_n which approach one.

Finally, we wish to point attention to the similarity between this iterative scheme to find the minimum the functional $\mu(\underline{L})$ and the familiar gradient or steepest descent methods.† Both schemes introduce a small, adjustable parameter ϵ into the problem (in the gradient scheme, ϵ is referred to as the "step size") and choose ϵ to assure a decrease in cost at each iteration. Therefore, eventual convergence to a local minimum is attained. There are many variations of the gradient method which can be used to improve rates of

---

†See Ref. 33 for a discussion of this method as well as for an extensive list of references pertaining to this subject.

convergence, suggesting that it may be possible to modify our iterative scheme to yield a more rapid convergence rate. Several modifications are presently under investigation, yet this still remains a subject which warrants a considerable amount of further research.

D. A NUMERICAL EXAMPLE

In order to ellucidate the iterative scheme proposed in the foregoing section we shall determine the matrices $\underline{L}_i^o$ for a second-order system. The system under consideration is time invariant and is characterized by the matrices

$$\underline{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \underline{B} = \underline{b} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \underline{C} = \underline{c}' = [1 \quad 0]$$

We shall take the initial time $t_o = 0$ and the terminal time $T = 2$ with $\underline{F} = \underline{0}$. The solution of the Riccati equation, $\underline{K}(t; 2, \underline{0})$, corresponding to these matrices is shown in Fig. 5.1 for $t \leq 2$.† The optimal gain matrix $\underline{L}^*(t) = \underline{b}'\underline{K}(t; 2, \underline{0})$ is

$$\underline{L}^*(t) = [k_{11}(t) \quad k_{22}(t)] \quad \text{for} \quad t \in [0, 2]$$

and the optimal cost $\mu(\underline{L}^*)$ is

$$\mu(\underline{L}^*) = \text{tr } \underline{K}(0) = 2.5144$$

---

† These results were obtained by using the discrete optimization technique of Section III.C.

We shall examine the piecewise constant suboptimization problem for two cases and compute, using our iterative scheme, the suboptimal gain matrices. The two cases are a) N=1 with $t_o = 0$, $t_1 = 2$ and b) N=2 with $t_o = 0$, $t_1 = 1$, $t_2 = 2$. In both cases we shall begin our iterations with the initial guess $\underline{L}_i^a$ as we discussed earlier. All calculations were done on a PDP-1 computer.

a) $N = 1$; $t_o = 0$, $t_1 = 2$.

For this case we seek the constant matrix $\underline{L}_1^o$ and the suboptimal cost μ^o. For our initial iteration we chose

$$\underline{L}_1^1 = [.4842 \quad .4423] = \underline{L}_1^a$$

The results of applying our iterative scheme are tabulated in Table 5a. It was found that for $\epsilon_n = 1$ at each iteration, the cost μ^{n+1} was always smaller than μ^n. Consequently it was never necessary to reduce ϵ_n, and $\underline{L}_1^{n+1} = \hat{\underline{L}}_1^n$. The iterative scheme converged to a suboptimal cost

$$\mu^o = 2.6283$$

(which is within 5 percent of the optimal cost) and to

$$\underline{L}_1^o = [.8095 \quad 1.1668]$$

This result is displayed graphically in Fig. 5.2.

In this example it was found that $\mu(\underline{L})$ is relatively insensitive to small changes in $\underline{L}(\cdot)$ in the vicinity of $\underline{L}(\cdot) = \underline{L}^o(\cdot)$, since μ^n had converged to μ^o before $\underline{L}_1^n$ converged to $\underline{L}_1^o$. It was also noted

that for a wide range of initial iterates, $\underline{L}_1^1$, the iterative scheme displayed the same convergence properties: for $\epsilon_n = 1$, μ^n monotonically decreased to μ^o and $\underline{L}_1^n \rightarrow \underline{L}_1^o$.

b) $N = 2$; $t_o = 0$, $t_1 = 1$, $t_2 = 2$

In this case the suboptimal gain matrix $\underline{L}^o(\cdot)$ is constant over each of the intervals $(0,1]$ and $(1,2]$ so that we seek the matrices $\underline{L}_1^o$ and $\underline{L}_2^o$ as well as the suboptimal cost μ^o. For an initial iteration we chose

$$\underline{L}_1^1 = [.8043 \quad .8013] = \underline{L}_1^a$$

$$\underline{L}_2^1 = [.1640 \quad .0833] = \underline{L}_2^a$$

The numerical results obtained with our iterative scheme are shown in Table 5b. As in case a), μ^n monotonically decreased to μ^o with $\epsilon_n = 1$ at each iteration (so that $\underline{L}_i^{n+1} = \hat{\underline{L}}_i^n$ for $i = 1,2$). The suboptimal cost μ^o was found to be

$$\mu^o = 2.5490$$

which is smaller than that of case a), showing the improvement of choosing $N = 2$ over $N = 1$, and is only 2 percent larger than the optimal cost $\mu(\underline{L}^*)$. The matrices $\underline{L}_i^o$ are

$$\underline{L}_1^o = [.8723 \quad 1.0442]$$

$$\underline{L}_2^o = [.1810 \quad .0652]$$

and are displayed graphically in Fig. 5.3.

The results in Table 5b show that μ^n converges to μ^o in 2 iterations while $\underline{L}_i^n$ converges in 6 iterations to $\underline{L}_i^o$. This again shows the insensitivity in $\mu(\underline{L})$ to small changes in $\underline{L}(\cdot)$ about $\underline{L}(\cdot) = \underline{L}^o(\cdot)$. As was also noticed in case a), choosing different initial iterates still resulted in a rapid monotone convergence of μ^n to μ^o for $\epsilon_n = 1$.

These results point to the feasibility of applying the proposed iterative scheme to determine the piecewise constant suboptimal gain matrices. The potential use of this technique for designing suboptimal regulator systems is great, yet much remains to be done in its improvement, modification and analysis. Research along these lines is currently being persued, and a computer program is being written which will calculate the matrices $\underline{L}_i^o$ for an n-th order, time invariant system Σ.

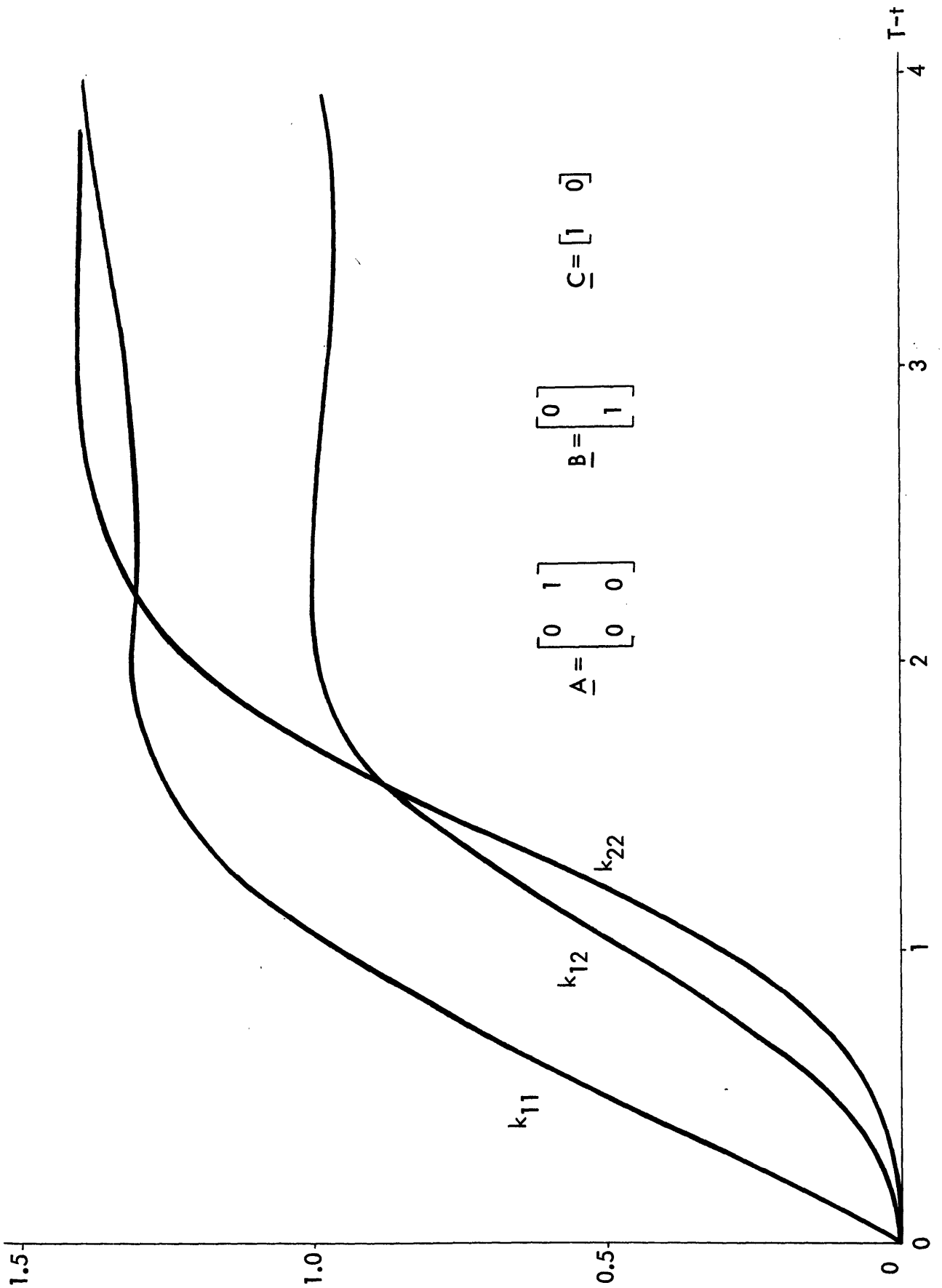


Fig. 5.1 The Riccati Equation Solution for $t<T$, $T=2$

n	$(\underline{L}^n)^{11}$	$(\underline{L}_1^n)^{12}$	μ^n
1	.4842	.4423	2.9475
2	.5843	.9378	2.6791
3	.6989	1.1035	2.6377
4	.7582	1.1519	2.6303
5	.7859	1.1646	2.6288
6	.7985	1.1673	2.6284
7	.8044	1.1675	2.6284
8	.8071	1.1673	2.6283
9	.8084	1.1671	2.6283
10	.8090	1.1670	2.6283
11	.8093	1.1669	2.6283
12	.8095	1.1668	2.6283

Table 5a. Iterative Scheme Results for Case a)

$N = 1,\ t_o = 0,\ t_1 = 2,\ \epsilon_n = 1$

n	$(\underline{L}_1^n)^{11}$	$(\underline{L}_1^n)^{12}$	$(\underline{L}_2^n)^{11}$	$(\underline{L}_2^n)^{12}$	μ^n
1	.8043	.8013	.1640	.0833	2.5673
2	.8497	1.0058	.1725	.0634	2.5490
3	.8662	1.0388	.1788	.0640	2.5490
4	.8708	1.0438	.1806	.0650	2.5490
5	.8720	1.0442	.1809	.0652	2.5490
6	.8723	1.0442	.1810	.0652	2.5490

Table 5b. Iterative Scheme Results for Case b)

$N = 2,\ t_o = 0,\ t_1 = 1,\ t_2 = 2,\ \epsilon_n = 1$

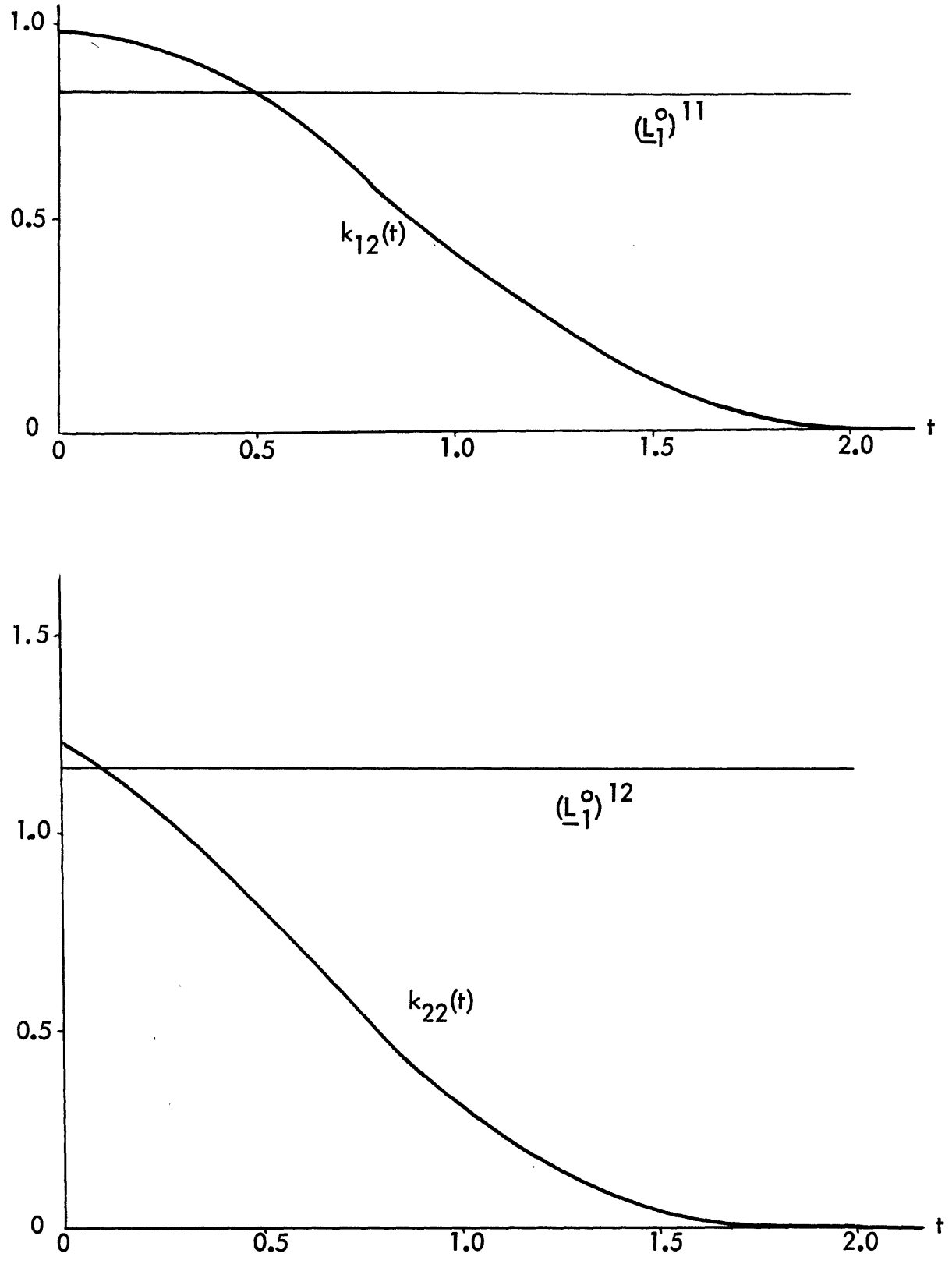
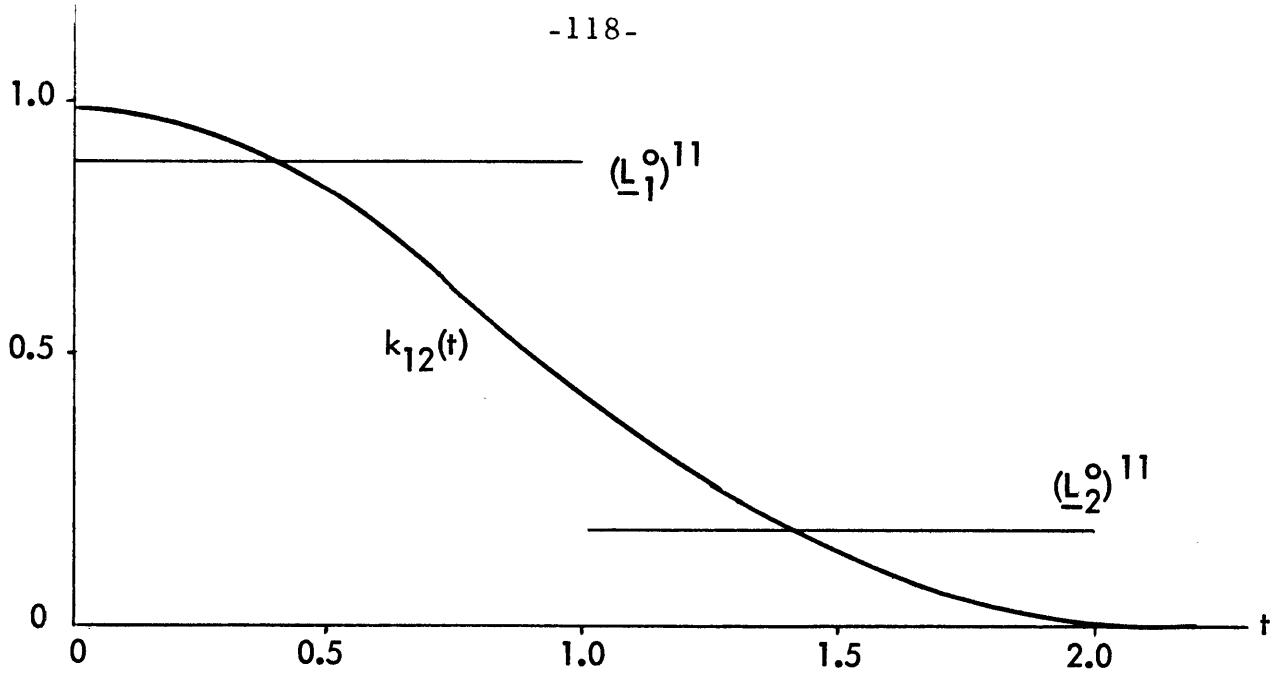


Fig. 5.2 Optimal and Suboptimal Gain Matrices for Case a
$N=1,\ t_o=0,\ t_1=2$



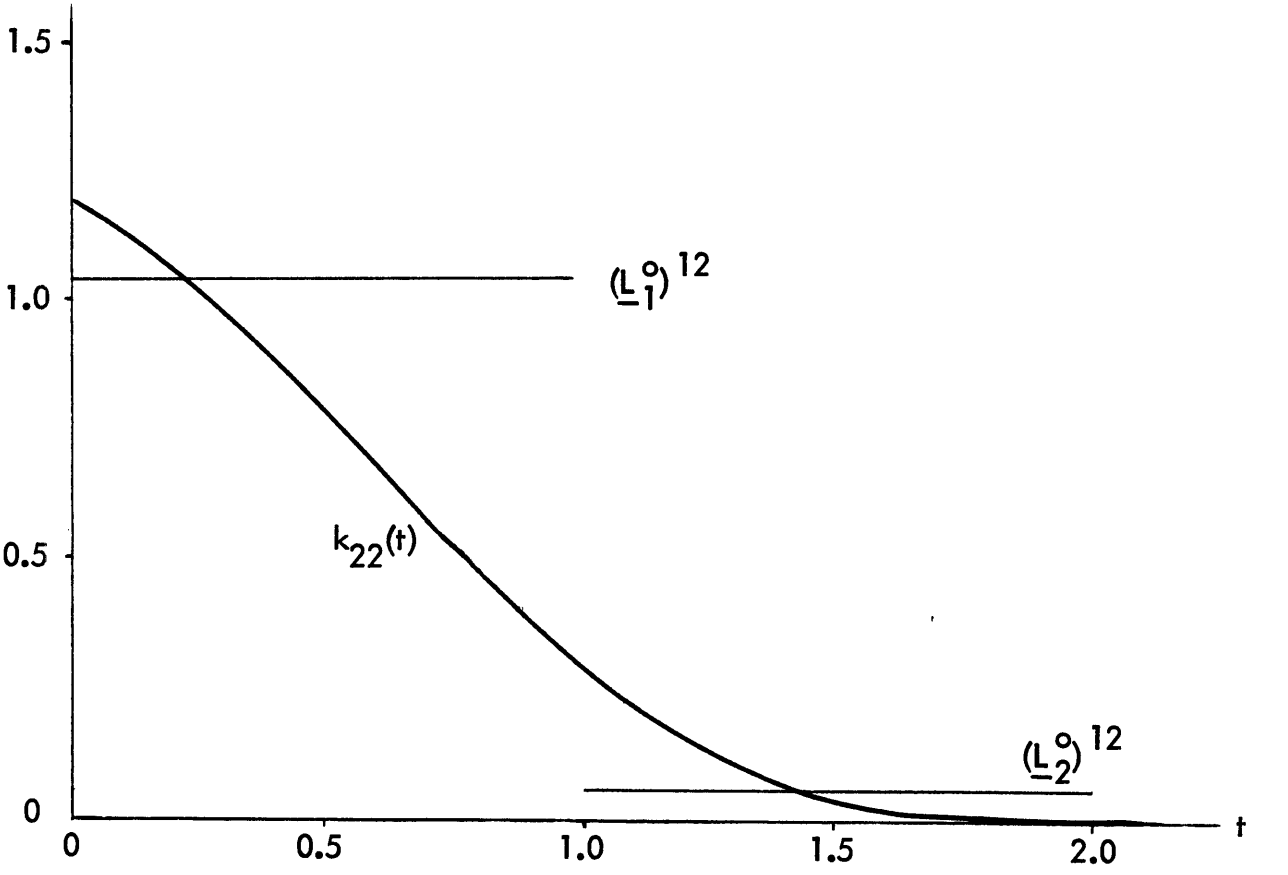


Fig. 5.3 Optimal and Suboptimal Gain Matrices for Case b
$N=2,\ t_o=0,\ t_1=1,\ t_2=2$

CHAPTER VI

TOPICS FOR FURTHER RESEARCH

Throughout this report we have mentioned, and in some cases briefly discussed, subjects which warrant further investigation. The majority of these topics are directly related to the results presented in Chapters IV and V, and range from purely theoretical to entirely numerical studies. In this chapter we shall enumerate these additional research problems while categorizing them for the benefit of future investigators.

A. THEORETICAL STUDIES

1. Sufficient Conditions for $\underline{L}^o(\cdot) \in \Lambda_{NM}$ to Minimize $\mu(\underline{L})$

Theorem 10 gives only necessary conditions for $\underline{L}^o(\cdot) \in \Lambda_{NM}$ to minimize $\mu(\underline{L}) = \text{tr } \underline{V}_L(t_o)$. It is then appropriate to inquire whether these conditions are also sufficient.

One approach to a sufficiency proof is by showing that Eq. (4.32) has a unique solution. If such is the case then, per force, this solution must necessarily minimize $\mu(\underline{L})$.† However, this method of attack can be restrictive since there may exist several solutions of Eq. (4.32). Under these latter circumstances, the sufficiency of Theorem 10 is guaranteed by merely showing that if $\underline{L}^o(\cdot)$ satisfies Eq. (4.32) then

$$\mu(\underline{L}^o) \leq \mu(\underline{L}) \qquad \text{for all } \underline{L}(\cdot) \in \Lambda_{NM},$$

with equality holding if and only if $\underline{L}(\cdot)$ satisfies Eq. (4.32). A useful tool in such an investigation is given by Lemma 8 in which we derived an expression for $\mu(\underline{L}) - \mu(\underline{L}^o)$.

---

† With the underlying assumption that there exists an element of Λ_{NM} which minimizes $\mu(\underline{L})$.

If, in general, Theorem 10 is not a sufficient as well as necessary condition, we would then like to determine the additional constraints (on the functions $a_{ij}(t)$ or on the system Σ) which must be placed into the problem statement in order to guarantee the sufficiency of Theorem 10.

2. Optimal Determination of the Times t_i, $i = 0, 1, \ldots, N$

Throughout this report we have assumed that the set of times $\{t_i, i = 0, 1, \ldots, N\}$ are specified in advance of solving for $\underline{L}^o(\cdot)$. This a priori choice of t_i has its disadvantages. It would be more desirable if these times were left free to be chosen in such a manner as to minimize $\mu(\underline{L})$. In other words the new sub-optimization problem would be to choose the set of times t_i and the element $\underline{L}^o(\cdot) \in \Lambda_{NM}$ which minimize $\mu(\underline{L})$. We assume that the functions $a_{ij}(t)$ are pre-specified.

One very important special case which merits such an investigation is when the gain matrices are constrained to be piecewise constant as in Chapter V. In addition to determining the optimal set of piecewise constant gains, we then also seek the optimal set of times t_i. Hence we regard $\mu(\underline{L})$ as a function of the N matrices $\underline{L}_i$, $i = 0, 1, \ldots, N-1$ and the N-1 times t_i, $i = 1, \ldots, N-1$. This problem is particularly important if the storage limitation of our computer dictates that we can store only N matrices. Choosing the times t_i, $i = 1, \ldots, N-1$ in an optimal fashion is then equivalent to making the most beneficial use of the existing storage facility.

In any particular problem of this sort it is possible to make a reasonable choice for the times t_i by merely looking at the nature of the time variation of $\underline{L}^*(t) = \underline{B}'(t)\underline{K}(t; T, \underline{F})$. Over portions of the control interval $[t_o, T]$ where the time variation of $\underline{L}^*(t)$ is slight (i.e. $\| \frac{d}{dt} \underline{L}^*(t) \|$ is small) we can allow the times t_i to be more widely dispersed than over regions where $\underline{L}^*(t)$ varies with time to a greater extent. In any case we would like to obtain necessary (and perhaps sufficient) conditions for the set $\{t_i\}$ to minimize $\mu(\underline{L})$.

3. Determining Rates of Convergence for $\underline{L}^o_M(\cdot)$

In Section IV.C we introduced a set of functions $\{\phi_j\}$ which were complete on the interval $[0, \Delta]$, where $\Delta = (T-t_o)/N$, and we then let

$$\Lambda_{NM} = \{\underline{L}(\cdot): \underline{L}(t) = \sum_{j=1}^{M} \phi_j(t-i\Delta)\underline{L}_{ij} \text{ for } t \in (t_i, t_i+\Delta] \; i = 0, 1, \ldots, N-1\}$$

with the suboptimal gain matrix being defined as

$$\underline{L}^o_M(\cdot) = \underset{\underline{L}(\cdot) \in \Lambda_{MN}}{\arg\min.} \; \mu(\underline{L})$$

We showed that $\underline{L}^o_M(\cdot)$ converged to $\underline{L}^*(\cdot)$ in a mean square sense as $M \to \infty$ and that $\mu^o_M \to \mathrm{tr}\, \underline{K}(t_o)$.

We would like to be able to determine (for a given set $\{\phi_j\}$) bounds on the rates of convergence of these quantities. Such bounds will obviously depend upon both the original optimization problem itself and on the choice of the functions $\phi_j(t)$. It is then reasonable to ask, given the system Σ and the integer N, for the set of functions $\{\phi_j\}$ which

maximize the speed of convergence. Under such circumstances we could have $\underline{L}^{o}_{M}(\cdot)$ and μ^{o}_{M} being quite satisfactory approximations to $\underline{L}^{*}(\cdot)$ and $\text{tr}\,\underline{K}(t_o)$, respectively, for relatively small M.

If, for a specific optimization problem we could find such a set of functions, the results would indeed be fruitful. We would then have a control law which is simple to implement, yet one which results in near optimal system performance.

Experience with the Ritz method[19] of the calculus of variations has substantiated similar claims for numerous problems in the field of physics. However, there has been no application of the techniques of this direct method to the solution of optimal control problems.

B. NUMERO-THEORETIC INVESTIGATIONS

1. Computer Storage Versus System Performance

In Chapter V where the gain matrix $\underline{L}(\cdot)$ was constrained to be piecewise constant, we showed that $\underline{L}^{o}_{N}(\cdot)$ converged to $\underline{L}^{*}(\cdot)$ as $N \to \infty$. The implication of allowing N to be large is that we have a correspondingly large amount of computer storage at our convenience. This, unfortunately is rarely the case. Suppose, therefore, that initial storage limitations dictate an a priori choice of N. With N fixed we can then proceed to determine the choice of matrices $\underline{L}_i$, $i = 0, 1, \ldots, N-1$ and the set of times $t_1, t_2, \ldots, t_{N-1}$ which minimize $\mu(\underline{L})$. In this manner we are using the alloted computer storage to its maximum advantage. For a fixed N we therefore define

$$\underline{L}_N^o(\cdot) = \arg \min_{t_1, \ldots, t_{N-1}} \{ \min_{\underline{L}(\cdot) \in \Lambda_N} \operatorname{tr} \underline{V}_L(t_o) \} \tag{6.1}$$

and

$$\mu_N^o = \operatorname{tr} \underline{V}_L(t_o) \Big|_{\underline{L}(\cdot) = \underline{L}_N^o(\cdot)} \tag{6.2}$$

Clearly, $\mu_{N+1}^o \leq \mu_N^o$ for all N and by Theorem 11 we have that $\mu_N^o \to \operatorname{tr} \underline{K}(t_o)$ as well as $\underline{L}_N^o(\cdot) \Rightarrow \underline{L}^*(\cdot)$.

The question to which we now address ourselves is the following. If we allot more memory facility to the task of storing the elements of the matrices $\underline{L}_i$ it then becomes possible for us to increase the integer N. Correspondingly, the number μ_N^o will be closer to its minimum value of $\operatorname{tr} \underline{K}(t_o)$, and we also expect $\underline{L}_N^o(\cdot)$ to be a finer approximation to $\underline{L}^*(\cdot)$. However, the increase in storage can only be justified by a measurable decrease in μ_N^o. If the difference $\mu_N^o - \mu_{N+1}^o$ is insignificant then we have reached a point of diminishing returns insofar as our sub-optimization problem is concerned. Therefore, the correlation between N, the rate of decrease of μ_N^o to $\operatorname{tr} \underline{K}(t_o)$, and the rate at which $\underline{L}_N^o(\cdot)$ converges to $L^*(\cdot)$ become matters of paramount importance in view of their relationship to the cost of additional storage registers.

2. Accuracy of $\underline{L}_N^o(\cdot)$ Versus Computer Storage Limitations

The storage of $\underline{L}_N^o(\cdot)$, or equivalently the matrices $\underline{L}_i^o$, $i = 0, 1, \ldots, N-1$, in the memory banks of a digital computer requires that we specify the numerical accuracy to which we wish to carry the elements of these matrices. This is necessary so that a sufficient number of core registers may be allocated to the task of storing each of the $(r.n)N$ numbers.

$$(\underline{L}_i^o)^{k\ell} \quad ; i = 0, 1, \ldots, N-1; \; k = 1, \ldots, r \; ; \; \ell = 1, \ldots, n \tag{6.3}$$

Since a number is stored in a digital manner (i.e. in base 2) this implies that one core element is necessary to store each significant figure † of $(\underline{L}_i^o)^{k\ell}$. Therefore, if we wish to store the elements of the suboptimal gain matrix to six significant figures, we require $6 \cdot (r \cdot n)N$ core elements.

Let us now suppose that we decide to store these numbers to an accuracy of only four significant figures. We then require only $4 \cdot (r.n)N$ core elements. What this implies, therefore, is that we can increase the number N by a factor of $3/2$ while keeping the number of core elements constant. On the one hand increasing N will result in a lower value for μ_N^o, but on the other hand this increase in performance is likely to be offset by rounding off the elements of $\underline{L}_i^o$ to two less significant figures. However, if the control system is <u>insensitive</u> to slight perturbations in $\underline{L}_N^o(\cdot)$ it is reasonable to expect that an improvement in system performance can be made by simply storing numbers to a lesser degree of accuracy. This suggests a study of the sensitivity of μ_N^o to perturbations in $\underline{L}_N^o$. The object of this study would be to determine, given a fixed storage capacity, the value of N and the number of significant figures with which to store $\underline{L}_N^o(\cdot)$ so as to minimize $\mu(\underline{L})$. (There are hardware factors to consider here also, such as the accuracy of the Digital-Analog converter, etc.).

---

†The number of significant figures is often referred to as "word length".

3. Improvement of Iterative Schemes to Determine $\underline{L}_N^o(\cdot)$

In Section V.D we developed and discussed an algorithm for the determination of the matrices $\underline{L}_i^o$, $i = 0, 1, \ldots, N-1$ for the case when the times $t_o, t_1, \ldots, t_N$ were prespecified. The computational scheme is basically a "gradient" technique. At each iteration we choose a convergence parameter ϵ_n † to assure a decrease in cost. Therefore, we will eventually converge to $\underline{L}_N^o(\cdot)$, which is the element of Λ_N which is the element of Λ_N which absolutely minimizes the cost functional $\mu(\underline{L})$.

There has been a considerable amount of research done in the past, dealing with gradient techniques.[33, 34] These investigations treat subjects ranging from gradient scheme modifications for more rapid convergence to prescriptions for choosing the step size at each iteration. In view of the wealth of knowledge which exists in this area, it is feasible that one could apply these ideas to improve or modify the basic gradient scheme of Section V.D.

There is another way to look at the piecewise-constant suboptimization problem posed in Chapter V which may suggest a different computational approach to the problem of finding $\underline{L}_N^o(\cdot)$. The suboptimization problem can be cast into a framework which suggests the application of the discrete minimum principle. To do this we proceed as follows. We assume that the times t_i are fixed and we let $\underline{L}(\cdot)$ be an element of Λ_N. Let $\underline{\Phi}(t, t_o)$ be the transition matrix corresponding to $\underline{L}(\cdot)$. Then, as in Eq. (5.17), we may write, since $\underline{L}(\cdot)$ is piecewise-constant,

† often referred to as the "step size".

$$\underline{\Phi}(t, t_o) = \underline{\Phi}(t, t_i)\,\underline{\Phi}(t_i, t_o) \quad \text{for } t \in (t_i, t_{i+1}] \tag{6.4}$$

Let us denote $\underline{\Phi}(t_i, t_o)$ by $\underline{\Phi}_i$, noting that $\underline{\Phi}_o = \underline{I}$, viz,

$$\underline{\Phi}(t_i, t_o) = \underline{\Phi}_i \quad \text{for } i = 0, 1, \ldots, N-1 \tag{6.5}$$

so that in particular

$$\underline{\Phi}_{i+1} = \underline{\Phi}(t_{i+1}, t_i)\,\underline{\Phi}_i \tag{6.6}$$

We now write $\mu(\underline{L})$ as a sum of integrals, noting that $\underline{L}(t)$ is piecewise-constant. Hence

$$\mu(\underline{L}) = \text{tr } \underline{V}_L(t_o) = \text{tr} \int_{t_o}^{t_N} \underline{\Phi}'(t, t_o)\left[\underline{C}'(t)\underline{C}(t) + \underline{L}_i'\underline{L}_i\right]\underline{\Phi}(t, t_o)\,dt$$

$$= \sum_{i=0}^{N-1} \text{tr} \int_{t_i}^{t_{i+1}} \underline{\Phi}'(t, t_o)\left[\underline{C}'(t)\underline{C}(t) + \underline{L}_i'\underline{L}_i\right]\underline{\Phi}(t, t_o)\,dt \tag{6.7}$$

Substituting Eqs. (6.4) and (6.5) into the above yields

$$\mu(\underline{L}) = \sum_{i=0}^{N-1} \text{tr}\left\{\underline{\Phi}_i' \int_{t_i}^{t_{i+1}} \underline{\Phi}'(t, t_i)\left[\underline{C}'(t)\underline{C}(t) + \underline{L}_i'\underline{L}_i\right]\underline{\Phi}(t, t_i)dt \cdot \underline{\Phi}_i\right\} \tag{6.8}$$

$\underline{\Phi}(t, t_i)$ depends solely on $\underline{L}_i$, hence we define the positive semidefinite matrix

$$\underline{Q}(\underline{L}_i) = \int_{t_i}^{t_{i+1}} \underline{\Phi}'(t, t_i)\left[\underline{C}'(t)\underline{C}(t) + \underline{L}_i'\underline{L}_i\right]\underline{\Phi}(t, t_i)\,dt \tag{6.9}$$

and we obtain

$$\mu(\underline{L}) = \sum_{i=0}^{N-1} \operatorname{tr}\ \underline{\Phi}_i' \underline{Q}(\underline{L}_i) \underline{\Phi}_i \qquad (6.10)$$

The minimization problem is now the following: "Given the cost functional $\mu(\underline{L})$ where $\underline{\Phi}_i$ is generated by the difference equation

$$\underline{\Phi}_{i+1} = \underline{\Phi}(t_{i+1}, t_i)\underline{\Phi}_i \quad ; \quad \underline{\Phi}_o = \underline{I} \qquad (6.11)$$

Determine the sequence $\underline{L}_o, \underline{L}_1, \ldots, \underline{L}_{N-1}$ which minimizes $\mu(\underline{L})$".

As reformulated above the piecewise constant suboptimization problem suggests the use of the discrete minimum principle. The "discrete dynamical system" is Eq. (6.11), with $\underline{\Phi}_i$ being the system "state" at time t_i. The "control variable" at time t_i is the matrix $\underline{L}_i$. To apply a minimum principle for the case of "state matrices" such as above, we define the inner product $< \cdot, \cdot >$ between two matrices by

$$< \underline{A}, \underline{B} > = \operatorname{tr} \underline{A}\,\underline{B}' \qquad (6.12)$$

which is a valid inner product over a matrix space. Under this assumption, the application of the minimum principle is straightforward. If we introduce a "costate matrix" the minimum principle will yield a set of two matrix difference equations with split boundary conditions. By developing iterative schemes for their solution (which is no easy task), it is then possible to determine the matrices $\underline{L}_i^o$, $i = 0, 1, \ldots, N-1$ by means other than a pure gradient technique.

4. Development of Iterative Schemes for $M \geq 2$

In Section IV.E we derived necessary conditions for $\underline{L}^o(\cdot) \in \Lambda_{NM}$ to minimize $\mu(\underline{L})$. These conditions are embodied in Theorem 10, via Eq. (4.32). We would like to obtain a means of solving Eq. (4.32) for

$\underline{L}^{o}(.)$ once the times $\{t_i\}$ and the functions $\{a_{ij}(t)\}$ are specified. In particular we wish to investigate the case for which $M > 1$, since for $M = 1$ the analysis of Chapter V is pertinent.

One of the more important cases for which $M > 1$ is that characterized by $M = 2$ and

$$\left.\begin{aligned} a_{i1}(t) &= 1 \\ a_{i2}(t) &= t - t_i \end{aligned}\right\} \quad \text{for } t \in (t_i, t_{i+1}] \text{ , } i = 0, 1, \ldots, N-1 \tag{6.13}$$

The importance and interest in deriving algorithms for this case arises because a gain matrix belonging to the set Λ_{N2} is <u>piecewise-linear</u> over any interval $(t_i, t_{i+1}]$. A piecewise-linear function may be implemented by simply performing real-time linear interpolation in a feedback loop with a small, special purpose digital computer. Therefore, an analysis of the problem for $M = 2$ is supported by engineering feasibility, despite the increase in technical difficulty in going from $M = 1$ to $M = 2$.

CHAPTER VII

SUMMARY

In this chapter we summarize our approach to the development and study of suboptimal linear regulator problems, and stress, in a categorical fashion, the contributions of the thesis.

Our initial task was to formulate, in a precise manner, the well-known and often-studied optimal linear regulator problem. We discussed the solution to this optimal control problem in terms of the solution $\underline{K}(t; T, \underline{F})$ to the matrix Riccati differential equation. We showed that the optimal control may be constructed as a linear, time-varying feedback law given by

$$\underline{u}^*(\underline{x}, t) = -\underline{B}'(t)\, \underline{K}(t; T, \underline{F})\underline{x} = -\underline{L}^*(t)\underline{x} \qquad (7.1)$$

In addition, we presented several well-known properties of the Riccati equation solution, first for when the terminal time T is finite and then for $T = \infty$. In the latter case, we gave conditions assuring the stability of the optimal closed-loop system.

Having presented the reader with an understanding of the form of the optimal solution, we turned, in Chapter III, to methods for implementing the control law (7.1). We showed that due to the computational instability of the Riccati equation solutions, one cannot accurately compute $\underline{K}(t; T, \underline{F})$ in an on-line manner for $t > t_o$. This fact forces us to implement the optimal control by prestoring the

elements of $\underline{L}^*(t)$ on tape and playing the tape back upon command in real time to generate the control law (7.1). Therefore, the Riccati equation solution is computed off-line, before the control system is placed into operation. This lead to our study of numerical techniques for the off-line computation of $\underline{K}(t; T, \underline{F})$. We presented three known algorithms which are based upon approximating the nonlinear Riccati differential equation by a nonlinear difference equation, and we discussed some of the advantages and disadvantages of each scheme. We then developed an iterative scheme for determining $\underline{K}(t; T, \underline{F})$ which was an extension (to the matrix case) of Kalaba's method of successive approximations. By introducing the concept of a "cost matrix," and solving a sequence of linear differential equations, we obtained a sequence of iterates which converged monotonically to $\underline{K}(t; T, \underline{F})$.

In Chapter IV we discussed the engineering difficulties associated with storing the optimal feedback gain matrix $\underline{L}^*(t)$ on tape for $t_o \leq t \leq T$. Motivated by engineering feasibility, we then prescribed a time structure for the feedback gain matrix $\underline{L}(t)$ by requiring $\underline{L}(\cdot) \in \Lambda_{NM}$. We discussed at length the implications of such a constraint from a practical point of view as well as from a mathematical viewpoint. In Section IV.C we motivated the choice of a cost functional $\mu(\underline{L})$, and developed the new concept of a "suboptimal linear regulator problem." We showed that under certain assumptions, the solution

of this suboptimal problem converged to the solution of the optimal linear regulator problem as $M \to \infty$. We finally derived necessary conditions which the solution of the suboptimal problem must satisfy, as well as some properties of the suboptimal solution itself.

In Chapter V we examined the important special case for which the feedback gain matrix is constrained to be piecewise constant over the control interval $[t_o, T]$. We discussed the implications inherent in this constraint insofar as they relate to the storage limitations of a digital computer which may be used to implement the suboptimal feedback control law. We showed in Theorem 11, that if the storage facility of the computer is increased (i.e., if $N \to \infty$), the suboptimal piecewise-constant gain matrix $\underline{L}_N^o(t)$ approaches the optimal gain matrix $\underline{L}^*(t)$. We then proceeded to apply the necessary conditions for suboptimality derived in Chapter IV, and for a fixed value of N, we developed an iterative scheme for determining the suboptimal gain matrix $\underline{L}_N^o(\cdot)$. We illustrated this computational method by way of a numerical example which demonstrated the algorithm's effectiveness as a suboptimal design tool.

In the following chapter we discussed several problems for further research which arise in the study of suboptimal linear regulator problems. Most of these problems, if solved, would have a direct practical application to the design of linear regulator systems. Some topics are currently being investigated, yet much remains to be done in exploring the properties of the suboptimal regulator problem.

CHAPTER VIII

CONCLUSIONS

The feasibility of taking practical engineering constraints into consideration when designing optimal linear regulator systems has been investigated. This study was approached by prespecifying the structural form of time-varying feedback gains, while leaving various free parameters to be chosen in an optimal fashion. In this manner, a "suboptimal linear regulator problem" was defined and necessary conditions for its solution were obtained by introducing the concepts of a cost matrix and of a gradient matrix of a trace function.

For the special case when the feedback gains were constrained to be piecewise constant over the control interval of interest, an algorithm was developed for determining the required suboptimal gains. Limited computer experience with this algorithm has demonstrated its effectiveness as a useful tool in the suboptimal design of regulator systems.

APPENDIX A

THE EXISTENCE AND UNIQUENESS OF THE SOLUTION TO THE RICCATI EQUATION[14]

In this appendix we present a clear, detailed proof of the existence and uniqueness theorem for the matrix Riccati differential equation which arises in the solution of the optimal linear regulator problem of Chapter II. Recall that this equation is

$$\frac{d}{dt}\underline{K}(t) = -\underline{K}(t)\underline{A}(t) - \underline{A}'(t)\underline{K}(t) - \underline{C}'(t)\underline{C}(t) + \underline{K}(t)\underline{B}(t)\underline{B}'(t)\underline{K}(t) \quad (A.1)$$

with the boundary condition $\underline{K}(T) = \underline{F}$.

We begin our investigation by examining the local properties of the Riccati equation. For ease in analysis we shall consider, instead of Eq. (A.1), the equation

$$\frac{d}{dt}\underline{K}(t) = -\underline{K}'(t)\underline{A}(t) - \underline{A}'(t)\underline{K}(t) - \underline{C}'(t)\underline{C}(t) + \underline{K}'(t)\underline{B}(t)\underline{B}'(t)\underline{K}(t) \quad (A.2)$$

with the boundary condition $\underline{K}(T) = \underline{F}$. Note that there is no loss of generality here, since Eq. (A.2) possesses a unique solution if and only if Eq. (A.1) does and the solutions are identical in such an event as $\underline{K}(t)$ will equal $\underline{K}'(t)$.

Let us denote by $\mathcal{L}(X,X)$ the set of bounded linear mappings from X into itself (i.e., the set of nxn matrices) with the norm of an element $\underline{P} \in \mathcal{L}(X,X)$ defined as the norm induced by the Euclidean norm on X. To be more specific, if $\underline{P} \in \mathcal{L}(X,X)$, then

$$\|\underline{P}\|^2 = \sup_{\underline{x} \neq \underline{0}} \frac{\langle \underline{P}\underline{x}, \underline{P}\underline{x} \rangle}{\langle \underline{x}, \underline{x} \rangle} = \sup_{\|\underline{x}\| = 1} \langle \underline{P}\underline{x}, \underline{P}\underline{x} \rangle$$

$$= \lambda_{max}(\underline{P}'\underline{P}) = \text{maximum eigenvalue of } \underline{P}'\underline{P} \quad (A.3)$$

Note that if $\underline{P}$ is symmetric, then

$$\|\underline{P}\| = |\lambda_{max}(\underline{P})|$$

$$= \sup_{\|\underline{x}\| = 1} |<\underline{x}, \underline{P}\underline{x}>| \qquad (A.4)$$

If we now return to the problem at hand and set

$$\underline{f}(t, \underline{K}) = -\underline{K}'\underline{A}(t) - \underline{A}'(t)\underline{K} - \underline{C}'(t)\underline{C}(t) + \underline{K}'\underline{B}(t)\underline{B}'(t)\underline{K} \qquad (A.5)$$

it is easy to show that $\underline{f}(t, \underline{K})$ is a locally Lipschitzian mapping of $(-\infty, \infty) \times \mathcal{L}(X, X)$ into $\mathcal{L}(X, X)$ which is integrable in t. In fact we can show

Proposition: If $\underline{K}_1$ and $\underline{K}_2$ are arbitrary elements of $\mathcal{L}(X, X)$ then

$$\|\underline{f}(t, \underline{K}_1) - \underline{f}(t, \underline{K}_2)\| \leq (2\|\underline{A}(t)\| + \|\underline{B}(t)\|^2 \|\underline{K}_1 + \underline{K}_2\|)\|\underline{K}_1 - \underline{K}_2\| \qquad (A.6)$$

Proof: Since $\underline{f}(t, \underline{K})$ is symmetric, we have by Eq. (A.4), that

$$\|\underline{f}(t, \underline{K}_1) - \underline{f}(t, K_2)\| = \sup_{\|\underline{x}\| = 1} |<\underline{x}, [-(\underline{K}_1 - \underline{K}_2)'\underline{A}(t) - \underline{A}'(t)(\underline{K}_1 - \underline{K}_2)$$

$$+ \underline{K}_1'\underline{B}(t)\underline{B}'(t)\underline{K}_1 - \underline{K}_2'\underline{B}(t)\underline{B}'(t)\underline{K}_2]\underline{x}>|$$

$$\leq \|(\underline{K}_1 - \underline{K}_2)'\underline{A}(t) + \underline{A}'(t)(\underline{K}_1 - \underline{K}_2)\| + \sup_{\|\underline{x}\| = 1} |<\underline{x}, (\underline{K}_1 + \underline{K}_2)'\underline{B}(t)\underline{B}'(t)(\underline{K}_1 - \underline{K}_2)\underline{x}^2>|$$

where we have used the identity

$$<\underline{x}, (\underline{A}'\underline{A} - \underline{B}'\underline{B})\underline{x}> = <\underline{x}, (\underline{A} + \underline{B})'(\underline{A} - \underline{B})\underline{x}> \quad \text{for all } \underline{x}$$

with $\underline{A} = \underline{B}'(t)\underline{K}_1$, $\underline{B} = \underline{B}'(t)\underline{K}_2$. However

$$\sup_{\|\underline{x}\| = 1} |<\underline{x}, (\underline{K}_1 + \underline{K}_2)'\underline{B}(t)\underline{B}'(t)(\underline{K}_1 - \underline{K}_2)\underline{x}>| \leq \|(\underline{K}_1 + \underline{K}_2)\underline{B}(t)\underline{B}'(t)(\underline{K}_1 - \underline{K}_2)\|$$

so that we finally obtain, upon substituting and using the triangle inequality, the required result

$$\|\underline{f}(t,\underline{K}_1) - \underline{f}(t,\underline{K}_2)\| \leq (2\|\underline{A}(t)\| + \|\underline{B}(t)\|^2\|\underline{K}_1+\underline{K}_2\|)\,\|\underline{K}_1-\underline{K}_2\|,$$

where we have also used the fact that for induced matrix norms

$$\|\underline{A}\underline{B}\| \leq \|\underline{A}\| \cdot \|\underline{B}\|.$$

The above proposition implies that there exists an open interval about T, say (r_1, s_1), in which Eq. (A.2) has a unique solution $\hat{\underline{K}}(t)$ satisfying the condition $\hat{\underline{K}}(T) = \underline{F}$ (see Ref. 15). In particular, then, $\hat{\underline{K}}(t)$ is the unique solution of the Riccati equation in the interval $(r_1, T]$, and by lemma 1 of Chapter II, $\hat{\underline{K}}(t)$ is positive semi-definite for $t \in (r_1, T]$.

Let us now denote by S the set of points $s \in (-\infty, T]$ such that Eq. (A.2) possesses a unique solution $\hat{\underline{K}}(t)$ defined on the closed interval $[s, T]$ with $\hat{\underline{K}}(T) = \underline{F}$. By the foregoing arguments we see that S is non-empty; we wish to show that, in fact, $S = (-\infty, T]$.

Let us assume to the contrary that $S \neq (-\infty, T]$ and let $\sigma \neq -\infty$ be the greatest lower bound (g.l.b.) of S. The solution $\hat{\underline{K}}(t)$ is then defined on the interval $(\sigma, T]$. If we could show that the mapping $t \rightarrow \underline{f}(t, \hat{\underline{K}}(t))$ were bounded on $(\sigma, T]$, then by 10.5.5 of Ref. 15, we would be able to (uniquely) extend the solution $\hat{\underline{K}}(t)$ to an interval $(\sigma_1, T]$ with $\sigma_1 < \sigma$. Hence σ could not be the g.l.b. of S and it would follow that $S = (-\infty, T]$, so that the Riccati equation could not have a finite escape time† for $t < T$.

To show that $\underline{f}(t, \hat{\underline{K}}(t))$ has finite norm over any interval $(s, T]$, $s \in S$, it is sufficient to show that $\hat{\underline{K}}(t)$ is bounded on $(s, T]$, $s \in S$. This is the objective of the following lemma:

---

† The existence of a finite escape time for the Riccati equation corresponds to the existence of a "conjugate point" in the classical calculus of variations (see Ref. 19).

Lemma: There exists a positive semi-definite matrix, $\underline{H}(t)$, bounded for all $t \in (-\infty, T]$, such that for all $t \in (s, T]$, $\|\underline{H}(t)\| \geq \|\hat{\underline{K}}(t)\|$

Proof: In order to exhibit a suitable $\underline{H}(t)$, we note first that $\hat{\underline{K}}(t)$ is positive semi-definite. Furthermore, by virtue of the fact that $\hat{\underline{K}}(t)$ is associated with the optimal control, we have

$$0 \leq \frac{1}{2} < \underline{x}, \hat{\underline{K}}(t)\underline{x} > = J(\underline{x}, t, T, \underline{u}(\cdot))\Big|_{\underline{u}=\underline{u}^*} \leq J(\underline{x}, t, T, \underline{u}(\cdot))\Big|_{\underline{u} \equiv \underline{0}}$$

If now, $\underline{x} \in X$ and $t \in (s, T]$, the solution of the system state differential equation (2.1), starting from $\underline{x}$ at time t and generated by the control $\underline{u}(\cdot) \equiv \underline{0}$, is given by $\underline{x}(\tau) = \underline{\Phi}(\tau, t)\underline{x}$ where $\underline{\Phi}(\tau, t)$ is the transition matrix corresponding to $\underline{A}(t)$, i.e.,

$$\frac{d}{d\tau}\underline{\Phi}(\tau, t) = \underline{A}(\tau)\underline{\Phi}(\tau, t) \; ; \; \underline{\Phi}(t, t) = \underline{I}$$

It then follows, by substituting $\underline{y}(\tau) = \underline{C}(\tau)\underline{x}(\tau)$ and $\underline{u}(\tau)$ into Eq. (2.3) for $J(\underline{x}, t, T, \underline{u}(\cdot))$, that

$$J(\underline{x}, t, T, \underline{u}(\cdot))\Big|_{\underline{u} \equiv \underline{0}} = \frac{1}{2} < \underline{x}, \underline{H}(t)\underline{x} >$$

where

$$\underline{H}(t) = \underline{\Phi}'(T, t)\underline{F}\underline{\Phi}(T, t) + \int_t^T \underline{\Phi}'(\tau, t)\underline{C}'(\tau)\underline{C}(\tau)\underline{\Phi}(\tau, t)\, d\tau \qquad (A.7)$$

The nxn matrix $\underline{H}(t)$ is positive semi-definite and has finite norm for all $t \in (-\infty, T]$. Consequently,

$$0 \leq < \underline{x}, \hat{\underline{K}}(t)\underline{x} > \leq < \underline{x}, \underline{H}(t)\underline{x} > \qquad (A.8)$$

for all $\underline{x} \in X$ and $t \in (s, T]$. Therefore, by Eq. (A.8) we conclude that $\|\hat{\underline{K}}(t)\| \leq \|\underline{H}(t)\|$ as claimed. (Note that the proof would not remain valid if $\hat{\underline{K}}(t)$ were not positive semi-definite). ‖

In view of the above lemma and the remarks preceding it, we have proven the following theorem:

Theorem: For all T and all positive semi-definite matrices $\underline{F}$, the equation

$$\dot{\underline{K}}(t) = -\underline{K}(t)\,\underline{A}(t) - \underline{A}'(t)\underline{K}(t) - \underline{C}'(t)\underline{C}(t) + \underline{K}(t)\underline{B}(t)\underline{B}'(t)\underline{K}(t)$$

has a unique, positive, semi-definite solution defined over the entire interval $(-\infty, T]$ which satisfies $\underline{K}(T) = \underline{F}$.

APPENDIX B

PROOF OF THEOREM 6

The linear, time-invariant system Σ is characterized by the equations

$$\Sigma: \quad \begin{aligned} \dot{\underline{x}}(t) &= \underline{A}\underline{x}(t) + \underline{B}\underline{u}(t) \\ \underline{y}(t) &= \underline{C}\,\underline{x}(t) \end{aligned}$$

and we wish to prove

Theorem 6: If Σ is completely controllable and completely observable then

(a) the algebraic equation

$$\underline{0} = \underline{K}\,\underline{A} + \underline{A}'\underline{K} + \underline{C}'\underline{C} - \underline{K}\underline{B}\,\underline{B}'\underline{K} \tag{B.1}$$

cannot possess a positive semi-definite solution, but may possess a positive definite solution.

(b) $\bar{\underline{K}} = \lim_{T \to \infty} \underline{K}(t; T, \underline{0})$ is the unique positive definite solution of Eq. (B.1).

(c) The optimal closed-loop system

$$\dot{\underline{x}}(t) = (\underline{A} - \underline{B}\underline{B}'\bar{\underline{K}})\,\underline{x}(t) \tag{B.2}$$

is asymptotically stable, i.e. $\mathrm{Re}\,\lambda_i(\underline{A} - \underline{B}\underline{B}'\bar{\underline{K}}) < 0$ and

$$V(\underline{x}) = \frac{1}{2} < \underline{x}, \bar{\underline{K}}\,\underline{x} > \tag{B.3}$$

is a suitable Lyapunov function.

Proof: (a) Let $\underline{K}_1 \geq \underline{0}$ be a positive semi-definite solution of Eq. (B.1) then

$$\dot{\underline{x}}(t) = (\underline{A} - \underline{B}\underline{B}'\underline{K}_1)\,\underline{x}(t) \qquad (B.4)$$

is the equation of the closed-loop system for Σ with $\underline{u}(t) = -\underline{B}'\underline{K}_1\underline{x}(t)$.

Let us now consider the scalar function

$$V(\underline{x}) = \frac{1}{2} < \underline{x}, \underline{K}_1\underline{x} >$$

which is non-negative since $\underline{K}_1 \geq 0$. The rate of change of $V(\underline{x})$ along trajectories of (B.4) is given by

$$\dot{V}(\underline{x}) = \frac{1}{2} < \underline{x}, (\underline{A}_1'\underline{K}_1 + \underline{K}_1\underline{A}_1)\,\underline{x} >$$

$$= -\frac{1}{2} < \underline{x}, (\underline{C}'\underline{C} + \underline{K}_1\underline{B}\underline{B}'\underline{K}_1)\,\underline{x} > ,$$

the last step following since $\underline{K}_1$ satisfies Eq. (B.1). $\dot{V}(\underline{x})$ is always negative along a trajectory of (B.4) unless $\underline{x}(t) \equiv \underline{0}$. To see this we note that $\dot{V}(\underline{x}) \equiv 0$ implies $-\underline{B}'\underline{K}_1\,\underline{x}(t) \equiv \underline{0} = \underline{u}(t)$. Hence, $\dot{V}(\underline{x}) \equiv 0$ implies

$$< \underline{x}(t), \underline{C}'\underline{C}\underline{x}(t) > \equiv 0$$

or

$$< \underline{C}\,\underline{\Phi}(t,\tau)\underline{x}, \underline{C}\,\underline{\Phi}(t,\tau)\,\underline{x} > \equiv 0$$

for all initial states $\underline{x}$ at time τ. However, by complete observability, $\underline{C}\,\underline{\Phi}(t,\tau)\,\underline{x} \equiv \underline{0}$ if and only if $\underline{x} = \underline{0}$, establishing the fact that $\dot{V}(\underline{x}) < 0$ for all $\underline{x} \neq \underline{0}$.

But now, since $\underline{K}_1$ is only positive semi-definite, there exists a vector $\underline{\xi}$ such that $V(\underline{\xi}) = 0$. Let $\underline{\phi}(t;\tau,\underline{\xi})$, $t \geq \tau$, denote the trajectory of Eq. (B.4) satisfying the initial condition $\underline{\phi}(\tau;\tau,\underline{\xi}) = \underline{\xi}$. Then since $\dot{V}(\cdot)$ is negative along any solution of Eq. (B.4), we conclude that

$$V(\underline{\phi}(t;\tau,\underline{\xi})) < 0 \quad \text{for } t > \tau$$

which contradicts the fact $\underline{K}_1 \geq \underline{0}$.

(b) $\bar{\underline{K}} = \lim_{T\to\infty} \underline{K}(t;T,\underline{0})$ is at least positive semi-definite and must satisfy Eq. (B.1). Hence, under the assumption of complete observability, the result of part (a) guarantees that $\bar{\underline{K}}$ is positive definite. Let us now suppose that there exists another matrix $\underline{K}_1 > \underline{0}$ which satisfies Eq. (B.1). Then the system

$$\dot{\underline{x}}(t) = (\underline{A} - \underline{B}\underline{B}'\underline{K}_1)\underline{x}(t)$$

is asymptotically stable (since the pair $\{\underline{A},\underline{C}\}$ is completely observable) and

$$V_1(\underline{x}) = \frac{1}{2} < \underline{x}, \underline{K}_1\underline{x} >$$

is a suitable Lyapunov function with

$$\dot{V}_1(\underline{x}) = -\frac{1}{2} < \underline{x}, \underline{C}'\underline{C}\underline{x} > - \frac{1}{2} < \underline{x}, \underline{K}\underline{B}\underline{B}'\underline{K}\underline{x} >$$

Therefore, $\text{Re}\ \lambda_i(\underline{A} - \underline{B}\underline{B}'\underline{K}_1) < 0$

If we now let $\delta\underline{K} = \bar{\underline{K}} - \underline{K}_1$ we find that $\delta\underline{K}$ satisfies the equation

$$\underline{0} = \delta\underline{K}[\underline{A} - \underline{B}\underline{B}'\bar{\underline{K}}] + [\underline{A} - \underline{B}\underline{B}'\underline{K}_1]'\delta\underline{K} \qquad (B.5)$$

But the matrix equation $\underline{X}\underline{A} + \underline{B}\underline{X} = \underline{0}$ has a unique solution, namely $\underline{X} = \underline{0}$, whenever $\lambda_i(\underline{A}) + \lambda_j(\underline{B}) \neq 0$ for all pairs i,j.[22] Now since both $(\underline{A} - \underline{B}\underline{B}'\bar{\underline{K}})$ and $(\underline{A} - \underline{B}\underline{B}'\bar{\underline{K}}_1)$ have eigenvalues with negative real parts, the required condition is satisfied and so $\underline{K}_1 = \bar{\underline{K}}$, showing that Eq. (B.1) has only one positive definite solution as asserted.

(c) This part follows immediately from Theorem 5, since complete controllability and observability in the constant case imply uniform c.c. and uniform c.o. respectively, and in fact the constant σ appearing in Definition 4 may be made arbitrarily small. ‖

APPENDIX C

COST MATRICES FOR LINEAR REGULATOR PROBLEMS

In this appendix we define the notion of a "cost matrix" for a linear regulator problem and derive several useful expressions for the difference between two cost matrices.

We deal with the linear system

$$\Sigma: \quad \begin{aligned} \dot{\underline{x}}(t) &= \underline{A}(t)\,\underline{x}(t) + \underline{B}(t)\,\underline{u}(t) \\ \underline{y}(t) &= \underline{C}(t)\,\underline{x}(t) \end{aligned}$$

and the quadratic cost functional

$$J(\underline{x}, t, T, \underline{u}(\cdot)) = \frac{1}{2}\langle \underline{x}(T), \underline{F}\underline{x}(T)\rangle + \frac{1}{2}\int_t^T \left[\langle \underline{y}(\tau), \underline{y}(\tau)\rangle + \langle \underline{u}(\tau), \underline{u}(\tau)\rangle\right] d\tau \tag{C.1}$$

where $T > t$ (we may have $T = \infty$ in which case we assume $\underline{F} = \underline{0}$).

Suppose $\underline{u}(t)$ is constrained to be a linear feedback control law of the form

$$\underline{u}(t) = -\underline{L}(t)\,\underline{x}(t) \tag{C.2}$$

where the rxn matrix $\underline{L}(\tau)$ is defined for all $\tau \in [t, T]$. We then propose

Definition: The cost matrix $\underline{V}(t)$ associated with $\underline{L}(t)$ is

$$\underline{V}(t) = \underline{\Phi}_L'(T, t)\underline{F}\underline{\Phi}_L(T, t) + \int_t^T \underline{\Phi}_L'(\tau, t)\left[\underline{C}'(\tau)\underline{C}(\tau) + \underline{L}'(\tau)\right]\underline{\Phi}_L(\tau, t)\,d\tau \tag{C.3}$$

where $\underline{\Phi}_L(\tau, t)$ is the transition matrix corresponding to $\underline{A}(\tau) - \underline{B}(\tau)\underline{L}(\tau)$, i.e., $\underline{\Phi}_L(\tau, t)$ satisfies

$$\frac{d}{d\tau}\underline{\Phi}_L(\tau, t) = \left[\underline{A}(\tau) - \underline{B}(\tau)\underline{L}(\tau)\right]\underline{\Phi}_L(\tau, t); \underline{\Phi}_L(t, t) = \underline{I} \qquad (C.4)$$

Note that $\underline{V}(t)$ has the major property that

$$J(\underline{x}, t, T, \underline{u}(\cdot))\Big|_{u = -\underline{L}(t)\underline{x}(t)} = \frac{1}{2} < \underline{x}, \underline{V}(t)\underline{x} > \qquad (C.5)$$

and, in addition, since this is non-negative for all $\underline{x} \in E_n$, the matrix $\underline{V}(t)$ is positive semi-definite. Furthermore, by differentiating both sides of Eq. C.3 with respect to t, noting that

$$\frac{d}{dt}\underline{\Phi}'_L(\tau, t) = -\left[\underline{A}(t) - \underline{B}(t)\underline{L}(t)\right]'\underline{\Phi}_L(\tau, t) \qquad (C.6)$$

we find that $\underline{V}(t)$ satisfies the linear differential equation

$$\dot{\underline{V}}(t) = -\left[\underline{A}(t) - \underline{B}(t)\underline{L}(t)\right]'\underline{V}(t) - \underline{V}(t)\left[\underline{A}(t) - \underline{B}(t)\underline{L}(t)\right] - \underline{C}'(t)\underline{C}(t) - \underline{L}'(t)\underline{L}(t) \qquad (C.7)$$

with the boundary condition

$$\underline{V}(T) = \underline{F} \qquad (C.8)$$

Suppose now that we have two control laws $\underline{u}_1 = -\underline{L}_1(t)\underline{x}(t)$ and $\underline{u}_2 = -\underline{L}_2(t)\underline{x}(t)$ and we wish to determine their relative merit with respect to the cost functional C.1. For a given $\underline{x} \in E_n$, the cost difference between using $\underline{u}_1$ and $\underline{u}_2$ is

$$J(\underline{x}, t, T, \underline{u}_1(\cdot)) - J(\underline{x}, t, T, \underline{u}_2(\cdot)) = \frac{1}{2} < \underline{x}, \left[\underline{V}_1(t) - \underline{V}_2(t)\right]\underline{x} > \qquad (C.9)$$

where $\underline{V}_1(t)$ and $\underline{V}_2(t)$ are the cost matrices associated with $\underline{L}_1(t)$ and $\underline{L}_2(t)$ respectively. Therefore, studying the cost

difference between control laws is tantamount to studying the difference in their associated cost matrices. We shall now derive several useful relationships for $\underline{V}_1(t) - \underline{V}_2(t)$. We use the notation $\underline{A}_i(t) = \underline{A}(t) - \underline{B}(t)\underline{L}_i(t)$ for $i = 1, 2$ and we let $\underline{\Phi}_i(\tau, t)$ denote the transition matrix corresponding to $\underline{A}_i(t)$.

The matrix $\underline{V}_1(t)$ satisfies the differential equation

$$\dot{\underline{V}}_1(t) = -\underline{A}_1'(t)\underline{V}_1(t) - \underline{V}_1(t)\underline{A}_1(t) - \underline{C}'(t)\underline{C}(t) - \underline{L}'_1(t)\underline{L}_1(t) \qquad (C.10)$$

with $\underline{V}_1(T) = \underline{F}$. Writing $\underline{A}_1(t) = \underline{A}_2(t) - \underline{B}(t)[\underline{L}_1(t) - \underline{L}_2(t)]$ yields

$$\dot{\underline{V}}_1(t) = -\underline{A}_2'(t)\underline{V}_1(t) - \underline{V}_1(t)\underline{A}_2(t) - \underline{C}'(t)\underline{C}(t) - \underline{L}_1'(t)\underline{L}_2(t) + [\underline{L}_1(t) - \underline{L}_2(t)]'\underline{B}(t)\underline{V}_1(t) + \underline{V}_1(t)\underline{B}(t)[\underline{L}_1(t) - \underline{L}_2(t)] \qquad (C.11)$$

But on the other hand $\underline{V}_2(t)$ satisfies

$$\dot{\underline{V}}_2(t) = -\underline{A}_2'(t)\underline{V}_2(t) - \underline{V}_2(t)\underline{A}_2(t) - \underline{C}'(t)\underline{C}(t) - \underline{L}_2'(t)\underline{L}_2(t) \qquad (C.12)$$

with $V_2(T) = \underline{F}$. Consequently, subtracting Eq. C.12 from C.11 yields, writing $\delta\underline{V}(t) = \underline{V}_1(t) - \underline{V}_2(t)$,

$$\delta\dot{\underline{V}}(t) = -\underline{A}_2'(t)\delta\underline{V}(t) - \delta\underline{V}(t)\underline{A}_2(t) + \underline{L}_2'(t)\underline{L}_2(t) - \underline{L}_1'(t)L_1(t) + [\underline{L}_1(t) - \underline{L}_2(t)]'\underline{B}'(t)\underline{V}_1(t) + \underline{V}_1(t)\underline{B}(t)[\underline{L}_1(t) - \underline{L}_2(t)] \qquad (C.13)$$

with $\delta\underline{V}(T) = \underline{0}$. We now add and subtract the term $(\underline{L}_1 - \underline{L}_2)'\underline{L}_2 + \underline{L}_2'(\underline{L}_1 - \underline{L}_2) = \underline{L}_1'\underline{L}_2 + \underline{L}_2'\underline{L}_1 - 2\underline{L}_2'\underline{L}_2$ from the right hand side of Eq. C.13 to obtain

$$\delta\dot{\underline{V}}(t) = -\underline{A}_2'(t)\delta\underline{V}(t) - \delta\underline{V}(t)\underline{A}_2(t) - [\underline{L}_1(t) - \underline{L}_2(t)]'\cdot[\underline{L}_1(t) - \underline{L}_2(t)]$$

$$+[\underline{L}_1(t) - \underline{L}_2(t)]'\cdot[\underline{B}'(t)\underline{V}_1(t) - \underline{L}_2(t)] + [\underline{B}'(t)\underline{V}_1(t) - \underline{L}_2(t)]'\cdot[\underline{L}_1(t) - \underline{L}_2(t)] \quad \text{(C. 14)}$$

The solution of Eq. C. 14 with the boundary condition $\delta\underline{V}(T) = \underline{0}$ is given by

$$\delta\underline{V}(t) = \int_t^T \underline{\Phi}_2'(\tau, t)[(\underline{L}_1 - \underline{L}_2)'(\underline{L}_1 - \underline{L}_2) - (\underline{L}_1 - \underline{L}_2)'(\underline{B}'\underline{V}_1 - \underline{L}_2)$$

$$-(\underline{B}'\underline{V}_1 - \underline{L}_2)'(\underline{L}_1 - \underline{L}_2)]\underline{\Phi}_2(\tau, t)d\tau \quad \text{(C. 15)}$$

where, for ease of notation in the integrand, $\underline{L}_1 \equiv L_1(\tau)$, etc.

By interchanging subscripts in Eq. C. 15 we obtain an expression for $\underline{V}_2(t) - \underline{V}_1(t)$. Multiplying this newly found expression by -1 yields another formula for $\delta\underline{V}(t) = \underline{V}_1(t) - \underline{V}_2(t)$, namely

$$\delta\underline{V}(t) = \int_t^T \underline{\Phi}_1'(\tau, t)[-(\underline{L}_1 - \underline{L}_2)'(\underline{L}_1 - \underline{L}_2) - (\underline{L}_1 - \underline{L}_2)'(\underline{B}\underline{V}_2 - \underline{L}_1)$$

$$-(\underline{B}'\underline{V}_2 - \underline{L}_1)'(\underline{L}_1 - \underline{L}_2)]\underline{\Phi}_1(\tau, t)\, d\tau \quad \text{(C. 16)}$$

and substituting

$$(\underline{B}'\underline{V}_2 - \underline{L}_1) = (\underline{B}'\underline{V}_2 - \underline{L}_2) - (\underline{L}_1 - \underline{L}_2)$$

into the second and third terms of the integrand results in

$$\delta\underline{V}(t) = \int_t^T \underline{\Phi}_1'(\tau, t)[(\underline{L}_1 - \underline{L}_2)'(\underline{L}_1 - \underline{L}_2) - (\underline{L}_1 - \underline{L}_2)'(\underline{B}'\underline{V}_2 - \underline{L}_2)$$

$$-(\underline{B}'\underline{V}_2 - \underline{L}_2)'(\underline{L}_1 - \underline{L}_2)]\underline{\Phi}_1(\tau, t)\, d\tau \quad \text{(C. 17)}$$

There are various other expressions which one can obtain for $\delta\underline{V}(t)$, either from Eqs. C.15 - C.17 or from the differential equations satisfied by $\underline{V}_1(t)$ and $\underline{V}_2(t)$. We list two more of these below for reference

$$\delta\underline{V}(t) = \int_t^T \underline{\Phi}_2'(\tau, t)\left[(\underline{L}_1 - \underline{B}'\underline{V}_1)'(\underline{L}_1 - \underline{B}'\underline{V}_1) - (\underline{L}_2 - \underline{B}'\underline{V}_1)'(\underline{L}_2 - \underline{B}'\underline{V}_1)\right]\underline{\Phi}_2(\tau, t)d\tau \tag{C.18}$$

$$\delta\underline{V}(t) = \int_t^T \underline{\Phi}_2'(\tau, t)\left[(\underline{L}_1 - \underline{L}_2)'(\underline{L}_1 - \underline{V}_1) + (\underline{L}_2 - \underline{V}_2)'(\underline{L}_1 - \underline{L}_2)\right]\underline{\Phi}_1(\tau, t)d\tau \tag{C.19}$$

Finally, we note that equations C.15 - C.19 are valid for all T, including the case $T = \infty$. However, in the latter case these expressions become meaningless if $\underline{L}_1$ and $\underline{L}_2$ are such that both $\underline{V}_1(t)$ and $\underline{V}_2(t)$ are unbounded,† an event which is impossible for finite T. Therefore, care must be exercised when using the above formulae for $T = \infty$.

---

† Since in this case we have the indeterminate form $(\infty - \infty)$.

APPENDIX D

PROOF OF THEOREM 8 AND COROLLARY

In this appendix we prove the monotone convergence of the sequence of successive approximations to $\underline{K}(t; T, \underline{F})$ as proposed in Theorem 8. For convenience this theorem is repeated as

> Theorem: Let $\underline{V}_{n+1}(t)$, $n = 0, 1, \ldots,$ be the cost matrix associated with $\underline{L}_{n+1}(t)$ where $\underline{L}_{n+1}$ is recursively determined by
>
> $$\underline{L}_{n+1}(t) = \underline{B}'(t)\underline{V}_n(t) \qquad (D.1)$$
>
> and where $\underline{L}_o(t)$ is arbitrary, with associated cost matrix $\underline{V}_o(t)$. Then
>
> (a) $\underline{K}(t; T, \underline{F}) \leq \underline{V}_{n+1}(t) \leq \underline{V}_n(t)$ for $n = 0, 1, \ldots$
>
> (b) $\lim_{n \to \infty} \underline{V}_n(t) = \underline{V}_\infty(t)$ exists
>
> (c) $\underline{V}_\infty(t) = \underline{K}(t; T, \underline{F})$

Proof: (a) To show that $\underline{V}_{n+1}(t) \leq \underline{V}_n(t)$ we apply Eq. (C.15) to our problem. Rewriting this equation with $\underline{L}_1(t) = \underline{L}_n(t)$ and $\underline{L}_2(t) = \underline{L}_{n+1}(t)$ yields

$$\underline{V}_n(t) - \underline{V}_{n+1}(t) = \int_t^T \underline{\Phi}'_{n+1}(\tau, t)\left[(\underline{L}_n - \underline{L}_{n+1})'(\underline{L}_n - \underline{L}_{n+1}) - (\underline{L}_n - \underline{L}_{n+1})'(\underline{B}'\underline{V}_n - \underline{L}_{n+1})\right.$$

$$\left. - (\underline{B}'\underline{V}_n - \underline{L}_{n+1})'(\underline{L}_n - \underline{L}_{n+1})\right]\underline{\Phi}_{n+1}(\tau, t)d\tau \qquad (D.2)$$

where $\underline{\Phi}_{n+1}(t, t_o)$ is the transition matrix associated with $\underline{A}_{n+1}(t) \stackrel{\Delta}{=} \underline{A}(t) - \underline{B}(t)\underline{L}_{n+1}(t) = \underline{A}(t) - \underline{B}(t)\underline{B}'(t)\underline{V}_n(t)$ for $n = 0, 1, \ldots$

From Eq. (D.2) we obtain by substitution, since $\underline{L}_{n+1}(t) = \underline{B}'(t)\underline{V}_n(t)$, that

$$\underline{V}_n(t)-\underline{V}_{n+1}(t) = \int_t^T \underline{\Phi}'_{n+1}(\tau, t)\left[\underline{L}_n(\tau)-\underline{L}_{n+1}(\tau)\right]'\left[\underline{L}_n(\tau)-\underline{L}_{n+1}(\tau)\right]\underline{\Phi}_{n+1}(\tau, t)d\tau \tag{D.3}$$

Consequently, we see that

$$\underline{V}_n(t) \geq \underline{V}_{n+1}(t)$$

and $\underline{V}_n(t) \equiv \underline{V}_{n+1}(t)$ if and only if $\underline{L}_n(t) \equiv \underline{L}_{n+1}(t)$. Finally, the fact that $\underline{K}(t; T, \underline{F}) \leq \underline{V}_{n+1}(t)$ follows immediately from Lemma 6 with $\underline{L}(t)$ taken to be $\underline{B}'(t)\underline{V}_n(t) = \underline{L}_{n+1}(t)$.

(b) To show that the sequence $\underline{V}_n(t)$ of positive-semi definite matrices has a limit as $n \to \infty$, we choose an arbitrary $\underline{x} \in E_n$. Then, by part (a), $<\underline{x}, \underline{V}_n(t)\underline{x}>$ is non-increasing as $n \to \infty$ and is uniformly bounded below by $<\underline{x}, \underline{K}(t; T, F)\underline{x}>$. Therefore

$$\lim_{n \to \infty} <\underline{x}, \underline{V}_n(t)\, \underline{x}>$$

exists for all $\underline{x} \in E_n$ since a bounded monotone sequence always has a limit. This implies that $\lim_{n \to \infty} \underline{V}_n(t)$ exists. To see this, first consider $\underline{x} = \underline{e}_i$ ($\underline{e}_i$ = i-th element of the standard set of basis vectors). Then $<\underline{x}, \underline{V}_n(t)\, \underline{x}> = (\underline{V}_n)_{ii}$. Letting $n \to \infty$, since $\lim <\underline{x}, \underline{V}_n(t)\underline{x}>$ exists, this implies that

$$\lim_{n \to \infty} (\underline{V}_n)_{ii}$$

exists for all i. Next, take $\underline{x} = \underline{e}_i + \underline{e}_j$ to show that the limit as $n \to \infty$ of the off diagonal terms of $\underline{V}_n(t)$ exist. Hence, we conclude that $\lim \underline{V}_n(t)$ exists as $n \to \infty$. Call this limit $\underline{V}_\infty(t)$, viz,

$$\lim_{n \to \infty} \underline{V}_n(t) = \underline{V}_\infty(t) \tag{D.4}$$

(c) Having proved the existence of $\underline{V}_\infty(t)$ for $t < T$, we wish to show that $\underline{V}_\infty(t)$ satisfies the Riccati equation and that $\underline{V}_\infty(t) = \underline{K}(t; T, \underline{F})$. Since $\underline{V}_{n+1}(t)$ is the cost matrix associated with $\underline{L}_{n+1}(t)$, $\underline{V}_{n+1}(t)$ satisfies the equation

$$\dot{\underline{V}}_{n+1}(t) = -\underline{V}_{n+1}(t)\left[\underline{A}(t)-\underline{B}(t)\underline{L}_{n+1}(t)\right]-\left[\underline{A}(t)-\underline{B}(t)\underline{L}_{n+1}(t)\right]'\underline{V}_{n+1}(t)$$

$$-\underline{C}'(t)\underline{C}(t) - \underline{L}'_{n+1}(t)\underline{L}_{n+1}(t)$$

or $$\dot{\underline{V}}_{n+1}(t) = -\underline{V}_{n+1}(t)\underline{A}_{n+1}(t)-\underline{A}'_{n+1}(t)\underline{V}_{n+1}(t)-\underline{C}'(t)\underline{C}(t)-\underline{V}_n(t)\underline{B}(t)\underline{B}'(t)\underline{V}_n(t) \tag{D.5}$$

Integrating both sides of Eq. (D.5) from $\tau = t$ to $\tau = T$ we obtain, since $\underline{V}_{n+1}(T) = \underline{F}$, that

$$\underline{V}_{n+1}(t)-\underline{F} = \int_t^T \left[\underline{V}_{n+1}(\tau)\underline{A}_{n+1}(\tau)+\underline{A}'_{n+1}(\tau)\underline{V}_{n+1}(\tau)+\underline{C}'(\tau)\underline{C}(\tau)+\underline{V}_n(\tau)\underline{B}(\tau)\underline{B}'(\tau)\underline{V}_n(\tau)\right] d\tau$$

Taking the limit of both sides of this expression we have (since $\underline{V}_n(\tau)$ is uniformly bounded we can take the limit under the integral sign by the bounded convergence theorem)

$$\underline{V}_\infty(t)-\underline{F} = \int_t^T \left[\underline{V}_\infty(\tau)\underline{A}(\tau)+\underline{A}'(\tau)\underline{V}_\infty(\tau)+\underline{C}'(\tau)\underline{C}(\tau)-\underline{V}_\infty(\tau)\underline{B}(\tau)\underline{B}'(\tau)\underline{V}_\infty(\tau)\right] d\tau$$

Hence, $\underline{V}_\infty(t)$, as an integral is continuous. Differentiating yields:

$$\dot{\underline{V}}_\infty(t) = -\underline{V}_\infty(t)\underline{A}(t) - \underline{A}'(t)\underline{V}_\infty(t) - \underline{C}'(t)\underline{C}(t) + \underline{V}_\infty(t)\underline{B}(t)\underline{B}(t)\underline{V}_\infty(t)$$

with $\underline{V}_\infty(T) = \underline{F}$. Hence $\underline{V}_\infty(t)$ satisfies the Riccati equation and, by uniqueness, we conclude

$$\underline{V}_\infty(t) = \underline{K}(t; T, \underline{F})$$

which completes the proof of the theorem. ‖

In the special case when Σ is stationary and $T = \infty$, the Riccati equation solution is $\underline{K}(t) = \bar{\underline{K}}$ = constant and the above theorem may be slightly extended to give a monotonic convergence scheme for determining $\bar{\underline{K}}$. We prove this fact as a corollary; the method of proof is very similar to that used by Wonham[30] to obtain a comparable result for the discrete-time regulator problem.

Corollary 1: Let the time-invariant system Σ be completely controllable and completely observable. Let $\underline{V}_n$, $n = 0, 1, \ldots$, be the (unique) positive definite solution of the linear algebraic equation

$$\underline{0} = \underline{A}_n'\underline{V}_n + \underline{V}_n\underline{A}_n + \underline{C}'\underline{C} + \underline{L}_n'\underline{L}_n \qquad (D.6)$$

where, recursively,

$$\underline{L}_n = \underline{B}'\underline{V}_{n-1} \quad \text{for } n = 1, 2, \ldots$$

$$\underline{A}_n = \underline{A} - \underline{B}\underline{L}_n$$

and where $\underline{L}_0$ is chosen such that the matrix $\underline{A}_0 = \underline{A} - \underline{B}\underline{L}_0$ has eigenvalues with negative real parts. Then,

(a) $\bar{\underline{K}} \leq \underline{V}_{n+1} \leq \underline{V}_n \leq \ldots$ for $n = 0, 1, \ldots$

(b) $\lim_{n \to \infty} \underline{V}_n = \bar{\underline{K}}$

Proof: (a) The cost matrix $\underline{V}_o(t)$, associated with $\underline{L}_o$ is given by

$$\underline{V}_o(t) = \int_t^\infty e^{\underline{A}_o'(t-\tau)} (\underline{C}'\underline{C} + \underline{L}_o'\underline{L}_o) e^{\underline{A}_o(t-\tau)} d\tau \qquad (D.7)$$

Since Re $\lambda_i(\underline{A}_o) < 0$, $\|\underline{V}_o(t)\| < \infty$ and so $\underline{V}_o(t)$ satisfies the equation

$$\dot{\underline{V}}_o(t) = -\underline{A}_o' \underline{V}_o(t) - \underline{V}_o(t) \underline{A}_o - \underline{C}'\underline{C} - \underline{L}_o'\underline{L}_o \qquad (D.8)$$

But $\underline{V}_o(t)$ is a constant matrix independent of t. To show this we simply make a change of variable in Eq. (D.8) for $\underline{V}_o(t)$. Letting $\xi = t_1 - t + \tau$, where t_1 is arbitrary, yields

$$\underline{V}_o(t) = \int_{t_1}^\infty e^{\underline{A}_o'(t_1-\xi)} (\underline{C}'\underline{C}+\underline{L}_o'\underline{L}_o) e^{\underline{A}_o'(t_1-\xi)} d\xi = \underline{V}_o(t_1)$$

Hence, $\underline{V}_o(t) = \underline{V}_o$ must be the unique† solution of the algebraic equation

$$\underline{0} = \underline{A}_o'\underline{V}_o + \underline{V}_o\underline{A}_o + \underline{C}'\underline{C} + \underline{L}_o'\underline{L}_o$$

We now let $\underline{L}_1 = \underline{B}'\underline{V}_o$ and $\underline{V}_1(t)$ be its associated cost matrix. Then, by using Eq. (C.15) with $T = \infty$, we obtain

$$\underline{V}_o - \underline{V}_1(t) = \int_t^\infty e^{\underline{A}_1'(\tau-t)} (\underline{L}_o-\underline{L}_1)'(\underline{L}_o-\underline{L}_1) e^{\underline{A}_1(\tau-t)} d\tau \qquad (D.9)$$

so that $\underline{V}_1(t) \leq \underline{V}_o$. Consequently, $\underline{V}_1(t)$ is bounded for all t and by the same reasoning used above for $\underline{V}_o$ we find that $\underline{V}_1(t) = \underline{V}_1 =$ constant and satisfies

$$\underline{0} = \underline{A}_1'\underline{V}_1 + \underline{V}_1\underline{A}_1 + \underline{C}'\underline{C} + \underline{L}_1\underline{L}_1 \qquad (D.10)$$

---

† Uniqueness is guaranteed since $\lambda_i(\underline{A}_o) + \lambda_i(\underline{A}_o) \neq 0$ for all pairs i, j.

Furthermore, since Σ is completely observable, $\underline{V}_1$ is finite if and only if $\text{Re}\,\lambda_i(\underline{A}_1) < 0$ for all i; hence $\underline{V}_1$ must be the unique solution of Eq. (D.10).

Recursively, we have that if $\underline{L}_n = \underline{B}'\underline{V}_{n-1}$ for $n = 1, 2, \ldots$ then the cost matrix $\underline{V}_n$ associated with $\underline{L}_n$ is the unique (positive definite)† solution of

$$\underline{0} = \underline{A}_n'\underline{V}_n + \underline{V}_n\underline{A}_n + \underline{C}'\underline{C} + \underline{L}_n'\underline{L}_n \qquad \text{(D.11)}$$

and

$$\underline{V}_{n+1} \leq \underline{V}_n \quad \text{for} \quad n = 0, 1, \ldots$$

Finally, $\overline{\underline{K}} \leq \underline{V}_n$ for all n since $\overline{\underline{K}}$ is associated with the optimal control.

(b) $\lim_{n \to \infty} \underline{V}_n = \underline{V}_\infty$ exists by the same method of proof used in part b of the above theorem. To show that $\underline{V}_\infty = \overline{\underline{K}}$ we substitute $\underline{A}_n = \underline{A} - \underline{B}\underline{B}'\underline{V}_{n-1}$ into Eq. (D.11) to obtain

$$\underline{0} = \underline{A}'\underline{V}_n + \underline{V}_n\underline{A} + \underline{C}'\underline{C} - \underline{V}_{n-1}\underline{B}\underline{B}'\underline{V}_n - \underline{V}_n\underline{B}\underline{B}'\underline{V}_{n-1} + \underline{V}_{n-1}\underline{B}\underline{B}'\underline{V}_{n-1}$$

The right hand side of this equation is uniformly bounded in n, so that taking the limit as $n \to \infty$ yields

$$\underline{0} = \underline{A}'\underline{V}_\infty + \underline{V}_\infty\underline{A} + \underline{C}'\underline{C} - \underline{V}_\infty\underline{B}\underline{B}'\underline{V}_\infty$$

$\underline{V}_\infty$ is positive definite and therefore by Theorem 6, $\underline{V}_\infty = \overline{\underline{K}}$ which concludes the proof. ‖

---

† Positive definiteness is assured if Σ = completely observable.

APPENDIX E

ON THE RELATIONSHIP BETWEEN NEWTON'S METHOD AND THE METHOD OF SUCCESSIVE APPROXIMATIONS TO DETERMINE $\underline{K}(t; T, \underline{F})$

In this section we shall present a non-rigorous discussion of the application of Newton's method to the solution of the Riccati equation. We shall show that for this problem, the iterative scheme of Newton's method is precisely equivalent to the method of successive approximations as discussed in Section III.D. For a detailed treatment of the application of Newton's method in function spaces see Reference 18. For further details of the relationships that exist between this method and Kalaba's method of successive approximations see Reference 17. For a more mathematically abstract discussion see Ref. 32.

Newton's method may be motivated as follows. Let $f(\cdot)$ be a continuously differentiable mapping of a Banach space D into another Banach space R, i.e.,

$$f(\cdot): \; D \rightarrow R \tag{E.1}$$

Let $f(\cdot)$ have a zero in D, i.e., there exists an element $\hat{x} \in D$ such that

$$f(\hat{x}) = 0 \in R \tag{E.2}$$

We now take any element $x_o \in D$. Under the assumption that the mapping $f(\cdot)$ is continuously differentiable in D, the element

$$f(x_o) = f(x_o) - f(\hat{x}) \tag{E.3}$$

which belongs to R, can be replaced by the approximation

$$f'_{x_o}(x_o - \hat{x}) \approx f(x_o) \tag{E.4}$$

where $f'_x(y)$ is the Frechet differential of $f(\cdot)$, evaluated at x, operating on y. It is a linear mapping from D into R and is defined

by (α = scalar),

$$f'_x(y) = \lim_{\alpha \to 0} \frac{f(x+\alpha y) - f(x)}{\alpha} \qquad (E.5)$$

Equation (E.4) provides a basis for assuming that the solution of the equation

$$f'_{x_o}(x_o - x) = f(x_o) \qquad (E.6)$$

will be close to $\hat{x}$. This last equation is linear in x so that its solution is

$$x_1 = x_o - \left[f'_{x_o}\right]^{-1}(f(x_o)) \qquad (E.7)$$

assuming the existence of the inverse mapping $\left[f'_{x_o}\right]^{-1}(\cdot)$.

Continuing the above process, we obtain after starting from the initial approximation x_o, the sequence $\{x_n\}$:

$$x_{n+1} = x_n - \left[f'_{x_n}\right]^{-1}(f(x_n)) \qquad n = 0, 1, .. \qquad (E.8)$$

Each x_n is an approximate solution of the equation $f(x) = 0$ and in general becomes more accurate with increasing n. The process of forming the sequence $\{x_n\}$ is known as "Newton's method".†

Equation (E.8) is an explicit equation for the (n+1)-st iterate x_{n+1} in terms of x_n. We can obtain an implicit relation for x_{n+1} by operating on both sides of (E.8) with $f'_{x_n}(\cdot)$. This results in

$$f'_{x_n}(x_n - x_{n+1}) = f(x_n) \qquad (E.9)$$

which is often easier to use than Eq. (E.8) because of the difficulty in obtaining the inverse mapping to $f'_{x_n}(\cdot)$.

In order to apply Newton's method, as outlined above, to solve the Riccati equation, we write the Riccati equation in the form

---

† If $f(\cdot)$ is a scalar valued function of a scalar argument Eq. (E.8) can be written as $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$ which is Newton's method in its ordinary form.

$$\underline{f}(\underline{V}) = \dot{\underline{V}} + \underline{A}'(t)\underline{V} + \underline{V}\underline{A}(t) - \underline{V}\underline{B}(t)\underline{B}'(t)\underline{V} + \underline{C}'(t)\underline{C}(t) \qquad \text{(E.10)}$$

where we take the Banach space D as the space of nxn matrix valued functions which are absolutely continuous on the interval $[t_o, T]$. We seek a solution of $\underline{f}(\underline{V}) = \underline{0}$ satisfying $\underline{V}(T) = \underline{F}$. We know that $\hat{\underline{V}} = \underline{K}(t; T, \underline{F})$ is in fact the solution and we shall apply Eq. (E.9) to form a sequence of iterates $\underline{V}_n \to \underline{K}$.

We first compute

$$\underline{f}'_{\underline{V}_n}(\underline{V}_n - \underline{V}_{n+1}) = \lim_{a \to 0} \frac{\underline{f}[\underline{V}_n + a(\underline{V}_n - \underline{V}_{n+1})] - \underline{f}(\underline{V}_n)}{a} \qquad \text{(E.11)}$$

Substituting (E.10) into (E.11) yields

$$\underline{f}'_{\underline{V}_n}(\underline{V}_n - \underline{V}_{n+1}) = \dot{\underline{V}}_n - \dot{\underline{V}}_{n+1} + (\underline{V}_n - \underline{V}_{n+1})\underline{A}(t) + \underline{A}'(t)(\underline{V}_n - \underline{V}_{n+1}) - (\underline{V}_n - \underline{V}_{n+1})\underline{B}(t)\underline{B}'(t)\underline{V}_n - \underline{V}_n\underline{B}(t)\underline{B}'(t)(\underline{V}_n - \underline{V}_{n+1}) \qquad \text{(E.12)}$$

Newton's method then sets

$$\underline{f}'_{\underline{V}_n}(\underline{V}_n - \underline{V}_{n+1}) = \underline{f}(\underline{V}_n) \qquad \text{(E.13)}$$

from which we obtain

$$-\dot{\underline{V}}_{n+1} + (\underline{V}_n - \underline{V}_{n+1})\underline{A}(t) + \underline{A}'(t)(\underline{V}_n - \underline{V}_{n+1}) - (\underline{V}_n - \underline{V}_{n+1})\underline{B}(t)\underline{B}'(t)\underline{V}_n - \underline{V}_n\underline{B}(t)\underline{B}'(t)(\underline{V}_n - \underline{V}_{n+1})$$

$$= \underline{A}'(t)\underline{V}_n + \underline{V}_n\underline{A}(t) - \underline{V}_n\underline{B}(t)\underline{B}'(t)\underline{V}_n + \underline{C}'(t)\underline{C}(t) \qquad \text{(E.14)}$$

Hence, the (n+1)-st iterate $\underline{V}_{n+1}$ satisfies the linear differential equation

$$\dot{\underline{V}}_{n+1}(t) = -\underline{V}_{n+1}(t)[\underline{A}(t) - \underline{B}(t)\underline{B}'(t)\underline{V}_n(t)] - [\underline{A}(t) - \underline{B}(t)\underline{B}'(t)\underline{V}_n(t)]'\underline{V}_{n+1}(t) - \underline{V}_n(t)\underline{B}(t)\underline{B}'(t)\underline{V}_n(t) - \underline{C}'(t)\underline{C}(t) \qquad \text{(E.15)}$$

with the boundary condition $\underline{V}_{n+1}(T)$ which we are free to choose. Note that the first approximation $\underline{V}_o(t)$ is arbitrarily chosen.

If we use the notation $\underline{L}_{n+1}(t) = \underline{B}'(t)\underline{V}_n(t)$ we see that (E. 15) expresses the exact same iterative scheme as does Theorem 8, since $\underline{V}_{n+1}(t)$ satisfying Eq. (E. 15) with $\underline{V}_{n+1}(T) = \underline{F}$ is the cost matrix associated with $\underline{L}_{n+1}(t)$.

Consequently, we have shown that the application of Newton's method to recursively determine $\underline{K}(t; T, \underline{F})$ gives exactly the same rule for forming the sequence of iterates $\{\underline{V}_n(t)\}$, $n = 1, 2, \ldots$ as does Theorem 8.

APPENDIX F

GRADIENT MATRICES OF TRACE FUNCTIONS

In this section we define what we mean by the gradient matrix of a trace function and we describe a procedure by which the gradient matrix may be obtained. Several examples are included which elucidate this method. The results presented in this appendix are an extension of those found in Reference 28, Section 8, to arbitrary trace functions.

Definition 1: Let $\underline{X}$ be an rxn matrix with elements x_{ij}, $i = 1, \ldots, r$; $j = 1, \ldots, n$. Let $f(\cdot)$ be a scalar, real-valued function of the x_{ij}, i.e.,

$$f(\underline{X}) = f(x_{11}, \ldots, x_{in}, x_{21}, \ldots, x_{2n}, \ldots) \qquad (F.1)$$

We then define the gradient matrix of $f(\underline{X})$ with respect to $\underline{X}$ as the rxn matrix $\frac{\partial f(\underline{X})}{\partial \underline{X}}$ whose ij-th element is given by

$$\frac{\partial f(\underline{X})}{\partial x_{ij}} \quad \text{for } i = 1, \ldots, r;\ j = 1, \ldots, n \qquad (F.2)$$

As an example, suppose that $\underline{X}$ is a 2x2 matrix and that $f(\underline{X}) = x_{11}^2 x_{21} + x_{21}^3 - x_{11} x_{22} x_{12} + 5x_{21}$, then

$$\frac{\partial f(\underline{X})}{\partial \underline{X}} = \begin{bmatrix} 2x_{11} x_{21} - x_{22} x_{12} & - x_{11} x_{22} \\ x_{11}^2 + 3x_{21}^2 + 5 & - x_{11} x_{12} \end{bmatrix}$$

We are interested in obtaining the gradient matrix of the trace of a matrix which depends on the matrix $\underline{X}$. We therefore define

Definition 2: $f(\cdot)$ is a trace function of the matrix $\underline{X}$ if $f(\underline{X})$ is of the form

$$f(\underline{X}) = \mathrm{tr}\left[\underline{F}(\underline{X})\right] \tag{F.3}$$

where $\underline{F}(\cdot)$ is a continuously differentiable mapping from the space of rxn matrices into the space of nxn matrices.

$\underline{F}(\underline{X})$ is a square matrix, so that its trace is well-defined. Typical examples of such functions are

(1) $\underline{F}(\underline{X}) = \underline{X}'\underline{X}$

(2) $\underline{F}(\underline{X}) = e^{\underline{A}+\underline{B}\underline{X}}$

(3) $\underline{F}(\underline{X}) = \underline{B}\underline{X}\,\underline{A}$

where in the above, $\underline{B}$ and $\underline{A}$ are nxr and nxn matrices respectively.

We now wish to indicate a procedure for determining $\dfrac{\partial f(\underline{X})}{\partial \underline{X}}$ when $f(\cdot)$ is a trace function. We shall not attempt to be mathematically rigorous; for a more precise discussion of matrix calculus (see Refs. 28 and 29).

We first note that the ij-th element of $\dfrac{\partial f(\underline{X})}{\partial \underline{X}}$ is $\dfrac{\partial f(\underline{X})}{\partial x_{ij}}$ which is defined by

$$\frac{\partial f(\underline{X})}{\partial x_{ij}} = \lim_{\epsilon \to 0} \frac{f(\underline{X} + \epsilon \Delta \underline{X}_{ij}) - f(\underline{X})}{\epsilon\, \delta x_{ij}} \tag{F.4}$$

where $\Delta \underline{X}_{ij}$ is an rxn matrix all of whose elements are zero except for its ij-th element which is given by δx_{ij}.

We now define the matrix differential $\Delta \underline{X}$ to be the matrix whose ij-th element is δx_{ij} (where the δx_{ij}'s are independent variations in

x_{ij}'s for $i = 1, \ldots, r, j = 1, \ldots, n$). We then have the following result concerning trace functions

Lemma: Let $f(\underline{X})$ be a trace function. Then if we can write

$$f(\underline{X} + \epsilon \Delta\underline{X}) - f(\underline{X}) = \epsilon \text{ tr}\left[\underline{M}(\underline{X})\Delta\underline{X}\right] \qquad (F.5)$$

as $\epsilon \to 0$, where $\underline{M}(\underline{X})$ is an $n \times r$ matrix, we have

$$\frac{\partial f(\underline{X})}{\partial \underline{X}} = \underline{M}'(\underline{X}) \qquad (F.6)$$

Proof: It is only necessary to show that Eq. (F.5) implies that $\frac{\partial f(\underline{X})}{\partial x_{ij}}$ is the ij-th element of $\underline{M}'(\underline{X})$. To see this we note that since the elements of $\Delta\underline{X}$ represent independent variations we can let $\Delta\underline{X} = \Delta\underline{X}_{ij}$. Then $\underline{M}(\underline{X})\Delta\underline{X} = \underline{M}(\underline{X})\Delta\underline{X}_{ij}$ is an $n \times n$ matrix all of whose elements are zero, except those in the j-th column. The j-th column of $\underline{M}(\underline{X})\Delta\underline{X}_{ij}$ is the i-th column of $\underline{M}(\underline{X})$ multiplied by δx_{ij}.

The trace of $\underline{M}(\underline{X})\Delta\underline{X}_{ij}$ is then the j-th element of the j-th row of this matrix which is given by

$$\text{tr } \underline{M}(\underline{X})\Delta\underline{X}_{ij} = m_{ji}(\underline{X}) \cdot \delta x_{ij}; \; j = 1, \ldots, n; \; i = 1, \ldots, r$$

Therefore, by Eq. (F.4)

$$\frac{\partial f(\underline{X})}{\partial x_{ij}} = m_{ji}(\underline{X})$$

But $m_{ji}(\underline{X})$ is simply the ij-th element of $\underline{M}'(\underline{X})$ and therefore by definition of the matrix $\frac{\partial f(\underline{X})}{\partial \underline{X}}$ we have

$$\frac{\partial f(\underline{X})}{\partial \underline{X}} = \underline{M}'(\underline{X})$$

as claimed. ‖

With the above lemma as an aid we outline a procedure for determining the gradient matrix of a trace function.

1. We are given the trace function $f(\underline{X}) = \text{tr}\, \underline{F}(\underline{X})$. Form

$$f(\underline{X} + \epsilon \Delta\underline{X}) - f(\underline{X}) = \text{tr}\left[\underline{F}(\underline{X}) + \epsilon \Delta\underline{X}) - \underline{F}(\underline{X})\right] \qquad (F.7)$$

2. Expand $F(\underline{X} + \epsilon \Delta\underline{X})$ for $\epsilon \to o$ as

$$\underline{F}(\underline{X} + \epsilon \Delta\underline{X}) = F(\underline{X}) + \epsilon \mathcal{L}(\Delta\underline{X}) \qquad (F.8)$$

where $\mathcal{L}(\Delta\underline{X})$ is a (continuous) linear mapping from the space of r x n matrices into the space of n x n matrices.

3. Using the properties

$$\text{tr}(\underline{Y}\,\underline{Z}) = \text{tr}(\underline{Z}\,\underline{Y})$$

$$\text{tr}(\underline{Y}) = \text{tr}(\underline{Y}')$$

write

$$\text{tr}\left[\underline{F}(\underline{X} + \epsilon\Delta\underline{X}) - \underline{F}(\underline{X})\right] = \epsilon\, \text{tr}\left[\mathcal{L}(\Delta\underline{X})\right] \qquad (F.9)$$

in the form

$$\text{tr}\left[\underline{F}(\underline{X} + \epsilon \Delta\underline{X}) - \underline{F}(\underline{X})\right] = \epsilon\, \text{tr}\left[\underline{M}(\underline{X})\, \Delta\underline{X}\right] \qquad (F.10)$$

where $\underline{M}(\underline{X})$ is an n x r matrix (whose elements are continuous in $\underline{X}$).

4. By lemma,

$$\frac{\partial f(\underline{X})}{\partial \underline{X}} = \underline{M}'(\underline{X})$$

The only questionable step in the above procedure is Step 3. It may or may not always be possible to manipulate $\text{tr}\,\mathcal{L}(\Delta\underline{X})$ into the form $\text{tr}\, \underline{M}(\underline{X})\Delta\underline{X}$. However, in all cases which have been investigated thus

far, this has been possible. We now illustrate the use of the above method for computing gradient matrices with a few examples.

Examples: In the following we assume that $\underline{X}$ is an r x n matrix, $\underline{C}$ an arbitrary r x r matrix, $\underline{B}$ an arbitrary n x r matrix and $\underline{A}$ an arbitrary n x n matrix. $\underline{A}$, $\underline{B}$ and $\underline{C}$ are independent of $\underline{X}$.

1. $\underline{F}(\underline{X}) = \underline{A}\underline{X}'\underline{C}\underline{X}$, so that

$$\underline{F}(\underline{X} + \epsilon\Delta\underline{X}) = \underline{A}\underline{X}'\underline{C}\underline{X} + \epsilon\left[\underline{A}(\Delta\underline{X})'\underline{C}\underline{X} + \underline{A}\underline{X}'\underline{C}(\Delta\underline{X})\right] + \epsilon^2\,\underline{A}(\Delta\underline{X})'\underline{C}(\Delta\underline{X})$$

Therefore as $\epsilon \to o$ we have

$$F(\underline{X} + \epsilon\Delta\underline{X}) = \underline{F}(\underline{X}) + \epsilon\,\mathcal{L}(\Delta\underline{X}) = \underline{F}(\underline{X}) + \epsilon\left[\underline{A}(\Delta\underline{X})'\underline{C}\underline{X} + \underline{A}\underline{X}'\underline{C}(\Delta\underline{X})\right]$$

But

$$\begin{aligned} \operatorname{tr}\mathcal{L}(\Delta\underline{X}) &= \operatorname{tr}\left[\underline{A}(\Delta\underline{X})'\underline{C}\underline{X} + \underline{A}\underline{X}'\underline{C}(\Delta\underline{X})\right] \\ &= \operatorname{tr}\left[(\Delta\underline{X})'\underline{C}\underline{X}\underline{A} + \underline{A}\underline{X}'\underline{C}(\Delta\underline{X})\right] \\ &= \operatorname{tr}\left[(\underline{A}'\underline{X}'\underline{C} + \underline{A}\underline{X}'\underline{C})(\Delta\underline{X})\right] \end{aligned}$$

Therefore, by the lemma,

$$\frac{\partial}{\partial\underline{X}}\operatorname{tr}(\underline{A}\underline{X}'\underline{C}\underline{X}) = \underline{C}\underline{X}\underline{A} + \underline{C}'\underline{X}\underline{A}'$$

2. Assume r = n so that $\underline{X}$ is a square matrix and let $\underline{F}(\underline{X}) = \underline{A}\underline{X}^{-1}\underline{B}$. Hence

$$\begin{aligned} \underline{F}(\underline{X} + \epsilon\Delta\underline{X}) &= \underline{A}(\underline{X} + \epsilon\Delta\underline{X})^{-1}\underline{B} \\ &= \underline{A}\left[I + \epsilon\underline{X}^{-1}(\Delta\underline{X})\right]^{-1}\underline{X}^{-1}\underline{B} \end{aligned}$$

But for $\epsilon \to 0$

$$[I + \epsilon \underline{X}^{-1}(\Delta X)]^{-1} = I - \epsilon \underline{X}^{-1}(\Delta \underline{X})$$

so that

$$\mathcal{L}(\Delta \underline{X}) = -\underline{A}\underline{X}^{-1}(\Delta \underline{X})\,\underline{X}^{-1}\underline{B}$$

and

$$\mathrm{tr}\,\mathcal{L}(\Delta \underline{X}) = -\mathrm{tr}\left[(\underline{X}^{-1}\underline{B}\underline{A}\underline{X}^{-1})(\Delta \underline{X})\right]$$

Consequently,

$$\frac{\partial}{\partial \underline{X}}\ \mathrm{tr}(\underline{A}\underline{X}^{-1}\underline{B}) = -(\underline{X}^{-1}\underline{B}\underline{A}\underline{X}^{-1})'$$

3. $\underline{F}(\underline{X}) = e^{(\underline{A}+\underline{B}\underline{X})}$, so that

$$\underline{F}(\underline{X} + \epsilon\,\Delta \underline{X}) = e^{\underline{A}+\underline{B}\underline{X} + \epsilon\,\underline{B}\,\Delta \underline{X}}$$

But from p 171, Ref. 22 we have that to first order in ϵ this is

$$F(\underline{X} + \epsilon\,\Delta \underline{X}) = e^{\underline{A}+\underline{B}\underline{X}} + \epsilon \int_0^1 e^{(\underline{A}+\underline{B}\underline{X})(1-\tau)}\underline{B}(\Delta \underline{X})e^{(\underline{A}+\underline{B}\underline{X})\tau}\,d\tau$$

Hence,

$$\mathcal{L}(\Delta \underline{X}) = \int_0^1 e^{(\underline{A}+\underline{B}\underline{X})(1-\tau)}\underline{B}(\Delta \underline{X})\, e^{(\underline{A}+\underline{B}\underline{X})\tau}\,d\tau$$

and so, since the trace operation commutes with integration, we obtain

$$\text{tr}\,\mathcal{L}(\Delta\underline{X}) = \text{tr}\int_0^1 e^{(\underline{A}+\underline{B}\underline{X})\tau}\, e^{(\underline{A}+\underline{B}\underline{X})(1-\tau)}\,d\tau \cdot \underline{B}(\Delta\underline{X})$$

$$= \text{tr}\; e^{(\underline{A}+\underline{B}\underline{X})}\underline{B}\cdot(\Delta\underline{X})$$

Therefore,

$$\frac{\partial}{\partial\underline{X}}\,\text{tr}\; e^{(\underline{A}+\underline{B}\underline{X})} = \underline{B}'e^{(\underline{A}+\underline{B}\underline{X})'}$$

4. $\underline{F}(\underline{X}) = \underline{A}e^{\underline{B}\underline{X}}$. We shall not go through the derivation of the gradient matrix but merely state the result which is

$$\frac{\partial}{\partial\underline{X}}\,\text{tr}\;\underline{A}e^{\underline{B}\underline{X}} = \underline{B}'e^{(\underline{B}\underline{X})'}\left[\int_0^1 e^{-(\underline{B}\underline{X})\tau}\,\underline{A}e^{(\underline{B}\underline{X})\tau}\,d\tau\right]'$$

It is of course possible to give many more examples of computing gradient matrices. However, all we wish to do is to give a flavor for the method outlined above. For further results see Ref. 28.

APPENDIX G

PROOF OF THEOREM 9

In this appendix we wish to prove Theorem 9 of Chapter IV which is repeated below for convenience. To spare the reader constant referral to the notation of Chapter IV we again define for fixed M and N

$$\Lambda_{NM} = \{\underline{L}(\cdot): \underline{L}(t) = \sum_{j=1}^{M} \phi_j(t-i\Delta)\underline{L}_{ij} \text{ for } t \in (t_i, t_i+\Delta] i = 0, 1, \ldots, N-1\} \tag{G.1}$$

where the functions $\phi_j(t)$ are complete in $\mathcal{L}^2[0,\Delta]$. We also define, as in Chapter IV,

$$\underline{L}_M^o(\cdot) = \underset{\underline{L}(\cdot) \in \Lambda_{NM}}{\text{arg. min}} \tag{G.2}$$

and

$$\mu_M^o = \mu(\underline{L})\Big|_{\underline{L}(\cdot) = \underline{L}_M^o(\cdot)} \tag{G.3}$$

where $\mu(\underline{L}) \overset{\Delta}{=} \text{tr } \underline{V}_L(t_o)$. We then prove

Theorem 9:

(i) $\lim_{M\to\infty} \mu_M^o = \text{tr } \underline{K}(t_o) = \mu(\underline{L}^*)$

(ii) $\lim_{M\to\infty} \|\underline{L}_M^o(\cdot) - \underline{L}^*(\cdot)\|_2 = \lim_{M\to\infty} \left[\int_{t_o}^{T} \|\underline{L}_M^o(t) - \underline{L}^*(t)\|^2 dt\right]^{1/2}$

where $\underline{L}^*(t) = \underline{B}'(t)\,\underline{K}(t; T, \underline{F})$

Using the fact that the $\{\phi_j\}$ are complete, we first prove condition (i) of this theorem which we restate as a lemma.

Lemma 1: $\lim_{M \to \infty} \mu_M^o = \text{tr } \underline{K}(t_o) = \mu(\underline{L}^*)$

Proof: The functional $\mu(\underline{L}) = \text{tr } \underline{V}_L(t_o)$ is continuous in $\underline{L}(\cdot)$. Hence, given any $\epsilon > 0$ there exists a $\delta(\epsilon)$ such that

$$\| \underline{L}_\alpha(\cdot) - \underline{L}_\beta(\cdot) \|_2 = \left[\int_{t_o}^{T} \| \underline{L}_\alpha(t) - \underline{L}_\beta(t) \|^2 dt \right]^{1/2} < \delta \tag{G.4}$$

implies

$$| \mu(\underline{L}_\alpha) - \mu(\underline{L}_\beta) | < \epsilon \tag{G.5}$$

†

We now let $\hat{\underline{L}}_M(\cdot) \epsilon \Lambda_{NM}$ such that

$$\| \hat{\underline{L}}_M(\cdot) - \underline{L}^*(\cdot) \|_2 < \delta$$

Such an $\hat{\underline{L}}_M$ exists for sufficiently large M since the elements of $\underline{L}^*(\cdot)$ are of class $\mathcal{L}^2[t_o, T]$ and the sequence $\{\phi_j\}$ is complete on each subinterval $[t_i, t_i + \Delta]$. We simply take M large enough so that

$$\max_i \left[\int_{t_i}^{t_i+\Delta} \| \hat{\underline{L}}_M(t) - \underline{L}^*(t) \|^2 \, dt \right]^{1/2} \leq \frac{\delta}{N}$$

We now let $\underline{L}_M^o(\cdot)$ be that element of Λ_{NM} which minimizes $\mu(\underline{L})$ as indicated by Eq. (G.2). Then, with the aid of Eqs. (G.4) and (G.5) we have

$$\text{tr } \underline{K}(t_o) \leq \mu_M^o \leq \mu(\hat{\underline{L}}_M) < \mu(\underline{L}^*) + \epsilon$$

---

† This fact immediately follows from either Eq. (C.15) or (C.17).

or $$\text{tr } \underline{K}(t_o) \leq \mu_M^o < \text{tr } \underline{K}(t_o) + \epsilon$$

But since ϵ is arbitrary it follows that

$$\lim_{M \to \infty} \mu_M^o = \text{tr } \underline{K}(t_o) \tag{G.6}$$

as claimed. ||

Now that we have proved the convergence of the sequence of numbers μ_M^o, representing the minima of the functional $\text{tr } \underline{V}_L(t_o)$ on the sets of functions Λ_{NM}, it is natural to try to prove the convergence of the sequence of functions $\underline{L}_M^o(\cdot)$ for which these minima are achieved. We first must prove

Lemma 2: Let $\underline{\Phi}_M(t, t_o)$ be the transition matrix corresponding to $\underline{L}_M^o(t)$, i.e., $\underline{\Phi}_M(t, t_o)$ satisfies

$$\frac{d}{dt}\underline{\Phi}_M(t, t_o) = [\underline{A}(t) - \underline{B}(t)\underline{L}_M^o(t)]\underline{\Phi}_M(t, t_o); \underline{\Phi}_M(t_o, t_o) = \underline{I}$$

then for all $t \in [t_o, T]$,

$$\lim_{M \to \infty} \underline{\Phi}_M(t, t_o) = \underline{\Phi}^*(t, t_o) \tag{G.7}$$

where $\underline{\Phi}^*(t, t_o)$ is the transition matrix corresponding to $\underline{L}^*(t)$.

Proof: We first show that $\| [\underline{L}_M^o(\cdot) - \underline{L}^*(\cdot)]\underline{\Phi}_M(\cdot, t_o) \|_2 \to 0$ as $M \to \infty$. We use Eq. (3.28) to write

$$\mu_M^o - \mu(\underline{L}^*) = \text{tr} \int_{t_o}^{T} \underline{\Phi}_M'(t, t_o) [\underline{L}_M^o(t) - \underline{L}^*(t)]' [\underline{L}_M^o(t) - \underline{L}^*(t)] \underline{\Phi}_M(t, t_o) dt$$

(Cont. on next page)

$$\geq \int_{t_o}^{T} \lambda_{max} \{\underline{\Phi}'_M(t, t_o)[\underline{L}^o_M(t)-\underline{L}^*(t)]' \cdot [\underline{L}^o_M(t)-\underline{L}^*(t)]\underline{\Phi}_M(t, t_o)\}\, dt$$

$$= \int_{t_o}^{T} \|[\underline{L}^o_M(t)-\underline{L}^*(t)]\underline{\Phi}_M(t, t_o)\|^2\, dt \qquad (G.8)$$

Hence,

$$\mu^o_M - \mu(\underline{L}^*) \geq (\|[\underline{L}^o_M(\cdot)-\underline{L}^*(\cdot)]\underline{\Phi}_M(\cdot, t_o)\|_2)^2 \to 0 \text{ as } M \to \infty$$

where we have the right hand side of this inequality tends to zero with $M \to \infty$ since $\mu^o_M \to \mu(\underline{L}^*)$ by Lemma 1.

We now write

$$\dot{\underline{\Phi}}_M(t, t_o) = [\underline{A}(t)-\underline{B}(t)\underline{L}^*(t)]\underline{\Phi}_M(t, t_o) - \underline{B}(t)[\underline{L}^o_M(t)-\underline{L}^*(t)]\underline{\Phi}_M(t, t_o)$$

Hence,

$$\underline{\Phi}_M(t, t_o) - \underline{\Phi}^*(t, t_o) = -\int_{t_o}^{t} \underline{\Phi}^*(t, \tau)\underline{B}(\tau)[\underline{L}^o_M(\tau)-\underline{L}^*(\tau)]\underline{\Phi}_M(\tau, t_o)d\tau \qquad (G.9)$$

Taking norms yields

$$\|\underline{\Phi}_M(t, t_o) - \underline{\Phi}^*(t, t_o)\| \leq \int_{t_o}^{t} \|\underline{\Phi}^*(t, \tau)\underline{B}(\tau)\| \cdot \|[\underline{L}^o_M(\tau)-\underline{L}^*(\tau)]\underline{\Phi}_M(\tau, t_o)\|\, d\tau$$

where we have used the fact that for induced matrix norms $\|\underline{A}\underline{B}\| \leq \|\underline{A}\| \cdot \|\underline{B}\|$. Applying the Cauchy-Schwarz inequality to the right hand side yields

$$\|\underline{\Phi}_M(t, t_o)-\underline{\Phi}^*(t, t_o)\| \leq [\int_{t_o}^{t} \|\underline{\Phi}^*(t, \tau)\underline{B}(\tau)\|^2 d\tau]^{1/2} \cdot [\int_{t_o}^{t} \|(\underline{L}^o_M-\underline{L}^*)\underline{\Phi}_M(t, t_o)\|^2 \cdot d\tau]^{1/2}$$

The first integral on the right hand side is bounded, say, by C_1. Hence

$$\|\underline{\Phi}_M(t, t_o) - \underline{\Phi}^*(t, t_o)\| \leq C_1 \left[\int_{t_o}^{T} \|(\underline{L}_M^o - \underline{L}^*)\underline{\Phi}_M(\tau, t_o)\|^2 d\tau \right]^{1/2}$$

$$= C_1 \|[\underline{L}_M^o(\cdot) - \underline{L}^*(\cdot)]\underline{\Phi}_M(\cdot, t_o)\|_2 \qquad (G.10)$$

But by our earlier argument, the norm appearing on the right hand side tends to zero as $M \to \infty$. Hence

$$\lim_{M \to \infty} \|\underline{\Phi}_M(t, t_o) - \underline{\Phi}^*(t, t_o)\| = 0$$

as claimed. ||

Having proven Lemma 2, which is the major step in the proof of $\underline{L}_M^o(\cdot) \to \underline{L}^*(\cdot)$, we can now go on to complete this proof.

Lemma 3: $\|\underline{L}_M^o(\cdot) - \underline{L}^*(\cdot)\|_2 \to 0$ as $M \to \infty$

Proof: Since $\underline{\Phi}_M(t, t_o) \to \underline{\Phi}^*(t, t_o)$ implies that $\underline{\Phi}_M(t, t_o)\underline{\Phi}'_M(t, t_o) \to \underline{\Phi}^*(t, t_o)\underline{\Phi}^*(t, t_o)$, we have that for all $t \in [t_o, T]$ there exists a constant C_2 such that

$$\lambda_{\min} \{\underline{\Phi}_M(t, t_o)\underline{\Phi}'_M(t, t_o)\} \geq C_2 > 0 \qquad (G.11)$$

uniformly in M. We now again make use of Eq. (3.28) to write

$$\mu_M^o - \mu(\underline{L}^*) = \int_{t_o}^{T} \text{tr} \{[\underline{L}_M^o(t) - \underline{L}^*(t)]' \cdot [\underline{L}_M^o(t) - \underline{L}^*(t)]\underline{\Phi}_M(t, t_o)\underline{\Phi}'_M(t, t_o)\} dt$$

(G.12)

We now apply the inequality

$$\lambda_{min}(\underline{A})\lambda_{max}(\underline{B}) \leq \lambda_{min}(\underline{A})\ tr\ \underline{B} \leq tr\ \underline{A}\underline{B} \leq \lambda_{max}(\underline{A})\ tr\ \underline{B}$$

which is valid if $\underline{A}$ and $\underline{B}$ are positive semi-definite, to Eq. (G.12). This results in

$$\mu_M^o - \mu(\underline{L}^*) \geq \int_{t_o}^{T} \lambda_{min}\left[\underline{\Phi}_M(t, t_o)\underline{\Phi}'_M(t, t_o)\right] \cdot \left\|\underline{L}_M^o(t) - \underline{L}^*(t)\right\|^2 dt$$

But using Eq. (G.11) yields

$$\mu_M^o - \mu(\underline{L}^*) \geq C_2 \cdot \left[\left\|\underline{L}_M^o(\cdot) - \underline{L}^*(\cdot)\right\|_2\right]^2$$

Hence the desired result is established since $\mu_M^o - \mu(\underline{L}^*) \to 0$. ‖

This completes the proof of Theorem 9. We have shown that the intuitively expected results are in fact true. We showed that $\underline{L}_M^o(\cdot)$ converged to $\underline{L}^*(\cdot)$ in a mean square sense, often written as $\underline{L}_M^o(\cdot) \Rightarrow \underline{L}^*(\cdot)$. It may also be true that $\underline{L}_M^o(t) \to \underline{L}^*(t)$ for all $t \in [t_o, T]$, (i.e., pointwise convergence), but this has not as yet been proven.

REFERENCES

1. Athans, M. and Falb, P.F., Optimal Control: An Introduction to the Theory and its Applications, McGraw-Hill Book Co., New York, 1966.

2. Kalman, R.E., "Contribution to the Theory of Optimal Control," Bol. Soc. Mat. Mexico, 102-119, (1960).

3. Wiener, N., The Extrapolation, Interpolation and Smoothing of Stationary Time Series, Technology Press, M.I.T., Cambridge, Mass., 1949.

4. Newton, Gould and Kaiser, Analytical Design of Linear Feedback Controls, John Wiley and Sons, Inc., New York, 1957.

5. Booton, R.C., "An Optimization Theory for Time-Varying Linear Systems with Non-Stationary Statistical Inputs," Proc. IRE, Vol. 40, 997-981, (1952).

6. Zadeh, L.A., and Ragazzini, J.R., "An Extension of Wiener's Theory of Predictions," J. Appl. Phys., Vol. 21, 945-655 (1950).

7. Kalman, R.E., "When is a Linear System Optimal?", J. Basic Engineering, (ASME Trans.), Vol. 86, 1-10, (1964).

8. Willis, B.H., "The Frequency Domain Solution of Regulator Problems," presented at the 1965 JACC, Troy, New York, June 22-25, 1965.

9. Pontryagin, et al., Mathematical Theory of Optimal Processes, John Wiley (Interscience), 1962.

10. Bellman, R., Dynamic Programming, Princeton University Press, Princeton, New Jersey, 1957.

11. Kalman, Ho, and Narendra, "Controllability of Linear Dynamical Systems," Contributions to Differential Equations, Vol. 1, 1962.

12. Kalman, R.E., "Mathematical Description of Linear Dynamical Systems," J. SIAM on Control, Ser. A., Vol. 1, No. 2, 152-192, (1963).

13. Kleinman, D.L., "On the Linear Regulator Problem and the Matrix Riccati Equation," Electronic Systems Laboratory Report 271, Mass. Inst. of Tech., (1966).

REFERENCES (Cont.)

14. Falb, P. L. and Kleinman, D. L., "Remarks on the Infinite Dimensional Riccati Equation," IEEE Trans. on Auto. Control, Vol. 11, No. 3, 534-537, (1966).

15. Dieudonne, J., Foundations of Modern Analysis, Academic Press, Inc., 1960.

16. Kalman, Englar, and Bucy, "Fundamental Study of Adaptive Control Systems," ASD Technical Report 61-27, (1962).

17. Bellman, R. and Kalaba, R., "Quasilinearization and Nonlinear Boundary-Value Problems," RAND Report No. R-438, (1965).

18. Kantorovich, L.V., and Akilov, G.P., Functional Analysis in Normed Spaces, Pergamon Press, New York, 1964.

19. Gelfand, I.M. and Fomin, S.V., Calculus of Variations, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1963.

20. Kalman, R.E., and Bertram, J.E., "Control System Analysis and Design Via the Second Method of Lyapunov," J. Basic Engineering, (ASME Trans.), Vol. 82, 371-393, (1960).

21. Reid, W.T., "Riccati Matrix Differential Equations," Pacific J. Math., Vol. 13, 665-685 (1963).

22. Bellman, R., Introduction to Matrix Analysis, McGraw-Hill Book Co., New York, 1960.

23. Henrici, P., Discrete Variable Methods for Ordinary Differential Equations, John Wiley and Sons Inc., New York, 1962.

24. Wonham, W.M., "On Stabilization of Linear Controllable Systems," to be published.

25. Athans, M. and Levine, W.S., "On the Numerical Solution of the Matrix Riccati Differential Equation Using a Runge-Kutta Scheme," Electronic Systems Laboratory Report 276, Mass. Inst. of Tech., (1966).

26. Kleinman, D.L. and Athans, M., "The Discrete Minimum Principle with Application to the Linear Regulator Problem," Electronic Systems Laboratory Report 260, Mass. Inst. of Tech., (1966).

27. Private correspondence with A. Levis, Mass. Inst. of Tech.

28. Athans, M. and Schweppe, F., "Gradient Matrices and Matrix Calculations," MIT Lincoln Laboratory Technical Note 1965-53, (1965).

REFERENCES (Cont.)

29. Bodewig, E., Matrix Calculus, North Holland Publishing Co. and Interscience Publishers, New York, 1956.

30. Wonham, W.M., Class notes for course 6.603, Mass. Inst. of Tech., Spring 1966.

31. Kalman, R.E. and Bucy, R.S., "New Results in Linear Filtering and Prediction Theory, J. Basic Engineering, (ASME Trans.), Vol. 83, 95-108, (1961).

32. Witsenhausen, H.S., "Some Iterative Methods Using Partial Order for Solution of Nonlinear Boundary Value Problems," MIT Lincoln Laboratory Technical Note 1965-18, (1965).

33. Kelley, H.J., "Methods of Gradients" in Optimization Techniques (Leitmann, G., ed.), Academic Press, New York, 1962.

34. Bryson, A.E., and Denham, W.F., "A Steepest-Ascent Method for Solving Optimum Programming Problems," J. of Applied Math., 247-257, (June, 1962).

BIOGRAPHICAL NOTE

The author, David L. Kleinman, was born in Brooklyn, New York, on January 4, 1942. He attended the Cooper Union for the Advancement of Science and Art in New York City, from 1958 to 1962. The B.E.E. degree was received in June, 1962.

Mr. Kleinman came to the Massachusetts Institute of Technology as a Sloan fellow in September, 1962. He received the S.M.E.E. degree in June, 1963, writing a master's thesis on fuel-optimal control systems. Since 1963, Mr. Kleinman has been working in the M.I.T. Electronic Systems Laboratory as a research assistant, specializing in automatic control theory and mathematics.

The author is presently a consultant to the GPS Instrument Co., Newton, Mass., and has recently accepted a full-time staff position with Bolt, Beranek and Newman, Inc. in Cambridge, Mass. Mr. Kleinman is a member of Tau Beta Pi, Eta Kappa Nu and Sigma Xi. Among his publications are

"Fuel Optimal Control of Second-Order Linear Systems with Bounded Response Time," NEREM Record, 1964, pp. 42-43.

"The Discrete Minimum Principle with Application to the Linear Regulator Problem," Electronic Systems Laboratory Report 260, M.I.T., Cambridge, Mass., 1966.

"Remarks on the Infinite Dimensional Riccati Equation," IEEE Trans. on Automatic Control, Vol. AC-11, No. 3, July, 1966.

"On the Linear Regulator Problem and the Matrix Riccati Equation," Electronic Systems Laboratory Report 271, M.I.T., Cambridge, Mass., 1966.