SIAM J. CONTROL AND OPTIMIZATION
Vol. 14, No. 2, February 1976

CONVERGENCE AND STABILITY PROPERTIES OF THE DISCRETE RICCATI OPERATOR EQUATION AND THE ASSOCIATED OPTIMAL CONTROL AND FILTERING PROBLEMS*

WILLIAM W. HAGER† AND LARRY L. HOROWITZ‡

Abstract. The convergence properties for the solution of the discrete time Riccati matrix equation are extended to Riccati operator equations such as arise in a gyroscope noise filtering problem. Stabilizability and detectability are shown to be necessary and sufficient conditions for the existence of a positive semidefinite solution to the algebraic Riccati equation which has the following properties: (i) it is the unique positive semidefinite solution to the algebraic Riccati equation, (ii) it is converged to geometrically in the operator norm by the solution to the discrete Riccati equation from any positive semidefinite initial condition, (iii) the associated closed loop system converges uniformly geometrically to zero and solves the regulator problem, and (iv) the steady state Kalman–Bucy filter associated with the solution to the algebraic Riccati equation is uniformly asymptotically stable in the large. These stability results are then generalized to time-varying problems; also it is shown that even in infinite dimensions, controllability implies stabilizability.

1. Introduction. The purpose of this paper is to prove that the convergence and stability properties associated with the Riccati difference equation in finite dimensions also hold for the Riccati operator equation in infinite dimensions. Many of the finite-dimensional results already in the literature will also be strengthened. The Riccati difference equation has been studied by Caines and Mayne [2], Lee, Chow and Barr [9], and Zabczyk [10].

In finite dimensions, the first paper proved that if a stabilizability and an observability assumption held, then the solution to the Riccati difference equation converged to a positive definite matrix solving the algebraic Riccati equation, and furthermore, the solution to the algrebraic equation was unique in the class of positive semidefinite matrices. Their proof, however, required the Heine–Borel theorem (a closed, bounded set of $n \times n$ matrices forms a compact set) so that the proofs could not be extended to the Riccati operator equation.

The paper by Lee, Chow and Barr then showed that in a Hilbert space environment, the solution to the quadratic cost control problem could be expressed in feedback form in terms of the solution to the Riccati operator equation, and when the system dynamics were stable, then there existed a solution to the algebraic Riccati equation.

Zabczyk weakened this stability condition to stabilizability and then showed that if the cost functional was positive definite in the state variable, then the solution to the algebraic Riccati operator equation was unique in the class of positive semidefinite operators and furthermore was the limit (in the operator norm) of the solution to the Riccati equation from any positive semidefinite initial condition.

This paper contains the results above as special cases. The observability condition of Caines and Mayne and the positive definiteness of the cost functional required by Zabczyk are weakened to detectability. The positive definiteness of

* Received by the editors August 22, 1974, and in revised form March 11, 1975.

† Department of Mathematics, University of South Florida, Tampa, Florida 33620.
‡ Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, Massachusetts, 02173.

the Riccati equation solution proved by Caines and Mayne is also proved in the infinite-dimensional framework. Furthermore, it is shown that the solution to the regulator problem associated with the Riccati equation is uniformly asymptotically stable in the large if a detectability condition is satisfied and there exists a positive semidefinite solution to the algebraic Riccati equation.

The paper reaches its climax in § 5, where it is proved that stabilizability and detectability are necessary and sufficient conditions for the existence of a positive semidefinite solution to the algebraic Riccati equation which has the following properties: (i) it is the unique positive semidefinite solution to the algebraic Riccati equation, (ii) it is converged to geometrically in the operator norm by the solution to the discrete Riccati equation from any positive semidefinite initial condition, and (iii) the associated closed loop system converges uniformly geometrically to zero and solves the regulator problem.

The stability of the Kalman–Bucy filter for time-varying infinite-dimensional systems under stabilizability and detectability is also treated in the Appendix. This weakens the conditions of controllability, observability, and nonsingularity of the transition operator that Deyst and Price [3] required in their proof of the stability of the solution to the time-varying filtering problem in finite dimensions.

The paper concludes with an illustration of the use of the Riccati operator equation in filtering the noise additively corrupting a gyroscope's output signal. In this example, the domain of the Riccati operator is an L^2-space.

2. Problem statement. Let $K(S, T, i)$ denote the solution to the *Riccati operator equation* given by

$$K(i-1) = A^*(i)\{K(i) - K(i)B(i)[R(i) + B^*(i)K(i)B(i)]^{-1}B^*(i)K(i)\}A(i) + Q(i) \tag{1}$$

with boundary condition $K(T) = S$, where i is an integer, $i \leqq T$, and the following operators appearing in (1) are uniformly bounded linear mappings on Hilbert spaces Y and U: $Q(i): Y \to Y$, $S: Y \to Y$, $A(i): Y \to Y$, $B(i): U \to Y$, and $R(i): U \to U$. (Throughout this paper, the term operator will mean a bounded linear operator.) The inner products on both Hilbert spaces will be denoted by $(\cdot, \cdot)$—the inner product being used should be clear from context. The norm of a vector $y \in Y$ is given by $\|y\| = (y, y)^{1/2}$ and the norm of a linear operator $P: Y \to Y$ is given by $\|P\| = \sup\{\|Py\| : \|y\| = 1\}$. The operator P^* denotes the adjoint of an operator P. P is said to be positive if it is positive semidefinite and self-adjoint; i.e., $P^* = P$ and $(y, Py) \geqq 0$ for all $y \in Y$. The operators $Q(i)$, $R(i)$, and S are assumed positive, and furthermore, $R(i)$ is assumed uniformly positive definite, i.e., $(u, R(i)u) \geqq a\|u\|^2$ for some $a > 0$ and for all $u \in U$, where "a" is independent of i. The notation $P_1 \geqq P_2$ and $P_1 > P_2$ means that $P_1 - P_2$ is positive semidefinite and positive definite respectively.

Associated with the Riccati equation is the *control problem*:

$$\underset{\{u(i)\}}{\text{Minimize}} \left[(Sy(T), y(T)) + \sum_{i=i_0}^{T-1} \{(y(i), Q(i)y(i)) + (u(i), R(i)u(i))\} \right], \tag{2}$$

$$\text{Subject to} \quad y(i+1) = A(i)y(i) + B(i)u(i), \tag{3}$$

$$y(i_0) = y_0 \in Y, \qquad u(i) \in U.$$

Let $J(S, T, i_0, y_0)$ denote the optimal value for the control problem above. As shown in [1] for finite-dimensional spaces,

$$J(S, T, i_0, y_0) = (y_0, K(S, T, i_0)y_0), \tag{4}$$

and the optimal control in feedback form is given by

$$u(i) = -[R(i) + B^*(i)K(S, T, i+1)B(i)]^{-1}B^*(i)K(S, T, i+1)A(i)y(i). \tag{5}$$

The extension of these results to Hilbert spaces is trivial as noted in [6], since the dynamic programming argument used in the derivation of (4) and (5) does not require finite-dimensionality and can be performed in a Hilbert space environment.

The cost function (2) is nonnegative, so (4) implies that $K(S, T, i) \geqq 0$ for all $i \leqq T$, and hence the inverse appearing in (1) and (5) exists and is bounded since $R(i) > 0$. Thus $K(S, T, i)$ is a positive operator for $i \leqq T$.

When (1) is time-invariant (i.e., $A(i) = A$, $B(i) = B$, etc.), then also associated with (1) is the *algebraic Riccati equation* (abbreviated ARE):

$$K = A^*[K - KB(R + B^*KB)^{-1}B^*K]A + Q. \tag{6}$$

Similarly associated with the control problem when the system is time invariant is the *regulator problem*

$$\underset{\{u(i)\}}{\text{Minimize}} \left[\sum_{i=0}^{\infty} (y(i), Qy(i)) + (u(i), Ru(i)) \right], \tag{7}$$

$$\begin{aligned} &\text{Subject to} \quad y(i+1) = Ay(i) + Bu(i), \\ &\qquad\qquad\quad y(0) = y_0 \in Y, \quad u(i) \in U. \end{aligned} \tag{8}$$

Let $J(y_0)$ denote the optimal cost for the regulator problem above.

The estimation problem, or dual problem corresponding to the control problem, is given in Appendix C.

For future reference, the following abbreviations are used throughout the paper:

ARE	algebraic Riccati equation
UASL	uniformly asymptotically stable in the large
ST	stabilizability
DT	detectability
CT	controllability
OB	observability

3. The assumptions. The following *stabilizability* and *detectability* assumptions will appear in the development. These conditions are first stated for time-invariant problems:

(ST) There exists an integer $r \geqq 1$, a constant q, and an operator L such that

$$\|(A - BL)^r\| < q < 1. \tag{9}$$

(DT) There exist integers $s, t \geqq 0$ and constants $0 \leqq d < 1$, $0 < b < \infty$, such that

whenever $\|A^i y\| \geqq d\|y\|$, then

$$\left(y, \sum_{i=0}^{s} A^{*i} Q A^i y\right) \geqq b(y, y). \tag{10}$$

When the problem is time varying, we replace L in (ST) by a sequence $\{L(i)\}$ of uniformly bounded linear operators and require

$$\left\| \prod_{i=k}^{k+r-1} (A(i) - B(i)L(i)) \right\| < q < 1 \tag{ST'}$$

for $k = 0, r, 2r, \cdots$.

Similarly in (DT) we replace A^i by $C(i + k, k)$, where $C(i, k) = A(i - 1) \cdot A(i - 2) \cdots A(k)$ and $C(i, i) = I$, the identity operator, and require that for all $k \geqq 0$, whenever $\|C(k + t, k)y\| \geqq d\|y\|$, then

$$\left(y, \sum_{i=0}^{s} C(k + i, k)^* Q(k + i) C(k + i, k) y\right) \geqq b(y, y). \tag{DT'}$$

Special cases of (ST) and (DT) are the *controllability* and *observability* conditions:

(CT) There exists an integer $r \geqq 0$ and a constant $0 < a < \infty$ such that

$$\left(y, \sum_{i=0}^{r} A^i B B^* A^{*i} y\right) \geqq a(y, y) \tag{11}$$

for all $y \in Y$.

(OB) There exists an integer $s \geqq 0$ and a constant $0 < b < \infty$ such that

$$\left(y, \sum_{i=0}^{s} A^{*i} Q A^i y\right) \geqq b(y, y) \tag{12}$$

for all $y \in Y$.

Note that (OB) is trivially a special case of (DT). At the end of § 4, it will also be shown that (CT) implies (ST).

Recall that in finite dimensions, the pair of matrices $[A, B]$ are said to be stabilizable if there exists a matrix L such that the spectral radius $\rho(A - BL)$ is less than 1. (A, B, and L are assumed to be $n \times n$, $n \times m$, and $m \times n$ respectively.) Similarly $[C, A]$ is detectable if $[A^*, C^*]$ is stabilizable. Note that it follows immediately that (ST) is equivalent to the condition $\rho(A - BL) < 1$ for some L since $\rho(P) = \lim_{k \to \infty} \|P^k\|^{1/k}$ (see [4, p. 567]).

In Appendix B, it is proved that in finite dimensions, (DT) is equivalent to the condition that $\rho(A^* - C^*L) < 1$ for some L where $Q = C^*C$.

4. The main results. The first lemma gives a uniform bound for the solution $K(S, T, i)$ of the Riccati equation (1).

LEMMA 1. *If* (ST') *holds, then there exists a constant c independent of i and T such that $K(S, T, i) < cI$ and $J(y) < c\|y\|^2$, where $J(y_0)$ is the optimal cost for the regulator problem* (7).

Proof. By the relation (4), the bound on $K(S, T, i)$ will be proved if the optimal cost in the control problem (2) can be bounded in terms of the initial condition y_0. Since the operators $A(\cdot)$, $B(\cdot)$, and $L(\cdot)$ are all uniformly bounded, there exists a constant c such that

$$\prod_{i=j}^{j+m} \|N(i)\| \leqq c \tag{13}$$

for all m satisfying $0 \leqq m \leqq r$, where $N(i) = A(i) - B(i)L(i)$. (Throughout this paper, c will denote a generic constant whose value does not depend on T or i and whose value in different equations may change.)

Using the control $u(i) = -L(i)y(i)$ in the system dynamics leads to the estimates

$$\|y(k+1)\| = \|N(k)y(k)\| = \left\| \prod_{i=0}^{k} (N(i))y_0 \right\| \leqq cq^{k/r}\|y_0\|, \tag{14}$$

where the last inequality follows by grouping the operators $N(i)$ into groups of r factors and then applying the bound (ST′). Since $u(i) = -L(i)y(i)$, then $u(i)$ obeys a similar estimate. Inserting these bounds on $u(i)$ and $y(i)$ into the cost functional (2) leads to a bound on $J(S, T, i, y_0)$ of the form $c\sum_{k=0}^{\infty} q^{2k/r}\|y_0\|^2$. Since $q < 1$, the geometric series is convergent and $J(S, T, i, y_0) < c\|y_0\|^2$ as desired. Since c is independent of T and i, then the bound on $J(y)$ also follows immediately. □

A sequence of operators P_k is said to *converge strongly* to P if $\lim_{k\to\infty} \|(P - P_k)y\| = 0$ for all $y \in Y$. An elementary property of operators is the following (see [4, p. 925]): Suppose $\{P_k\}$ is a sequence of uniformly bounded self-adjoint operators satisfying $P_k \leqq P_{k+1}$ for $k \geqq 0$; then $\{P_k\}$ converges strongly to a self-adjoint operator P satisfying $P_k \leqq P$ for all $k \geqq 0$. The sequence P_k *converges weakly* to P if $\lim_{k\to\infty} (z, (P_k - P)y) = 0$ for all $y, z \in Y$. It can be shown that this last condition is equivalent to requiring $\lim_{k\to\infty} (y, (P_k - P)y) = 0$ for all $y \in Y$.

For the remainder of this section, we will only be dealing with the time-invariant Riccati equation and control problem. In Appendix A, the question of stability for time varying systems is considered. Let $K(T, i)$ denote the solution to the time-invariant Riccati equation when the terminal condition vanishes ($S = 0$).

THEOREM 1. *If* $J(0, T, 0, y) < c\|y\|^2$ *for some* c *independent of* T, *then* $K(T, i)$ *converges strongly as* $T \to \infty$ *to a positive operator* P *that satisfies the* ARE.

Proof. Since (4) holds, then $K(T, i) < cI$ and hence $\|K(T, i)\|$ is uniformly bounded by c. Also, $(y, K(T_1, i)y) = J(0, T_1, i, y) \geqq J(0, T_2, i, y) = (y, K(T_2, i)y)$ whenever $T_1 \geqq T_2$ since increasing the terminal time cannot decrease the optimal cost. Thus by the remarks preceding the theorem, $K(T, i) \to P$ strongly as $T \to \infty$. If $F(K)$ denotes the right-hand side of (6), then (1) can be written as $K(T+1, i) = K(T, i-1) = F(K(T, i))$, where the first equality follows since the equation is time-invariant. Now $K(T+1, i) \to P$ strongly as $T \to \infty$ and furthermore by [4, p. 922], $F(K(T, i)) \to F(P)$ strongly as $T \to \infty$. Thus $P = F(P)$ and hence P solves the ARE. □

Combining Lemma 1 and Theorem 1 yields the following.

COROLLARY 1. *If* (ST) *holds, then* $K(T, i) \to P$ *strongly as* $T \to \infty$, *where* P *solves the* ARE.

Later it will be shown that when (DT) holds and there exists a positive solution to the (ARE), then (ST) holds.

The stability of the solution to the following system when P is a positive solution to the ARE will now be studied:

$$y(i+1) = Ay(i) + Bu(i), \quad y(0) = y_0, \quad u(i) = Fy(i), \tag{15}$$

$$F = -[R + B^*PB]^{-1}B^*PA. \tag{16}$$

The following system of inequalities and equalities plays an important role in the development:

$$-(u(i), B^*PAy(i)) = (u(i), [R + B^*PB]u(i)) \tag{17}$$

$$(y(i), Py(i)) \geqq (y(i), Py(i)) - (y(j), Py(j)) \tag{18}$$

$$= \sum_{k=i}^{j-1} (y(k), Py(k)) - (y(k+1), Py(k+1)) \tag{19}$$

$$= \sum_{k=i}^{j-1} (y(k), Py(k) - A^*PAy(k)) - (u(k), B^*PAy(k)) - (B^*PAy(k), u(k)) - (u(k), B^*PBu(k)) \tag{20}$$

$$= \sum_{k=i}^{j-1} (y(k), Qy(k)) - (u(k), B^*PBu(k)) - (u(k), B^*PAy(k)) \tag{21}$$

$$= \sum_{k=i}^{j-1} (y(k), Qy(k)) + (u(k), Ru(k)) \geqq 0. \tag{22}$$

Above, $j > i$ and (17) follows by multiplying $u(i) = Fy(i)$ by $[R + B^*PB]$ and (18), (20), (21), and (22) follow by the positivity of P, (15), the ARE that P satisfies, and (17), respectively.

THEOREM 2. *$J(0, T, 0, y) < c\|y\|^2$ for some constant c independent of T if and only if there exists a positive solution to the* ARE.

Proof. The theorem in the forward direction was proved by Theorem 1. Now suppose P is a positive solution to the ARE and let $y_s(i)$ and $u_s(i)$ be the state and control generated by (15). Then by the relation (18),

$$(y_0, Py_0) \geqq \sum_{k=0}^{T-1} (y_s(k), Qy_s(k)) + (u_s(k), Ru_s(k)). \tag{23}$$

Since P is bounded, then $J(0, T, 0, y) \leqq \|P\| \|y\|^2$. □

Recall that the dynamical system $x(k+1) = f(x(k), k)$, $x(i_0) = x_0$ is said to be *uniformly asymptotically stable in the large* (abbreviated UASL) with respect to x^* if the following holds [8]:

(i) Given $\varepsilon > 0$, there exists $\delta > 0$ such that $\|x^* - x_0\| \leqq \delta$ implies that $\|x(k) - x^*\| \leqq \varepsilon$ for any k, i_0 satisfying $k \geqq i_0$.

(ii) Given $\delta > 0$, there exists $\varepsilon > 0$ such that $\|x^* - x_0\| \leqq \delta$ implies $\|x(k) - x^*\| \leqq \varepsilon$ for any k, i_0 satisfying $k \geqq i_0$.

(iii) Given δ, $\varepsilon > 0$, there exists T such that $\|x(k) - x^*\| \leqq \varepsilon$ for all k, i_0, x_0 satisfying $k \geqq T + i_0$ and $\|x_0 - x^*\| \leqq \delta$.

THEOREM 3. *If* $K(T,0) \to P$ *strongly,* P *solves the* ARE, *and the system* (15) *is* UASL *with respect to the origin, then* P *is the unique solution to the* ARE *in the class of positive operators and* $K(S,T,i)$ *converges strongly to* P *as* $T \to \infty$ *for any* $S \geqq 0$. *Also the state and the control generated by* (15) *are the optimal solutions for the regulator problem and* (y_0, Py_0) *is the optimal cost.*

Proof. Let $\{y_s(i)\}$ and $\{u_s(i)\}$ be generated by (15) using P. Then (18) implies

$$(24) \qquad (y_0, Py_0) \geqq \sum_{k=0}^{T-1} (y_s(k), Qy_s(k)) + (u_s(k), Ru_s(k)) \geqq (y_0, K(T,0)y_0) \leqq J(y_0).$$

Since $K(T,0) \to P$, then as $T \to \infty$, the $\geqq$'s in (24) become $=$'s, and the last $\leqq$ implies that $\{u_s(i)\}$ must actually achieve the optimal cost in the regulator problem. Since the cost function (7) is a strictly convex function of $\{u(i)\}$, then $\{u_s(i)\}$ must be the unique optimal control sequence and (y_0, Py_0) is the optimal cost.

Now consider the following inequalities:

$$(25) \qquad (y_s(T), Sy_s(T)) + (y_0, Py_0) \geqq (y_s(T), Sy_s(T)) + \sum_{k=0}^{T-1} [(y_s(k), Qy_s(k)) + (u_s(k), Ru_s(k))]$$

$$(26) \qquad \geqq (y_0, K(S,T,0)y_0) \geqq (y_0, K(0,T,0)y_0).$$

The second inequality above follows since $(y_0, K(S,T,0)y_0)$ is the optimal cost in the control problem (2) and the third inequality follows since the optimal cost when $S = 0$ is bounded by the optimal cost when a nonnegative terminal cost is present. By assumption, the right side of (26) converges to (y_0, Py_0) as $T \to \infty$, and since the system (15) is UASL with respect to the origin, then $y_s(T) \to 0$ as $T \to \infty$. Thus all the inequalities in (25) become equalities as $T \to \infty$ and hence $K(S,T,0) \to P$ weakly. An elementary application of the Schwarz inequality for positive operators shows that weak convergence implies strong convergence (see again [4, p. 925]).

If $\bar{P}$ is any positive solution to the ARE, then it is easy to see that $K(\bar{P},T,0) = \bar{P}$ for all T and since $K(\bar{P},T,0) \to P$, then $\bar{P} = P$. □

Now it is shown that if (DT) holds, then the stability condition of Theorem 3 is satisfied.

THEOREM 4. *Suppose* P *is a positive operator solving the* ARE *and* (DT) *holds; then the solution to the system* (15) *is* UASL *with respect to the origin.*

Proof. It is shown that $\|y(k+i)\| \leqq c2^{-i/N}\|y(k)\|$ for some N, $c > 0$ independent of k and i, so that the theorem follows immediately.

Step 1. *Suppose that for some* i, $\|A^i y(i)\| \geqq d\|y(i)\|$, *where* d *was given in* (DT); *then there exists a constant* $m > 0$ *independent of* i *such that*

$$(27) \qquad (y(i), Py(i)) - (y(i+s+1), Py(i+s+1)) \geqq m\|y(i)\|^2.$$

Proof of Step 1. Let Δ^2 denote the left side of (27) and let c again denote a generic constant. By (18),

$$(28) \qquad \Delta^2 \geqq \sum_{k=i}^{i+s} (u(k), Ru(k)) \geqq a \sum_{k=i}^{i+s} \|u(k)\|^2,$$

where a satisfies $R > aI > 0$.

Letting $z(\cdot)$ denote the solution to $z(k+1) = Az(k), z(k=i) = y(i)$, then the error $e(k) = y(k) - z(k)$ satisfies, for $i \leqq k \leqq i+s$,

$$\|e(k+1)\| \leqq \|A\|\,\|e(k)\| + \|B\|\,\|u(k)\| \leqq \sum_{j=i}^{k} \|A\|^{k-j}\|B\|\,\|u(j)\| \tag{29}$$

$$\leqq c\sum_{j=i}^{k} \|u(j)\| \leqq c\left[\sum_{j=i}^{k} \|u(j)\|^2\right]^{1/2} \leqq c\Delta, \tag{30}$$

where the last set of inequalities follow by the Schwarz inequality and the bound (28) on the control.

The relation (18) also yields

$$\Delta^2 \geqq \sum_{k=i}^{i+s} (y(k), Qy(k)) = \sum_{k=i}^{i+s} (e(k)+z(k), Q(e(k)+z(k))) \tag{31}$$

$$\geqq \sum_{k=i}^{i+s} (y(i), A^{*k-i}QA^{k-i}y(i)) - 2\|e(k)\|\,\|Q\|\,\|A^{k-i}y(i)\| \tag{32}$$

$$\geqq b\|y(i)\|^2 - c\Delta\|y(i)\|, \tag{33}$$

where b was given in (10); the inequality (32) follows by the Schwarz inequality and (33) follows by the bound on $e(k)$ in (30). Completing the square in (33) leads to $\|y(i)\|^2 \leqq c\Delta^2$, the desired result.

Step 2. Suppose that $\|A^t y(i)\| \leqq d\|y(i)\|$ *for* $i = k, k+t, \cdots, k+nt$. *Then there exists a constant* M *independent of* n *and* k *such that* $\|y(i)\|^2 \leqq M\|y(k)\|^2$ *for* $k \leqq i \leqq k+nt$.

Proof of Step 2. For notational convenience, suppose $k = 0$. First let $j = lt$ where $0 \leqq l \leqq n$. Then

$$\|y(j)\| = \left\| A^t y(j-t) + \sum_{i=0}^{t-1} A^i Bu(j-1-i) \right\| \leqq d\|y(j-t)\| + c\sum_{i=0}^{t-1} \|u(j-1-i)\| \tag{34}$$

$$\leqq d\|y(j-t)\| + c\left(\sum_{i=0}^{t-1} \|u(j-1-i)\|^2\right)^{1/2} \tag{35}$$

$$\leqq d^l\|y(0)\| + c\left(\sum_{i=0}^{j-1} \|u(i)\|^2\right)^{1/2}, \tag{36}$$

where the Schwarz inequality was used to derive (35) and the last inequality follows by writing the solution to the difference inequality (35) as the convolution of the forcing term $c(\sum_{i=0}^{t-1}\|u(j-1-i)\|^2)^{1/2}$ with d^i and then applying the Schwarz inequality to the convolution; since $d < 1$, then the $\sum d^{2i}$ factor in the Schwarz inequality is bounded. Now by (18),

$$a\sum_{i=0}^{j-1} \|u(i)\|^2 \leqq \|P\|\,\|y(0)\|^2, \tag{37}$$

where $R > aI$. Inserting this bound in (36) yields the desired estimate for $j = lt$.

For $lt < j < (l+1)t$, the relation $y(k+1) = Ay(k) + Bu(k)$ combined with the bound (37) on the controls and the bound above on $\|y(lt)\|$ proves the estimate.

Step 3. *Suppose that* $s_{j+1} \geqq s_j$, $s_j \to \infty$ *as* $j \to \infty$ *and* $|s_j - s_{j+1}|$ *is bounded independent of* j. *Then there exists a constant* c *independent of* j *such that* $\|y(i)\| \leqq c\|y(s_j)\|$ *for* $s_j \leqq i \leqq s_{j+1}$ *and for all* j.

Proof of Step 3. $y(i) = A^{i-s_j}y(s_j) + \sum_{k=s_j}^{i-1} A^{i-k-1}Bu(k)$. Since $|i - s_j| \leqq |s_{j+1} - s_j|$ is uniformly bounded, then the bound on $y(i)$ follows immediately from a bound of the form (37) on the controls where $y(0)$ is replaced by $y(s_j)$ and the summation is from $k = s_j$ to $i - 1$.

Let $\sqrt{D}$ be the maximum constant given in Step 3 corresponding to those sequences of integers $\{s_j\}$ satisfying $s_{j+1} = s_j + s + 1$. Now choose N_1, N_2, and N_3 large enough that the following conditions hold:

$$\|P\|/mN_1 < \tfrac{1}{4}, \tag{38}$$

$$d^{N_2}\bar{M}^{1/2} + c(\|P\|/aN_3)^{1/2} < \tfrac{1}{2}, \tag{39}$$

$$\text{where } \bar{M} = \max\{M, MD\|P\|/m\}, \tag{40}$$

where m was given in (27), M appeared in Step 2, c is the same constant appearing on the right side of (36), D appeared above at the end of Step 3, and $d < 1$ is given in (DT). Let $N = N_1N_2M_3 \max(s+1, t)$.

Step 4. *There exists* $i \in [k, k+N]$ *such that* $\|y(i)\| < \frac{1}{2}\|y(k)\|$ *for any* $k \geqq 0$.

Proof. For notational convenience, choose $k = 0$. Construct a sequence $\{t_j\}$ and $\{f_j\}$ as follows: $t_0 = 0$; for $j \geqq 0$,

$$\text{if } \|A^t y(t_j)\| \leqq d\|y(t_j)\|, \quad \text{then } t_{j+1} = t_j + t, \quad f_j = 0,$$

$$\text{if } \|A^t y(t_j)\| > d\|y(t_j)\|, \quad \text{then } t_{j+1} = t_j + s + 1, \quad f_j = 1.$$

By (18), $(y(t_j), Py(t_j)) - (y(t_{j+1}), Py(t_{j+1})) \geqq 0$, so combining this with (27) yields

$$(y(t_j), Py(t_j)) - (y(t_{j+1}), Py(t_{j+1})) \geqq f_j m\|y(t_j)\|^2. \tag{41}$$

Let J be the first index with $t_j \geqq N$. Adding the inequalities (41) for $j = 0, 1, \cdots, J-1$ yields

$$\begin{aligned}\|P\|\,\|y(0)\|^2 \geqq (y(0), Py(0)) &\geqq (y(t_J), Py(t_J)) + \sum_{j=0}^{J-1} f_j m\|y(t_j)\|^2 \\ &\geqq \sum_{j=0}^{J-1} f_j m\|y(t_j)\|^2.\end{aligned} \tag{42}$$

If at least N_1 of the f_j do not vanish, then the sum on the right side of (42) is bounded below by $mN_1 \min \|y(t_j)\|^2$, where the min is over j such that $f_j = 1$. If $j = n$ achieves the minimum, then $\|y(t_n)\|^2 \leqq \|P\|\,\|y(0)\|^2/mN_1$. Hence Step 4 would follow by (38).

Now if less than N_1 of the f_j equal 1, then there is a sequence of N_2N_3 consecutive j's with $f_j = 0$ since $N = N_1N_2N_3 \max(s+1, t)$ and hence $J \geqq N_1N_2N_3$. Let $k_1 = t_j$ be chosen such that $f_{j+i} = 0$ for $0 \leqq i \leqq N_2N_3 - 1$ and either $f_{j-1} = 1$ or $t_j = 0$. Let $k_2 = t_l$ mark the end of this sequence of f_i's that vanish. By Step 2,

$\|y(i)\|^2 \leqq M\|y(k_1)\|^2$ whenever $k_1 \leqq i \leqq k_2$. If $f_{j-1} = 1$, then the inequalities $\|y(k_1)\|^2 \leqq D\|y(t_{j-1})\|^2 \leqq D\|P\|\,\|y(0)\|^2/m$ follow by Step 3, the choice of D above and (42). Combining these last two sets of inequalities yields $\|y(i)\|^2 \leqq MD\|P\|\,\|y(0)\|^2/m$ if $k_1 \neq 0$ and $\|y(i)\|^2 \leqq M\|y(0)\|^2$ if $k_1 = 0$. Thus $\|y(i)\|^2 \leqq \overline{M}\|y(0)\|^2$, where $\overline{M}$ is given in (40).

Divide $[k_1, k_2]$ into subintervals of length $N_2 t$. Since $|k_1 - k_2| \geqq N_2 N_3 t$, then there are $\geqq N_3$ of these subintervals. By (37), one of these subintervals $[r_1, r_2]$ must satisfy

$$\sum_{i=r_1}^{r_2} \|u(i)\|^2 \leqq \frac{\|P\|}{aN_3}\|y(0)\|^2 \tag{43}$$

(i.e., the smallest sum of the form (43) is bounded by the average sum).

For $j = r_1 + N_2 t = r_2$, the inequality (34) implies

$$\|y(r_2)\| \leqq d^{N_2}\|y(r_1)\| + c\left\{\sum_{i=r_1}^{r_2} \|u(i)\|^2\right\}^{1/2}. \tag{44}$$

Inserting the bounds above on $\|y(i)\|^2$ and (43) into (44) yields

$$\|y(r_2)\| \leqq \left[d^{N_2}\overline{M}^{1/2} + c\left(\frac{\|P\|}{aN_3}\right)^{1/2}\right]\|y(0)\| < \tfrac{1}{2}\|y(0)\|, \tag{45}$$

where the last inequality follows by (39). This completes Step 4 and the geometric convergence follows by combining Steps 3 and 4. □

COROLLARY 2. *If P is a positive solution to the* ARE *and* (DT) *holds, then P is the unique solution to the* ARE *in the class of positive operators and $K(S, T, i) \to P$ geometrically in the operator norm as $T \to \infty$ for any $S \geqq 0$. Also the state and the control generated by* (15) *are optimal solutions to the regulator problem and the solution to the system* (15) *converges to zero uniformly and geometrically.*

Proof. By Theorems 2 and 1, there exists a solution $\bar{P}$ to the ARE such that $K(T, i) \to \bar{P}$ strongly as $T \to \infty$. By Theorem 4, since (DT) holds, the system (15) is UASL with respect to the origin and hence by Theorem 3, $\bar{P} = P$ and (y_0, Py_0) is the optimal cost for the regulator problem.

Let $y(T, i)$ denote the optimal solution to the control problem (2) in the time-invariant case when $S = 0$ and $i_0 = 0$. It can be shown that $y(T, i) \to 0$ uniformly and geometrically as $T \to \infty$. This follows since (18) holds with $(y(i), Py(i))$ replaced by $(y(T, i), K(T, i)y(T, i))$, and hence all the steps of Theorem 4 are valid with $y(i)$ replaced by $y(T, i)$ and P replaced by $K(T, i)$. Note that the proof of Theorem 4 required a bound on $\|P\|$ and hence will require a uniform bound on $\|K(T, i)\|$ for the finite terminal-time case; however, since (y_0, Py_0) is the optimal cost for the regulator problem by Theorem 3, then $(y_0, Py_0) \geqq (y_0, K(T, i)y_0)$ and $\|P\| \geqq \|K(T, i)\|$. (Berberian [11] shows that if Z is a positive operator, then $\|Z\| = \sup\{(y, Zy): \|y\| = 1\}$.)

Now

$$(y_0, Py_0) \leqq (y_0, K(T, 0)y_0) + (y(T, T), Py(T, T)).$$

Combining this with (25) yields

$$\begin{aligned}(y_s(T), Sy_s(T)) + (y_0, Py_0) &\geqq (y_0, K(S, T, 0)y_0)\\ &\geqq (y_0, Py_0) - (y(T, T), Py(T, T)).\end{aligned}$$

Since there exist c, q satisfying $\|y(T, T)\|, \|y_s(T)\| \leqq cq^T\|y_0\|$ and $0 < q < 1$, then $|(y_0, Py_0 - K(S, T, 0)y_0)| \leqq cq^{2T}\|y_0\|^2$ for some $c > 0$. Hence Berberian's theorem can now be used to prove that $\|P - K(S, T, 0)\| \leqq cq^{2T}$.

The remaining results in this corollary follow from Theorems 3 and 4. □

To summarize the previous results we have the following theorem.

THEOREM 5. *If* (ST) *and* (DT) *hold, then* $K(S, T, i)$ *converges geometrically in the operator norm as* $T \to \infty$ *to a positive operator* P *that is the unique positive solution to the* ARE. *Also, the control and state generated by* (15) *is* UASL *with respect to the origin and is the unique solution to the regulator problem.*

When the control problem is observable, then any positive solution to the ARE is actually positive definite.

THEOREM 6. *Suppose* P *is a positive solution to the* ARE *and* (OB) *holds. Then* $P > 0$ *and is the unique solution to the* ARE *in the class of positive operators.*

Proof. By Step 1 of Theorem 4, whenever (10) holds, then (27) holds. When the control problem is observable, however, (10) holds all the time so $(y_0, Py_0) \geqq (y(s+1), Py(s+1)) + m\|y_0\|^2 \geqq m\|y_0\|^2$ for some $m > 0$. The fact that P is the unique positive solution to the ARE follows by Corollary 2. □

Now cases are presented where the converse of Corollary 1 holds.

THEOREM 7. *If there exists a positive solution* P *of the* ARE *such that the system* (15) *is* UASL *with respect to the origin, then* (ST) *holds.*

Proof. Define $G = A - B[R + B^*PB]^{-1}B^*PA$ and suppose $\|G^k\| \geqq 1$ for all $k \geqq 0$. Then there exists y_k such that $\|G^k y_k\| > \frac{1}{2}$ and $\|y_k\| = 1$. This contradicts condition (iii) in the definition of UASL and so there exists $r \geqq 0$ with $\|G^r\| < 1$. Now (ST) holds for $L = [R + B^*PB]^{-1}B^*PA$. □

COROLLARY 3. *If there exists a positive solution* P *to the* ARE *and* (DT) *holds, then* (ST) *holds.*

Proof. This follows immediately by Theorems 4 and 7.

THEOREM 8. *If* (CT) *holds, then* (ST) *holds.*

Proof. The solution to the system equation (3) is

$$y(r+1) = A^{r+1}y_0 + \sum_{i=0}^{r} A^i Bu(r-i) = A^{r+1}y_0 + M[u(0), \cdots, u(r)], \tag{46}$$

where M is the linear operator on the controls appearing in the middle of (46). Note that the range space of M contains the range space of MM^* and furthermore the operator MM^* is precisely the operator appearing in (11). Thus MM^* is positive definite and hence there exists a solution $\bar{y}$ to the equation $-A^{r+1}y_0 = MM^*y$. Hence the control sequence $M^*\bar{y}$ inserted in (46) yields $y(r+1) = 0$. From the equation that $\bar{y}$ satisfies and the positive definiteness of MM^*, $a\|\bar{y}\|^2 \leqq (\bar{y}, MM^*\bar{y}) = -(\bar{y}, A^{r+1}y_0) \leqq \|\bar{y}\|\,\|A\|^{r+1}\|y_0\|$ or $\|\bar{y}\| \leqq c\|y_0\|$, where "$a$" is given in (11).

Now choose Q, R to be any positive operators satisfying $R, Q > 0$. Using the control sequence $M^*\bar{y}$ for the controls $\{u(0), \cdots, u(r)\}$ and $u(j) = 0$ for $j > r$ results in $y(j) = 0$ for $j > r$ and the cost function (2) is bounded by $c\|y_0\|^2$ since $\|\bar{y}\| \leqq c\|y_0\|$ and the first $r+1$ controls are given by $M^*\bar{y}$. By Theorem 1, there exists a positive solution of the ARE and since $Q > 0$, then (DT) holds. Corollary 3 completes the proof. □

Remark. It also follows that the steady state Kalman–Bucy filter for the dual estimation problem corresponding to the control problem (2) in the time-invariant case is uniformly asymptotically stable in the large with respect to the origin when (DT) holds and P solves the ARE. The homogeneous part of the Kalman–Bucy filter (presented in Appendix C) is given by

$$\begin{aligned} x(n+1|n+1) &= (A^* - PB[R + B^*PB]^{-1}B^*A^*)x(n|n) \\ &= (A^* - PB[R + B^*PB]^{-1}B^*A^*)^{n+1}x(0|0) \\ &= \{A[A - B(R + B^*PB)^{-1}B^*PA]^n[I - B(R + B^*PB)^{-1}B^*P]\}^*x(0|0), \end{aligned}$$

where the last equation follows by taking the adjoint of the prior equation twice and then regrouping terms. Theorems 4 and 7 imply that

$$\|[A - B(R + B^*PB)^{-1}B^*PA]^k\| < 1$$

for k large enough. Thus it is easy to see that the homogeneous part of the Kalman–Bucy filter is UASL.

5. Necessary and sufficient conditions. The results of the previous section are now tied together in the following theorem.

THEOREM 9. *The following conditions are all equivalent*:

(a) (ST) *and* (DT) *hold.*

(b) *There exists a unique positive solution* P *to the* ARE. *For any* $S \geqq 0$, $K(S, T, i) \to P$ *geometrically in the operator norm as* $T \to \infty$, *and the solution to* (15) *both solves the regulator problem and is* UASL *with respect to the origin.*

(c) *There exists a positive solution to the* ARE *and* (DT) *holds.*

(d) (DT) *holds and* $J(0, T, 0, y) \leqq c\|y\|^2$ *for some* c *independent of* T.

Proof. By Theorem 2, (c) and (d) are equivalent. By Theorem 5, (a) implies (c) and by Corollary (2), (*c*) implies (b). The proof will be complete when it is shown that (b) implies (a).

If (b) holds, then by Theorem 7, (ST) holds. Now suppose (DT) is violated and let P be as given in (b). Then given any ε, T, t, there exists $y(\varepsilon, T, t)$ such that $\|y(\varepsilon, T, t)\| = 1$, $\|A^t y(\varepsilon, T, t)\| > \frac{1}{2}$, and $(y(\varepsilon, T, t), M(T)y(\varepsilon, T, t)) \leqq \varepsilon$, where $M(T) = \sum_{i=0}^{T-1} A^{*i}QA^i$.

Now fix t and define $F(P) = A - B[R + B^*PB]^{-1}B^*PA$. It is easy to see that there exist constants $c, \delta > 0$ depending on P such that $\|F(P) - F(P')\| \leqq c\|P - P'\|$ whenever $\|P - P'\| \leqq \delta$. Let $y(\varepsilon, T, t, i)$ and $y_s(\varepsilon, T, t, i)$ denote the solutions to $y(i+1) = F(K(T, i+1))y(i)$, $y(0) = y(\varepsilon, T, t)$ and $z(i+1) = F(P)z(i)$, $z(0) = y(\varepsilon, T, t)$ respectively.

The error $e(\varepsilon, T, t, i) = y_s(\varepsilon, T, t, i) - y(\varepsilon, T, t, i)$ is the solution $e(i)$ to the equation

$$\begin{aligned} e(i+1) &= F(K(T, i+1))e(i) + [F(P) - F(K(T, i+1))]y_s(\varepsilon, T, t, i) \\ &= \sum_{j=0}^{i} \left(\prod_{k=j+2}^{i+1} F(K(T, k)) \right) \delta F(T, j) y_s(\varepsilon, T, t, j), \end{aligned}$$

where $\delta F(T, i) = F(P) - F(K(T, i+1))$ and $e(0) = 0$.

Since the system (15) is UASL with respect to the origin, and $\|y_s(\varepsilon, T, t, 0)\| = \|y(\varepsilon, T, t)\| = 1$, then $\|y_s(\varepsilon, T, t, i)\|$ is bounded uniformly in ε, T, t, and i. By Theorem 3, (y_0, Py_0) is the optimal cost in the regulator problem and hence $P \geqq K(T, i) \geqq 0$ for $i \leqq T$ and $\|K(T, i)\|$ is bounded uniformly in T and i. Also, note that if $a > 0$ satisfies $R > aI$, then $\|[R + B^*ZB]^{-1}\| \leqq 1/a$ for any positive operator Z and hence $\|F(K(T, i))\|$ is uniformly bounded. Combining these uniform bounds with the fact that t is fixed and $\|\delta F(T, i-1)\| = \|F(K(T, i)) - F(P)\| \leqq c\|K(T, i) - P\| \to 0$ as $T \to \infty$, implies that for T sufficiently large, $\|e(\varepsilon, T, t, t)\| \leqq \frac{1}{8}$ (independent of ε).

Now hold T fixed and consider the following lemma.

LEMMA 2. *Suppose $\mathscr{A}(\cdot, \cdot)$ is a continuous bilinear form on $U \times U$ satisfying $\mathscr{A}(u, u) \geqq a\|u\|^2$ for all $u \in U$ and some $a > 0$ independent of u, and $\mathscr{B}(\cdot)$ is a bounded linear operator. Then the problem to minimize $J(u) = \mathscr{A}(u, u) + \mathscr{B}(u)$ over $u \in U$ has a unique solution $u^* \in U$ and $\|u - u^*\|^2 \leqq (J(u) - J(u^*))/a$.*

Proof of lemma. The existence and uniqueness of u^* and the necessary condition $2\mathscr{A}(u - u^*, u^*) + \mathscr{B}(u - u^*) = 0$ for all $u \in U$ follows by [12]. If $J(u)$ is expanded about u^* and the necessary condition is applied, then the following holds: $J(u) - J(u^*) = \mathscr{A}(u - u^*, u - u^*) \geqq a\|u - u^*\|^2$. □

Note that the control problem (2) satisfies the conditions for the lemma since $(u(k), Ru(k)) \geqq a\|u(k)\|^2$ (where $R > aI > 0$), $(y(k), Qy(k)) \geqq 0$, and the cost functional is a quadratic in $\{u(i)\}$ when $\{y(i)\}$ is expressed in terms of $\{u(i)\}$. Thus if $J(\varepsilon, T, t)$ denotes the optimal cost in (2) when the initial condition is $y(0) = y(\varepsilon, T, t)$ and $S = 0$, and $J_0(\varepsilon, T, t)$ is the cost generated by the control sequence $u(k) = 0$ for $k \geqq 0$ starting from the same initial condition, then the relation $\varepsilon \geqq (y(\varepsilon, T, t), M(T)y(\varepsilon, T, t)) = J_0(\varepsilon, T, t) \geqq J(\varepsilon, T, t) \geqq 0$ implies that $\varepsilon/a \geqq (J_0(\varepsilon, T, t) - J(\varepsilon, T, t))/a \geqq \sum_{i=0}^{T-1} \|u(\varepsilon, T, t, i)\|^2$, where $u(\varepsilon, T, t, i)$ is the optimal control sequence for the control problem (2) corresponding to the initial condition $y(0) = y(\varepsilon, T, t)$. (Recall that the solution to $y(i+1) = F(K(T, i+1))y(i)$, $y(0) = y(\varepsilon, T, t)$, which was labeled $y(\varepsilon, T, t, i)$ above, is also the solution to the control problem (2) and so the notation above for the optimal control is compatible with the notation for the optimal state.)

Let $y_0(\varepsilon, T, t, i)$ denote the solution to $y(i+1) = Ay(i)$, $y(0) = y(\varepsilon, T, t)$. Then the error $e_0(\varepsilon, T, t, i) = y(\varepsilon, T, t, i) - y_0(\varepsilon, T, t, i)$ satisfies the equation $e(i+1) = Ae(i) + Bu(\varepsilon, T, t, i)$, $e(0) = 0$. Using the above bound on the controls implies that for ε sufficiently small, $\|e_0(\varepsilon, T, t, t)\| < \frac{1}{8}$.

To summarize,

$$\begin{aligned}\|y_s(\varepsilon, T, t, t) - y_0(\varepsilon, T, t, t)\| &\leqq \|y_s(\varepsilon, T, t, t) - y(\varepsilon, T, t, t)\| \\ &\quad + \|y(\varepsilon, T, t, t) - y_0(\varepsilon, T, t, t)\| \\ &\leqq \tfrac{1}{8} + \tfrac{1}{8} = \tfrac{1}{4}.\end{aligned}$$

By assumption, $\|y_0(\varepsilon, T, t, t)\| = \|A^t y(\varepsilon, T, t)\| > \frac{1}{2}$. Thus for all t, it is possible to choose T large enough and ε small enough so that $\|y_s(\varepsilon, T, t, t)\| > \frac{1}{4}$. However, this violates the assumption that the system (15) is UASL with respect to the origin (see condition (iii) in the definition of UASL). Hence (DT) must hold. □

6. A gyroscope noise filtering problem. A practical problem motivating the study of operator Riccati equations is the gyroscope noise filtering problem which is described briefly below. The additive noise corrupting gyroscopic output readings are observed experimentally to often possess a $1/f$ behavior in power spectral density over a wide band of frequency. To model this noise as the output of a linear system, a continuum of first order linear systems are used with time constants, r, of the linear systems described by a probability density function $p(r)$. The filtering problem is equivalent via duality to solving the following operator control problem.

The state $y(k) \in Y$ is given by a pair $[b(\cdot, k), a(k)]$, where $a(k)$ is an $m \times 1$ vector and $b(\cdot, k) \in L^2([r_1, r_2])$; i.e.,

$$\int_{r_1}^{r_2} b(r, k)^2 \, dr < \infty.$$

The limits r_1 and r_2 satisfy $0 < r_1 < r_2 < \infty$. The inner product on Y is given by

$$(y_1, y_2) = \int_{r_1}^{r_2} b_1(t) b_2(t) \, dt + a_1^* a_2,$$

where $y_1 = [b_1(\cdot), a_1]$ and $y_2 = [b_2(\cdot), a_2]$. The controls $u(k) \in U$ are scalars and the inner product on U is simply multiplication. The operators A and B in the system dynamics (3) are given by

$$A[b(\cdot), a] = [e^{-z/(\cdot)} b(\cdot), \bar{A}a],$$

$$B[u] = [p(\cdot)u, hu],$$

where $p(\cdot)$ is bounded and measurable, $\bar{A}$ is an $m \times m$ matrix, h is an $m \times 1$ vector, and $z > 0$.

The cost functional is

$$\left[(Sy(T), y(T)) + \sum_{k=0}^{T-1} \int_{r_1}^{r_2} Q(r) b(r, k)^2 \, dr + a(k)^* Q a(k) + u(k)^2 d \right],$$

where $Q \geqq 0$ is an $m \times m$ matrix, $Q(r) \geqq c > 0$ is a bounded measurable function, $d > 0$ is a scalar, and $S \geqq 0$ is a positive semidefinite operator.

Note that this problem is not controllable and, in fact, inserting the operators A and B into the controllability condition (11) results in

$$\sum_{i=0}^{r} \left\{ \int_{r_1}^{r_2} p(r) \, e^{-zi/r} b(r) \, dr \right\}^2 \geqq a \int_{r_1}^{r_2} b(r)^2 \, dr$$

for some $a > 0$ and for all $b \in L^2([r_1, r_2])$. This is clearly impossible (for example, consider a sequence of functions $\{b_j(\cdot)\}$ converging to a delta function). The L^2 part of the system dynamics, however, trivially satisfies the stabilizability condition with $L = 0$ since $e^{-z/r} \leqq e^{-z/r_2} < 1$ for $r_1 \leqq r \leqq r_2 < \infty$. The L^2 part of the system dynamics is also observable for $s = 0$ since $Q(r) \geqq c > 0$. Thus if the linear system $a(k+1) = \bar{A}a(k) + hu(k)$ is stabilizable and the matrix $[\sum_{k=0}^{s} \bar{A}^{*k} Q \bar{A}^k] > 0$ for some s, then all the theorems in § 4 apply.

More details on the gyroscope problem are given in [6].

Appendix A. Stability of the time varying Kalman–Bucy filter and control problem solution. The stability result in Theorem 4 can be generalized to the time-varying case, where

$$y(i+1) = A(i)y(i) + B(i)u(i),$$
$$y(0) = y_0 \in Y,$$
$$u(i) = F(S, T, i)y(i),$$
$$F(S, T, i) = -[R(i) + B^*(i)K(S, T, i+1)B(i)]^{-1}B^*(i)K(S, T, i+1)A(i).$$

Note that since $K(S, T, \cdot)$ is only defined on $[-\infty, T]$, then $y(\cdot)$ is only defined on $[0, T]$, and hence it no longer makes sense to ask whether $y(\cdot)$ is stable. However, (ST′) and (DT′) are sufficient to prove the following properties for $y(y_0, T, \cdot)$, the solution to the system above:

(i) Given $\varepsilon > 0$, there exists $\delta > 0$ such that $\|y_0\| \leqq \delta$ implies that $\|y(y_0, T, i)\| \leqq \varepsilon$ whenever $T \geqq i \geqq 0$ (δ independent of T).

(ii) Given $\delta > 0$, there exists $\varepsilon > 0$ such that $\|y(y_0, T, i)\| \leqq \varepsilon$ whenever $\|y_0\| \leqq \delta$ and $T \geqq i \geqq 0$ (ε independent of T).

(iii) Given $\varepsilon, \delta > 0$, there exists T' such that $\|y(y_0, T, i)\| \leqq \varepsilon$ whenever $\|y(y_0, T, j)\| \leqq \delta$ and $T \geqq i \geqq j + T'$ (T' independent of T).

This is essentially the same as the definition of UASL except that the index for $y(\cdot)$ must be confined to the range $0 \leqq i \leqq T$. The proof of these results is identical to the proof of Theorem 4. Note that the condition (18) holds with $(y(k), Py(k))$ replaced by $(y(k), K(S, T, k)y(k))$. All the steps of Theorem 4 are valid in the time-varying case with $K(S, T, j)$ replacing P. Since a bound was required on $\|P\|$ in various places in the proof, we must now require that $K(S, T, j)$ be bounded uniformly in T and j. Lemma 1, however, shows that when (ST′) holds, $\|K(S, T, j)\|$ is bounded uniformly.

As in the remark at the end of § 4, it follows that the Kalman–Bucy filter for the dual estimation problem corresponding to the control problem is uniformly asymptotically stable in the large with respect to the origin in the time-varying case when (ST′) and (DT′) hold.

Appendix B. (DT) and detectability. We now show that in finite dimensions, (DT) is equivalent to the condition $\rho(A^* - C^*L) < 1$ for some L where $Q = C^*C$. Hautas proves [5] that this last condition is equivalent to requiring that every unstable eigenvector of A is observable, i.e., when e is an eigenvector of A corresponding to the eigenvalue λ and $|\lambda| \geqq 1$, then $Ce \neq 0$.

PROPOSITION B.1. *Every unstable eigenvector of A is observable if and only if* (DT) *holds for $Q = C^*C$.*

Proof. If (DT) holds, $Ae = \lambda e$, $\|e\| = 1$, and $|\lambda| \geqq 1$, then $\|A^t e\| \geqq \|e\|$ and hence by (10),

$$\left(e, \sum_{i=0}^{s} A^{*i}C^*CA^i e\right) = \sum_{i=0}^{s} |\lambda|^{2i}\|Ce\|^2 \geqq b > 0.$$

Thus $Ce \neq 0$.

Conversely, suppose every unstable eigenvector of A is observable. It is easy to see that the nullspace of $M(k) = \sum_{i=0}^{k} A^{*i}C^*CA^i$ is contained in the nullspace of $M(k-1)$. Thus the nullspace is a decreasing function of k and there exists some integer s such that the nullspace is unchanged for $k \geqq s$.

We only treat the case where there is a complete set of normalized eigenvectors $\{e_k\}$ corresponding to the eigenvalues $\{\lambda_j\}$. The changes necessary for defective eigenvalues are summarized at the end of the proof.

If d is any constant satisfying $0 < d < 1$, then the following result is now proved:

(*) There exists an integer $t \geqq 0$ such that whenever $\|A^t y\| \geqq d\|y\|$, then the expansion $y = \sum a_k e_k$ has $a_k \neq 0$ for some unstable eigenvector e_k.

Form the matrix $N = [e_1, e_2, \cdots, e_n]$. Since the $\{e_k\}$ are independent, then N^{-1} exists and hence if $\|y\| = 1$ and $x = (a_1, a_2, \cdots, a_n)^*$ is defined by $x = N^{-1}y$, then $\|x\|^2 = \sum |a_k|^2 \leqq \|N^{-1}\|^2$. Define $a = \|N^{-1}\|$ and choose t large enough so that if λ_k is a stable eigenvalue, then $|\lambda_k|^t < d/(na)$. If $\|y\| = 1$ and $\|A^t y\| \geqq d$, then expanding y in terms of $\{e_k\}$ leads to $\|A^t y\| = \|A^t \sum a_k e_k\| = \|\sum a_k \lambda_k^t e_k\| \geqq d$. Suppose that a_k vanishes for all the unstable eigenvectors. Then the bounds $|a_k| \leqq a$ and $\|e_k\| = 1$ imply that $\|\sum a_k \lambda_k^t e_k\| \leqq \sum |\lambda_k|^t a < d$, where the last inequality follows since the previous sum is only over stable eigenvalues. This is a contradiction, and hence a_k cannot vanish for all the unstable eigenvectors.

Let f be any vector that minimizes $(y, M(s)y)$ over all real vectors satisfying $\|y\| = 1$ and $\|A^t y\| \geqq d$, and suppose that the optimal value of this minimization problem is zero. If it is not zero, then (DT) is immediately satisfied. Recall that a positive semidefinite matrix can be expressed as $D^T D$ so that $(f, M(s)f) = 0$ if and only if $M(s)f = 0$. Thus $M(k)f = 0$ for $k \geqq s$ since the nullspace of $M(k)$ is invariant for $k \geqq s$. Since $\|A^t f\| \geqq d$, then $a_j \neq 0$ for some unstable component in the expansion $f = \sum a_k e_k$. Let λ_j be the eigenvalue of the biggest modulus such that $a_j \neq 0$, and first let us assume that λ_j is real. Then $\lim_{k\to\infty} \lambda_j^{-k} A^k f = e$, where e is an unstable eigenvector (note that any nonzero linear combination of eigenvectors corresponding to a given eigenvalue is also an eigenvector corresponding to the same eigenvalue). Thus $\lim_{k\to\infty} |\lambda_j|^{-2k}(f, A^{*k}C^*CA^k f) = \|Ce\|^2$. Since $M(k)f = 0$ for $k \geqq s$, then $CA^k f = 0$ for $k \geqq s$ and hence $Ce = 0$. This violates the assumption that none of the unstable eigenvectors of A lies in the nullspace of C.

If λ_j occurs in a complex conjugate pair, then $|\lambda_j|^{-k} CA^k f \to C(e^{ik\theta}e + e^{-ik\theta}\bar{e})$ where $\bar{e}$ is the complex conjugate of e. Since $\theta \neq 0, \pi$, then as $k \to \infty$,we conclude that two linearly independent combinations of $\bar{e}$ and e lie in the nullspace of C (i.e., there exists a subsequence k_j of the k's that converges to a vector in the nullspace of C. Then consider $k_j' = k_j + 1$ and extract another convergent subsequence). Hence $Ce = C\bar{e} = 0$ which is again impossible.

We now summarize the changes for the case of defective eigenvalues. Write A in Jordan canonical form as $A = NDN^{-1}$, where D has eigenvalues on the diagonal and either 1's or 0's on the upper subdiagonal. Let $\{e_k\}$ denote the columns of N. The proof above is almost unaltered until the point where it was shown that $Ce = 0$ which violated the condition that the unstable eigenvectors cannot lie in the nullspace of A. Note now that e may no longer be an eigenvector; however, if e_j is not an eigenvector, then one property of the Jordan decomposition

above is that $Ae_j = \lambda_j e_j + e_{j-1}$. Thus $A^k e_j$ for k large enough will have a component which is an unstable eigenvector. The remainder of the proof (which is still very complicated) involves looking at convergent subsequences as above in the case of a complex conjugate pair of eigenvalues. □

Appendix C. The estimation problem. The estimation problem corresponding to the control problem (2) is now presented. Consider the following linear system and observation sequence $\{z(i)\}$:

$$x(i+1) = A^*(i)x(i) + w(i),$$

$$z(i) = B^*(i)x(i) + v(i),$$

where (i) $w(i)$, $x(i) \in Y$, $z(i)$, $v(i) \in U$, (ii) $x(0)$ is a random variable with mean x_0 and covariance Σ_0 satisfying $E[(Fx(0))^2] = F\Sigma_0 F^*$ for any $F: Y \to R$ = the real numbers, (iii) $\{w(i)\}$ and $\{v(i)\}$ are zero mean white noise with covariances $\{Q(i)\}$ and $\{R(i)\}$ satisfying $E[(Fw(i))^2] = FQ(i)F^*$ and $E[(Gv(i))^2] = GR(i)G^*$ for any $F: Y \to R$ and $G: U \to R$ respectively. Also, $x(0)$, $\{w(i)\}$, and $\{v(i)\}$ are assumed uncorrelated.

The estimation problem is to find a sequence of vectors $\{\hat{x}(i|i)\}$ that minimizes $E[(\hat{x}(i|i) - x(i), \hat{x}(i|i) - x(i))]$, where the estimate $\hat{x}(i|i)$ is based on the observations to time i. The Kalman–Bucy filter corresponding to the estimation problem is given by

$$\hat{x}(n+1|n+1) = A^*(n)\hat{x}(n|n) + \Sigma(n+1|n)B(n+1)[R(n+1) + B^*(n+1)$$
$$\cdot \Sigma(n+1|n)B(n+1)]^{-1}[z(n+1) - B^*(n+1)A^*(n)\hat{x}(n|n)],$$

$$\hat{x}(0|0) = x_0,$$

where $\Sigma(n+1|n)$ is generated by

$$\Sigma(n+1|n) = A^*(n)[\Sigma(n|n-1) - \Sigma(n|n-1)B(n)[R(n) + B^*(n)\Sigma(n|n-1)B(n)]^{-1}$$
$$\cdot B^*(n)\Sigma(n|n-1)]A(n) + Q(n),$$

$$\Sigma(0, -1) = \Sigma_0.$$

REFERENCES

[1] R. Boudarel, J. Delmas and P. Guichet, *Dynamic Programing and its Application to Optimal Control*, Academic Press, New York, 1971, pp. 46–48.

[2] P. E. Caines and D. Q. Mayne, *On the discrete time matrix Riccati equation of optimal control*, Internat. J. Control, 12 (1970), pp. 785–794.

[3] J. J. Deyst, Jr. and C. F. Price, *Conditions for asymptotic stability of the discrete minimum-variance linear estimator*, IEEE Trans. Automatic Control, AC-13 (1968), pp. 702–705.

[4] N. Dunford and T. Schwartz, *Linear Operators*, Interscience, New York, 1963.

[5] M. L. J. Hautas, *Stabilization, controllability, and observability of linear autonomous systems*, Indag. Math., 32 (1970), pp. 448–455.

[6] L. L. Horowitz, *Optimal filtering of gyroscopic noise*, Ph.D. thesis, Mass. Inst. of Tech., Cambridge, Mass., 1974.

[7] R. E. Kalman and J. E. Bertram, *Control system analysis and design via the second method of Lyapunov: I. Continuous time systems*, Trans. ASME, J. Basic Engrg., 82 (1960), pp. 371–393.

[8] ———, *Control system analysis and design via the second method of Lyapunov: II. Discrete time systems*, Ibid., 82 (1960), pp. 394–400.

[9] K. Y. Lee, S. N. Chow and R. O. Barr, *On the control of discrete-time distributed parameter systems*, this Journal, 10 (1972), pp. 361–376.
[10] J. Zabczyk, *Remarks on the control of discrete-time distributed parameter systems*, this Journal, 12 (1974), pp. 721–733.
[11] S. K. Berberian, *Introduction to Hilbert Space*, Oxford University Press, New York, 1961.
[12] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.