N91-22316

Optimal Controllers for Finite Wordlength Implementation

K. Liu, R. Skelton

Purdue University

West Lafayette, IN 47907

ABSTRACT

When a controller is implemented in a digital computer, with A/D and D/A conversion, the numerical errors of the computation can drastically affect the performance of the control system. There exists realizations of a given controller transfer function yielding arbitrarily large effects from computational errors. Since, in general, there is no upper bound, it is important to have a systematic way of reducing these effects. Optimum controller designs are developed which take account of the digital round-off errors in the controller implementation and in the A/D and D/A converters. These results provide a natural extension to the LQG theory since they reduce to the standard LQG controller when infinite precision computation is used. But for finite precision the separation principle does not hold.

I. INTRODUCTION

LQG controllers are normally designed under the assumption that computer implemention will be perfect (this is the infinite wordlength assumption for state variable computation). However, real control systems are subject to the effects of finite wordlength computation. These round-off errors should not be ignored in the design of the controller. The influence of these errors on the control system and the optimum controller design considering their effects are the subjects of this paper.

We consider the problems that arise with fixed-point arithmetic and the finite word length of digital computers. This paper was motivated by the work of Kladiman and Williamson [1989]. Mullis and Roberts [1976] and Hwang [1977] in the field of signal processing first revealed the fact that the influence of round-off errors on digital filter performance depends on the realization chosen for the filter implementation. To minimize round-off errors these papers suggest a special coordinate transformation T prior to filter (or controller) synthesis.

This is in stark contrast to frequency domain approaches to control, which regard as irrelevant (and hence is completely ignored) the state space realization of the controller transfer function.

The idea of applying a coordinate transformation prior to controller synthesis has been applied to Kalman filter and LQG controller design problems, Williamson [1985], Kladiman and Williamson [1989]. One may select the wordlength of the computer to insure that the resulting degradation in the performance from round-off error is less than a certain percentage of the ideal behavior of the standard Kalman filter or LQG controller without round-off error. This approach was adapted by Sripad [1981] in the design of Kalman filters, and later by Moroney, et. al [1983] for LQG controller design. In these papers the standard Riccati equations are solved, *followed* by a coordinate transformation to reduce the effects of round-off errors. We shall call these controllers LQG_T to indicate a standard LQG controller followed by an "optimal" coordinate transformation T. This transformation *depends* on the control gains, hence, we put the word optimal above in quotes, because the standard LQG gain is not the optimal gain for the round-off error problem. The optimum solution is to design the controller which *directly* takes into account the round-off errors associated with a finite word length implementation, rather than merely performing a coordinate transformation T on the LQG controller *after* it is designed. The optimal state estimation problem was solved by Williamson [1985]. This leads to a modified Kalman filter. The problem of optimum LQG controller design in the presence of round-off error was studied by Kadiman and Williamson [1989]. This paper worked with upper bounds and numerical results showed improvement over earlier work, but their algorithm does not provide the necessary conditions for an optimal solution. This paper provides the necessary conditions and a controller design algorithm for the solution of this problem. We shall call this controller LQG_{FW}.

With a fixed point implementation, the states of the LQG_{FW} controller are properly scaled to reduce the possibility of overflow. There are many scaling criteria available. The method we shall use is the variance oriented procedure, l_2-norm scaling [Hwang 1977]. We assume round-off errors are additive. This tends to be supported by the literature on state quantization, whereas quantization of coefficients leads to multiplicative errors [Williamson 1985].

The organization of the paper is as follows. In Section 2, the problem of LQG controller design in the presence of round-off errors is formulated. The importance of the coordinates of the controller will be discussed in Section 3. Section 3 summarizes the needed results from [Kadiman and Williamson 1989], and our new results on upperbounds of finite wordlength effects. It is shown that the portion of the LQG cost contributed by these errors will range from arbitrarily large to an achievable lower bound with the variation of the realization of the controller (variation of the choice of coordinates). The coordinate achieving the lower bound is described. In Section 4, the optimization problem is discussed in terms of chosing both the controller parameter matrices and the realization coordinate simultaneously. The necessary conditions are derived for the optimization problem. An algorithm is then presented for the designs of the optimal LQG_{FW} controller. The standard LQG and the LQG_{FW} controller are compared in Section 5. Some conclusions appear in Section 6.

II. Round-Off Error and LQG Controller Design Problem

In this section, we formulate the LQG controller design problem when round-off errors are present. The formulation procedure follows the original ideas of Mullis [1976], Hwang [1977] and the ideas of Williamson [1985], Kadiman and Williamson [1989]. Let us assume, for the study of round-off error, the discrete controller is designed from a discrete model of the plant to be controlled. We then introduce a model for finite wordlength effects into the discrete design problem.

Considering the following discrete-time model of a time-invariant plant:

$$\begin{cases} x_p(k+1) &= A_p x_p(k) + B_p u(k) + D_p w_p(k) \\ z_p(k) &= M_p x_p(k) + v_p(k) \\ y_p(k) &= C_p x_p(k) \end{cases} \tag{1}$$

where x_p is the state n_p–vector, u, y_p and z_p are the control n_u–vector, output n_y–vector, measurement n_z–vector, v_p and w_p are assumed to be mutually independent, zero mean, discrete white Gaussian noises with covariance matrices V_p and W_p, respectively.

The controller that one might <u>desire</u> to implement is described by following equations:

$$\begin{cases} x_c(k+1) &= A_c x_c(k) + B_c z_p(k) \\ u(k) &= C_c x_c(k) + D_c z_p(k) \end{cases} \tag{2}$$

where x_c is the controller state n_c–vector, u and z_p are the control and measurement vectors described in the plant model. In a finite wordlength digital computer, the controller state x_c and measurement variable z_p will be quantized at each time of computation. Considering the quantization process, computation (1) and (2) cannot be accomplished. Instead the computation is described by

$$\begin{cases} x_p(k+1) = A_p x_p(k) + B_p Q[u(k)] + D_p w_p(k) \\ z_p(k) = M_p x_p(k) + v_p(k) \\ y_p(k) = C_p x_p(k) \end{cases} \tag{3a}$$

$$\begin{cases} x_c(k+1) = A_c Q[x_c(k)] + B_c\, Q[z_p(k)] \\ u(k) = C_c Q[x_c(k)] + D_c\, Q[z_p(k)] \end{cases} \tag{3b}$$

where $Q[\cdot]$ stands for the quantization process. Assuming an additive property of the round-off error, we can model the quantization process by:

$$Q[u(k)] = u(k) + e_u(k) \qquad \text{D/A} \tag{4a}$$

$$Q[x_c(k)] = x_c(k) + e_x(k) \qquad \text{control computer} \tag{4b}$$

$$Q[z_p(k)] = z_p(k) + e_z(k) \qquad \text{A/D} \tag{4c}$$

where e_u is the round-off error resulting from D/A conversion, $e_x(k)$ is the error resulting from quantization and $e_z(k)$ is the error resulting from A/D conversion. We do not claim that this assumption is always justified, but we invoke this common assumption in this paper, since one cannot *optimize* with respect to coefficient errors directly. One can only *evaluate* designs with respect to coefficient errors. There are many such evaluations in filter theory, and we shall add our own numerical evaluation in this paper. All such evidence points to a conclusion that controller structures that are good with respect to state quantization tend to also be good with respect to coefficient quantization.

It was shown [Sripad 1977] that, under sufficient excitation conditions, the round-off error $e_x(k)$ can be modeled as a zero mean, white noise independent of $w_p(k)$ and $v_p(k)$, with covariance matrix E_x,

$$E_x = qI, \quad q \triangleq \frac{1}{12} 2^{-2\beta} , \tag{5a}$$

where β is the wordlength of the control computer. Similarly, we assume the D/A conversion error $e_u(k)$ and the A/D conversion error $e_z(k)$ to be zero mean, mutually independent white noise and also independent of $w_p(k)$, $v_p(k)$ and $e_x(k)$ with covariance matrices E_u and E_z,

$$E_u = q_u I, \quad q_u \triangleq \frac{1}{12} 2^{-2\beta_u} \tag{5b}$$

$$E_z = q_z I, \quad q_z \triangleq \frac{1}{12} 2^{-2\beta_z} \tag{5c}$$

where β_u and β_z are the wordlengths of D/A and A/D converters. Substitute (4) into (3) to obtain a closed-loop system model including finite wordlength effects,

$$\begin{cases} x_p(k+1) = A_p x_p(k) + B_p u(k) + D_p w_p(k) + B_p e_u(k) \\ z_p(k) = M_p x_p(k) + v_p(k) \\ y_p(k) = C_p x_p(k) \end{cases} \tag{6a}$$

$$\begin{cases} x_c(k+1) = A_c x_c(k) + B_c z_p(k) + A_c e_x(k) + B_c e_z(k) \\ u(k) = C_c x_c(k) + D_c z_p(k) + C_c e_x(k) + D_c e_z(k) \end{cases} \tag{6b}$$

We seek the controller to minimize the following cost function

$$J = \lim_{k \to \infty} E\{y_p^*(k) Q_p y_p(k) + u^*(k) R u(k)\} \tag{7}$$

where u and y_p are again control and output vectors, and Q_P and R are the weighting matrices.

After combining (6a) and (6b), and using the following notation for the vectors and matrices:

$$x(k) = \begin{bmatrix} x_p(k) \\ x_c(k) \end{bmatrix}; \; y(k) = \begin{bmatrix} y_p(k) \\ u(k) \end{bmatrix}; \; A = \begin{bmatrix} A_p & 0 \\ 0 & 0 \end{bmatrix}, \; B = \begin{bmatrix} B_p & 0 \\ 0 & I \end{bmatrix}, \; C = \begin{bmatrix} C_p & 0 \\ 0 & 0 \end{bmatrix}$$

$$D = \begin{bmatrix} D_p \\ 0 \end{bmatrix}; \; G = \begin{bmatrix} D_c & C_c \\ B_c & A_c \end{bmatrix}; \; I_0 = \begin{bmatrix} 0 & 0 \\ I & 0 \end{bmatrix}; \; I_1 = \begin{bmatrix} I \\ 0 \end{bmatrix}; \; I_2 = \begin{bmatrix} 0 \\ I \end{bmatrix};$$

$$M = \begin{bmatrix} M_p & 0 \\ 0 & I \end{bmatrix}, \; Q = \begin{bmatrix} Q_p & 0 \\ 0 & R \end{bmatrix}$$

the closed-loop system is compactly described by

$$
\begin{aligned}
x(k+1) &= [A + BGM]x(k) + Dw_p(k) + BGI_1 v_p(k) + BGI_2 e_x(k) + BGI_1 e_z(k) + BI_1 e_u(k) \\
y(k) &= [C + I_0 GM]x(k) + I_0 GI_1 v_p(k) + I_0 GI_2 e_x(k) + I_0 GI_1 e_z(k)
\end{aligned}
\tag{9}
$$

and the cost function (7) may be written

$$
J = \lim_{k \to \infty} E\{y^*(k) Q y(k)\} . \tag{10}
$$

Now, substitute (9) into (10), since $e_u(k)$, $e_x(k)$, $e_z(k)$, $w_p(k)$, and $v_p(k)$ are mutually independent,

$$
\begin{aligned}
J = \operatorname{tr}\{X[C + I_0 GM]^* Q[C + I_0 GM]\} + \operatorname{tr}\{V_p (I_0 GI_1)^* Q(I_0 GI_1)\} \\
+ \operatorname{tr}\{E_x (I_0 GI_2)^* Q I_0 GI_2)\} + \operatorname{tr}\{E_z (I_o GI_1)^* Q(I_o GI_1)\}
\end{aligned}
\tag{11a}
$$

where X is the state covariance satisfying:

$$
\begin{aligned}
X = [A + BGM]X[A + BGM]^* + DW_p D^* + (BGI_1)V_p(BGI_1)^* \\
+ (BGI_2)E_x(BGI_2)^* + (BGI_1)E_z(BGI_1)^* + BI_1 E_u (BI_1)^*
\end{aligned}
\tag{11b}
$$

We can decomposite J in eqn. (11a) into two terms:

$$
J = J_{wv} + J_e \tag{12a}
$$

where

$$
J_{wv} \triangleq \operatorname{tr}\{X_1 [C + I_0 GM]^* Q[C + I_0 GM]\} + \operatorname{tr}\{(V_p + E_z)(I_0 GI_1)^* Q(I_0 GI_1)\} \tag{12b}
$$

$$
X_1 = [A + BGM]X_1 [A + BGM]^* + DW_p D^* + (BGI_1)(V_p + E_z)(BGI_1)^* + BI_1 E_u (BI_1)^* \tag{12c}
$$

and

$$
J_e \triangleq \operatorname{tr}\{X_e [C + I_0 GM]^* Q[C + I_0 GM]\} + \operatorname{tr}\{E_x (I_0 GI_2)^* Q(I_0 GI_2)\} \tag{12d}
$$

$$
X_e = [A + BGM]X_e [A + BGM]^* + (BGI_2)E_x(BGI_2)^* \tag{12e}
$$

where $X = X_1 + X_e$. J_{wv} is the portion of the performance index contributed by disturbances $e_u(k)$, $e_z(k)$, $w_p(k)$ and $v_p(k)$. J_e is the portion contributed solely by round-off error $e_x(k)$.

To prevent the overflow in controller state variable computation, we must properly scale the state variables. We use the l_2-norm scaling procedure which is written as:

$$[X_1(2, 2)]_{ii} = s \quad i = 1, \cdots, n_c \tag{13}$$

where $X_1(2, 2)$ is the (2.2) subblock matrix of X_1 matrix (the controller subblock), and $[\cdot]_{ii}$ stands for the ith diagonal element of the matrix. Equation (13) requires that the controller state variables have variance equal to s when the closed-loop system is excited only by outside disturbance and measurement noise. We call (13) the scaling constraint.

Therefore, the optimization problem is

$$\min_G J = \min_G \, (J_{wv} + J_e) \ , \tag{14}$$

subject to (12-13).

III. Contribution of Round-Off Error to the LQG Performance Index

In this section, we discuss the J_e term in (12a) and defined by eqn. (12d) which is the portion of the LQG cost function contributed by round-off errors. This portion of the cost function is coordinate dependent. It is unbounded from above, (that is, it can be arbitrarily large), but it has an achievable lower bound, which can be achieved in an optimal coordinate. The lower bound result was obtained by [Moroney et. al. 1983] and [Kadiman and Williamson 1989]. The construction of this optimal coordinate is discussed in this section, where we assume G is some given matrix (we shall optimize G later).

We will first present three key lemmas, which form the basis for the results of this section.

Lemma 1. [Mullis and Roberts 1976, Hwang 1977].

Given any $n \times n$ *matrix* M, *there exist a (non-unique) unitary matrix* U *such that* $(UMU^*)_{jj} = s$ *for all j, if and only if* $\mathrm{tr}(M) = sn$

□

Lemma 2. [well known]

For any two positive definite matrices P *and* Q, *let* $\lambda_i\,[\cdot]$ *denote the* i^{th} *eigenvalue of matrix* $[\cdot]$. *Then,*

a) $\lambda_i[QP] > 0$ for all i

b) The $\lambda_i[QP]$ *are invariant under the transformation* $\tilde{P} = TPT^*$ *and* $\tilde{Q} = T^{-*}QT^{-1}$ *where* T *is nonsingular.*

□

Lemma 3.

Let a scalar J be defined by

$$J \triangleq \mathrm{tr}\{TT^*P\} \tag{15a}$$

where the $n_p \times n_p$ *nonsingular matrix* T *is* constrained by

$$(T^{-1}T^{-*})_{ii} = s \quad \text{for all } i \tag{15b}$$

and P *is a positive definite matrix. Then over the set of all nonsingular matrices* T *constrained by* (15b),

a) J *is not bounded from above.*

b) J *is bounded from below* ($J \geq \underline{J}$) *by*

$$\underline{J} \triangleq \frac{1}{sn_p}[\mathrm{tr}(\sqrt{P})]^2 \tag{16a}$$

where

$$P = \sqrt{P}\,\sqrt{P} \tag{16b}$$

and $\sqrt{P}$ *is symmetric.*

c) $\underline{J}$ *in* (16a) *is achievable by the matrix* T*:*

$$T = \underline{T} \triangleq U_t \Pi_t V_t^* \tag{17a}$$

where U_t, V_t *are unitary,* Π_t *diagonal, satisfying*

$$U_t \Pi_t^{-2} U_t^* = \frac{sn_p\sqrt{P}}{\mathrm{tr}(\sqrt{P})} \tag{17b}$$

$$[V_t \Pi_t^{-2} V_t^*]_{ii} = s \ \text{ for all } i\ . \tag{17c}$$

□

Statements b) and c) are minor modifications of the results obtained by [Mullis and Roberts 1976] and [Hwang, 1977]. The proof of a) appears in Appendix A. An algorithm for solving (17b), (17c) is given in Appendix B.

The contribution of finite wordlength error in the cost function is described by equations (12d) and (12e). This J_e term can also be written as:

$$J_e = tr\{K_e(BGI_2)E_x(BGI_2)^*\} + tr\{E_x(I_0GI_2)^*Q(I_0GI_2)\} \tag{18a}$$

$$K_e = [A + BGM]^*K_e[A + BGM] + [C + I_0GM]^*Q[C + I_0GM] \ . \tag{18b}$$

Since $E_x = qI$, we then have:

$$J_e = q\,tr\{(BGI_2)^*K_e(BGI_2) + (I_0GI_2)^*Q(I_0GI_2)\} \ . \tag{19}$$

We can easily check that the (2, 2)th subblock matrix of K_e (the controller subblock $K_e(2, 2)$) satisfies:

$$K_e(2,\ 2) = (BGI_2)^*K_e(BGI_2) + (I_0GI_2)^*Q(I_0GI_2) \ . \tag{20}$$

Substituting (20) into (19) reduces (19) to

$$J_e = q tr[K_e(2,\ 2)] \ .$$

Hence, the minimization of J_e reduces to the problem:

$$\min J_e, \quad J_e = q tr\{K_e(2,\ 2)\} \tag{21}$$

subject to (18b), (13) and (12c). From the singular value decompositions

$$X_1(2,\ 2) = U_x^*\Sigma_x U_x \tag{22a}$$

$$\Sigma_x^{1/2}U_xK_e(2,\ 2)U_x^*\Sigma_x^{1/2} = U_k^*\Sigma_k U_k \tag{22b}$$

then U_x, U_k are unitary, Σ_x, Σ_k are diagonal and

$$\Sigma_k \triangleq \mathrm{diag}\,\{\dots\ \lambda_i[K_e(2,\ 2)X_1(2,\ 2)]\ \dots\}\ . \tag{22c}$$

Suppose we begin our study with the closed-loop coordinate transformation T as:

$$T = \begin{bmatrix} I & 0 \\ 0 & U_x^* \Sigma_x^{1/2} U_k^* \end{bmatrix} . \tag{23}$$

Then, after this coordinate transformation as suggested by Kadiman and Williamson [1989]:

$$\overline{X}_1(2,\ 2) = (U_x^* \Sigma_x^{1/2} U_k^*)^{-1} X_1(2,\ 2)(U_x^* \Sigma_x^{1/2} U_k^{-*})^{-*} = I \tag{24}$$

$$\overline{K}_e(2,\ 2) = (U_x^* \Sigma_x^{1/2} U_k^*)^* K_e(2,\ 2)(U_x^* \Sigma_x^{1/2} U_k^*) = \Sigma_k\ . \tag{25}$$

If we take one more controller coordinate transformation T_c, the index J_e and its constraint equations, (after we substitute (24) and (25) into (13) and (21)), become

$$J_e = q\mathrm{tr}[T_c T_c^* \Sigma_k] \tag{26a}$$

$$[T_c^{-1} T_c^{-*}]_{ii} = s, \quad i = 1,\ \cdots,\ n_c\ . \tag{26b}$$

Since, from Lemma 2, Σ_k in (22c) is coordinate independent, we may ignore the K_e and X_1 calculations (18b) and (12c) and concentrate on T_c in (26). Then, by applying Lemma 3 on equation (26), we have following theorem.

Theorem 1. *The round-off error term* J_e *in the LQG performance index* (12d) *and* (12e), *and constrained by the scaling constraint eqn.* (12c), (13), *is controller coordinate dependent. It is* unbounded from above *when the realization coordinate varies arbitrarily. It is* bounded from below *by the following lower bound:*

$$\underline{J}_e = \frac{q}{sn_c}\,\mathrm{tr}\,\Sigma_k \tag{27}$$

The lower bound is achieved by the following controller coordinate transformation:

$$\underline{T}_c = U_x^* \Sigma_x^{1/2} U_k^* U_t \Pi_t V_t^* \tag{28a}$$

where U_x, U_k, U_t, V_t *are unitary matrices,* Σ_x, Π_t *are diagonal matrices, subject to the* *constraints:*

$$X_1(2,2) = U_x^* \Sigma_x U_x \tag{28b}$$

$$\Sigma_x^{1/2} U_x K_e(2,2) U_x^* \Sigma_x^{1/2} = U_k^* \Sigma_k U_k \tag{28c}$$

$$U_t \Pi_t^{-2} U_t^* = \frac{s n_c \, \Sigma_k}{\mathrm{tr}\Sigma_k} \tag{28d}$$

$$[V_t \Pi_t^{-2} V_t^*]_{ii} = s, \qquad i = 1, \cdots, n_c \quad . \tag{28e}$$

□

To find the optimal coordinate transformation $\underline{T}_c$ in (28a), we must solve (28d), (28e) to obtain U_t, Π_t, V_t. The equations (28d), (28e) are, however, special cases of (17b), (17c), where P is the diagonal matrix Σ_k. An algorithm is given in Appendix B to compute the U_t, Π_t, V_t needed for (28a).

The conclusion of this section is that the problem $\min\limits_{T_c} J_e$ is solved by the coordinate transformation given by (28a).

IV. LQG Controller Design in the Presence of Round-Off Errors

As discussed in Section II, when round-off error is present, the LQG performance index can be decomposed into two terms. One term contains the influence of disturbance and measurement noise, the other term is contributed by round-off errors. Although the first term is not influenced by the coordinate of the controller, the second term is critically dependent on the coordinate. An optimal coordinate transformation is given by (28a). With the scaling requirement of the controller state variables to prevent overflow, we have a different optimization problem now for controller design comparing to the original optimal control design problem without round-off errors. In this section, we will discuss the controller design.

Let us first present a useful result.

Lemma 4. *Suppose* $J_{kx} \triangleq \sum_{l=1}^{n} \sqrt{\lambda_l[K(i,\, i)X(j,\, j)]}$ *where* $K(i, i)$ *and* $X(j, j)$ *are the* (i, i)th *subblock of K and* (j, j)th *subblock of* X *respectively. Define*

$$\nabla_k J_{kx} \triangleq \frac{\partial}{\partial K} J_{kx}\,, \qquad \nabla_x J_{kx} \triangleq \frac{\partial}{\partial X} J_{kx}$$

then:

a) $\nabla_k J_{kx}(p,q) = 0$ *when* $p \neq i$ *or* $q \neq i$ (29a)

$$\nabla_k J_{kx}(p,q) = \frac{1}{2}\sum_{l=1}^{n} \frac{[E^{-1}(i,\, j)]^{*}_{l\text{th-row}}[E(i,\, j)]^{*}_{l\text{th-col}}X(j,\, j)}{\sqrt{\lambda_l[K(i,\, i)X(j,\, j)]}} \quad \textit{when } p = i \textit{ and } q = i \tag{29b}$$

b) $\nabla_x J_{kx}(p, q) = 0$ *when* $p \neq j$ *or* $q \neq j$ (29c)

$$\nabla_x J_{kx}(p,\, q) = \frac{1}{2}\sum_{l=1}^{n} \frac{K(i,\, i)[E^{-1}(i,\, j)]^{*}_{l\text{th-row}}[E(i,\, j)]^{*}_{l\text{th-col}}}{\sqrt{\lambda_l[K(i,\, i)X(j,\, j)]}} \quad \textit{when } p = j \textit{ and } q = j \tag{29d}$$

where $\nabla_k J_{kx}(p, q)$ *and* $\nabla_x J_{kx}(p, q)$ *are the* (p, q)th *subblock of* $\nabla_k J_{kx}$ *and* $\nabla_x J_{kx}$, $E(i, j)$ *is the eigenvector matrix of matrix* $K(i, i)X(j, j)$

□

The proof of the lemma is given in Appendix A.

The LQG controller design problem, when finite wordlength effects are taking into account, are described by the equations (12-14). This is denoted as the LQG_{FW} controller. However, the scaling constraint (13) can be always satisfied by properly choosing the coordinates of the controller, so the problem breaks up into two parts: Finding G and finding its optimal coordinate transformation T_c to satisfy (12), (13) and (14). On the strength of Section 3, we can therefore write the optimization problem as

$$\min_{G,T_c} J = \min_{G,T_c} (J_{wv} + J_e) = \min_{G}[\min_{T_c} (J_{wv} + J_e)]$$

since J_{wv} is constant in terms of the variation of T_c, we have

$$\min_{G,T_c} J = \min_{G} [J_{wv} + \min_{T_c} J_e] \tag{30}$$

Assume $\underline{J}_e \triangleq \min_{T_c} J_e$ is given by (27), from Theorem 1. Hence, the equivalent LQG_{FW} design problem becomes

$$\min_{G}[J_{wv} + \underline{J}_e] \ , \tag{30a}$$

subject to (12c) and (18b) where

$$J_{wv} = \mathrm{tr} X_1 (C + I_0 GM)^* Q (C + I_0 GM) + \mathrm{tr}(V_p + E_z)(I_0 G I_1)^* Q (I_0 G I_1) \tag{30b}$$

$$\underline{J}_e = \frac{q}{sn_c} (\mathrm{tr}\, \Sigma_k)^2 \tag{30c}$$

where Σ_k is defined by (22c), and the transformation T_c which yields $\underline{J}_e$ is given by the algorithm in Appendix B, and may be computed only after the optimal G is obtained from (30). The following theorem states the necessary conditions of the optimization problem (30).

Theorem 2:

Necessary conditions for G *to be the solution of the optimal controller design problem* (30) *are:*

$$[A+BGM]X_1[A+BGM]^* + DW_pD^* + (BGI_1)(V_p+E_z)(BGI_1)^* + BI_1E_u(BI_1)^* - X_1 = 0 \quad (31a)$$

$$[A+BGM]^*K_e[A+BGM] + [C+I_0GM]^*Q[C+I_0GM] - K_e = 0 \quad (31b)$$

$$[A+BGM]^*K_2[A+BGM] + [C+I_0GM]^*Q[C+I_0GM] - K_2 + \nabla_x = 0 \quad (31c)$$

$$[A+BGM]K_3[A+BGM]^* - K_3 + \nabla_k = 0 \quad (31d)$$

$$(I_0^*QI_0 + B^*K_2B)G(MX_1M^* + I_1(V_p+E_z)I_1^*) + (I_0^*QI_0 + B^*K_eB)GMK_3M^* + \\ + B^*(K_2AX_1 + K_eAK_3)M^* = 0 \quad (31e)$$

where ∇_x *has 4 subblocks as*

$$\nabla_x(i,j) = 0 \qquad i \neq 2 \text{ or } j \neq 2$$

$$\nabla_x(2,2) = \frac{q}{sn_c} \operatorname{tr} \Sigma_k \left\{ \sum_{i=1}^{n_c} \frac{K_e(2,2)[E^{-1}]_{irow}^{-1*}[E]_{icol}^*}{\sqrt{\Sigma_{k_{ii}}}} \right\}$$

and ∇_k *also has* 4 *subblocks as*

$$\nabla_k(i,j) = 0 \qquad i \neq 2 \text{ or } j \neq 2$$

$$\nabla_k(2,2) = \frac{q}{sn_c} \operatorname{tr} \Sigma_k \left\{ \sum_{i=1}^{n_c} \frac{[E^{-1}]_{irow}^*[E]_{icol}^* X_1(2,2)}{\sqrt{\Sigma_{k_{ii}}}} \right\}$$

where E *is the matrix of eigenvectors of the matrix* $K_e(2,2)\, X_1(2,2)$.

□

The proof of theorem 2 is given in Appendix A.

Remark 1: The only terms in (31) which are affected by q are the two terms in (31c) and (31d) denoted by ∇_x, ∇_k. Hence setting $\beta = \infty$ gives q = 0, $\nabla_k = 0$, $\nabla_x = 0$, $K_3 = 0$, $K_2 = K_e$. Hence, eqs. (31) reduce to the standard LQG design by setting $\beta = \infty$. In this case, the 11 block of (31a) reduces to the Kalman filter Riccati equation, and the 22 block of (31c) reduces to the control Riccati equation.

Remark 2: We shall denote the controller satisfying (31) as the $\overline{LQG_{FW}}$ controller to indicate that the LQG_{FW} controller requires an additional step; the computation of $\underline{T}_c$ from Appendix B. Now, we have following LQG_{FW} controller design algorithm:

The LQG_{FW} Algorithm

Step 1: Solve G from equations (31a)-(31e). This gives the $\overline{LQG_{FW}}$ controller.

Step 2: Compute $\underline{T}_c = U_x^* \Sigma_x^{1/2} U_k^* U_t \Pi_t V_t^*$ by solving $U_x, \Sigma_x, U_k, U_t, \Pi_t, V_t$ from (28b)-(28e), using the G obtained in Step 1.

Step 3: $\tilde{G} = \begin{bmatrix} I & 0 \\ 0 & \underline{T}_c^{-1} \end{bmatrix} G \begin{bmatrix} I & 0 \\ 0 & \underline{T}_c \end{bmatrix}$ is the optimal LQG_{FW} controller for implementation.

□

Remark: A natural algorithm to suggest in Step 1 is as follows. Suppose one desires to design a LQG_{FW} controller for 10 bit arithmetic.

(i) Solve (31a)-(31e) for $\beta_i = \infty$, (hence, the standard LQG controller).

(ii) On the next iteration set β_i = 32 (or whatever gives a reasonably small number for ∇_x, ∇_k.

(iii) Iterate by indexing β_i. Change β_i by no more than one bit on each iteration. This gives an "answer" in 32-10 = 22 iterations (but this manner of choosing step sizes in not guaranteed to be sufficient to yield the optimal answer).

This is a "natural" homotopy method, since β is a natural choice for a homotopy parameter.

V. Computation Examples

We consider an Euler Bernoulli beam modeled by its first 5 bending modes with 2 inputs and 2 outputs. The modal frequencies appear in TABLE 1. In discrete controller design, the discrete model is represented by the matrices $\{A_p, B_p, C_p, D_p, M_p, W_p, V_p\}$ in equation (1). These matrices are given in Appendix C for a uniform sample time $\Delta t = 0.018$ sec. The LQG cost function is given by equation (7) with

$$Q_p = 0.99I \qquad R = 0.01I \ .$$

The wordlength of the control computer is assumed to be 4 bits. Since the effects of D/A and A/D conversion errors on the control system simply modify the effects of system disturbance and measurement noise, we ignore these errors in the example. Both the standard LQG controller and the LQG_{FW} controller are computed for the system.

	Frequency	Damping Factor
Mode 1	3.4987e+00	9.9994e-03
Mode 2	1.3995e+01	2.1301e-02
Mode 3	3.1488e+01	4.5600e-02
Mode 4	5.5979e+01	8.0400e-02
Mode 5	8.7468e+01	1.2530e-01

TABLE 1. Frequencies and Damping Factors of the Euler-Bernoulli Beam Example

The standard LQG controller of course was designed without consideration of round-off errors ($\beta = \infty$) and is labeled controller "LQG" in the TABLES. Controllers denoted "LQG_{Ti}" $i = 1, \cdots, 4$ are the same as the LQG, but for a coordinate transformation on the controller after G is computed. The matrices $\{A_c, B_c, C_c, D_c\}$ associated with the LQG_{T1} controller are shown in Appendix C. In different coordinates T_i, TABLE 2 shows the finite wordlength contribution J_e in the closed-loop system cost, using the standard LQG controller. In the optimal coordinate T_1 (controller LQG_{T1}) the cost J_e is about 500 times smaller than the cost in the original coordinate design (controller LQG). This improvement is equivalent to increasing the wordlength of the control computer by about 5 bits ($5 = \frac{1}{2} \log_2 500$). The effect of computational errors J_e in two commonly used coordinates, Normalized Observable Hessenberg Coordinates [Skelton 1988] and Phase Variable Coordinates, are also given in TABLE 2. The fact that Phase Variable Coordinates are bad for computation is consistent with other findings in filter synthesis [Williamson 1990]. The extreme high costs of the controller in a particular coordinate (LQG_{T4}) in TABLE 2 serves only to demonstrate that the cost J_e can become unbounded for some coordinates. The choice of coordinate T_4 was rather arbitrary and will not be described or discussed further.

Controller	Controller Coordinates	Cost J_e
LQG_{T1}	Optimal	9.793
LQG_{T2}	Normalized Obs. Hess.	2.692×10^2
LQG	Plant Coordinates	4.862×10^3
LQG_{T3}	Phase Variable	9.486×10^3
LQG_{T4}	Coordinate "X"	1.472×10^8

TABLE 2. Standard LQG Controller in
Different Coordinates

The LQG_{FW} controller was designed by the LQG_{FW} algorithm given in Section 4. The controller matrices $\{A_c, B_c, C_c, D_c\}$ of this controller also appear in Appendix C. TABLE 3 shows the computed costs of the standard LQG controller, the transformed LQG controller (LQG_{T1}), and the LQG_{FW} controller (The "LQG_{FW} with coefficient error" will be discussed later). The costs for three different groups of excitations are computed in each case. The applicable disturbances for J, J_y, and J_u include plant disturbance w, sensor noise v, and finite wordlength error e. The applicable disturbance for J_e, J_{ey}, J_{eu} is only e, and for J_{wv}, J_{wvy}, J_{wvu} are only w_p and v_p (no finite wordlength effects). Hence, these sums apply to the various cost decompositions; J_y is the output term of J (the total cost), J_u is the control term in J, hence $J = J_y + J_u$. J_{wvy} is the output term of J_{wv} (the contribution of v_p and w_p in J), where $J_y = J_{wvy} + J_{ey}$ and $J_e = J_{ey} + J_{eu}$, $J = J_{wv} + J_e$. J_{wvu} is the control term of J_{wv} and $J_u = J_{wvu} + J_{eu}$. As we can

Disturbances Applied	Costs	LQG Controller	LQG_{T1} Controller	LQG_{FW} Controller	LQG_{FW} with coeff. errors
All v, w, and e	J	4.8827e+03	3.0589e+01	2.1207e+01	2.4695e+01
	J_y	2.8053e+03	2.3458e+01	2.0798e+01	2.4232e+01
	J_u	2.0774e+03	7.1303e+00	4.0941e-01	4.631e-01
e only	J_e	4.8621e+03	9.9302e+00	2.0067e-01	1.4071e-01
	J_{ey}	2.7850e+03	3.1790e+00	1.3841e-01	1.0275e-01
	J_{eu}	2.0771e+03	6.7512e+00	6.2267e-02	3.7961e-02
v and w only	J_{wv}	2.0659e+01	2.0659e+01	2.1006e+01	2.4554e+01
	J_{wvy}	2.0279e+01	2.0279e+01	2.0659e+01	2.0279e+01
	J_{wvu}	3.7912e-01	3.7912e-01	3.4715e-01	4.2514e-01

TABLE 3. Evaluation of LQG Controllers in Plant Coordinates, Optimal Coordinate and of the LQG_{FW} Controller

see in the TABLE 3, even when the standard LQG controller is in its optimal coordinate (LQG_{T1}), the J_e portion of the cost is still about 33% of the total cost (9.9302 compared to 30.589). By using the new LQG_{FW} controller design algorithm, we reduce the J_e portion of the cost 50 times, compared to the LQG_{T1} controller and 24,110 times compared to the LQG controller. In the latter case, this is equivalent to increasing the wordlength of the control computer by about 7 bits, That is, controller LQG_{FW} will give the same performance using 4 bit arithmetic that LQG gives using 11 bits. Furthermore this improvement in output performance is accompanied by a *reduction* in control effort $RMS = \sqrt{.40941}$ vs. $RMS = \sqrt{2077.4}$. To conclude this point, we see that if both controllers use 4 bits, the difference in RMS output performance is an order of magnitude ($\sqrt{20.798}$ vs. $\sqrt{2805.3}$). This kind of improvement in performance can mean the difference between feasibility and infeasibility of some control missions.

With the new controller, the round-off portion J_e of the cost is only 0.85% of the total cost as opposed to 33% for LQG. Now let us discuss the cost J_{wv}, which would be the total cost if the closed-loop system was *only* excited by measurement noise v_p and disturbance w_p. That is, suppose the LQG_{FW} controller was *designed* for 4 bits, but *evaluated* using infinite bits. These are the conditions of the standard LQG design, since there are no disturbances *in the evaluation.* J_{wv} of the LQG_{FW} controller is a little higher than that of standard LQG controller. The output term of the cost is also a little higher and the control term a little lower. These indicate that the LQG_{FW} controller is a little more conservative than the designed standard LQG controller. This compromise in nominal performance allows robustness to computational errors. Note in TABLE 3, that the quantities that are optimized by the theory (under the given conditions) are shaded.

In the design of the LQG_{FW} controller, the equations (31a) to (31e) were solved iteratively by a gradient method. The standard LQG controller in its optimal coordinate (LQG_{T1}) was used as the initial controller design for starting the iterative process. Figs. 1-3 illustrate the convergence process for the LQG_{FW} algorithm, plotting the total cost J, the wordlength cost J_e, the the output J_y and input J_u performances, versus iteration. The optimal coordinate transformation played a crucial role in reducing the round-off errors (reducing the error by 3-4 orders of magnitude) as shown in Fig. 2. This was expected because the transformation was formulated in the optimization problem. The LQG_{FW} controller was obtained after about 300 iterative computations, but note from Figs. 1-3 that after 120 iterations one might have stopped with little loss.

Coefficient Errors

In the introduction we promised some evaluation of the effects of coefficient errors. We argued that even though the LQG_{FW} controller is optimized only for state quantization it performs well with coefficient quantization as well. To show this we introduced coefficient errors in the controller by using 4 bit precision instead of infinite precision in the controller coefficients. The key issue here is this. Quantization errors in the state degrades performance, but does not destabilize, since the effect of e is just a disturbance (note that all controllers in TABLEs 1 and 2 are stable). Coefficient errors can easily destabilize. Figure 4 shows the closed loop pole locations using the standard LQG regulator (using infinite precision). The system is stable as marked by the x's. When the controller coefficients are implemented using only 4 bit arithmatic, some poles as indicated by the o's in Fig. 4, are outside the unit circle. Hence the standard LQG controller is unstable using a 4 bit control computer.

Fig. 5 shows the improvement in the LQG controller by its optimal coordinate transformation before synthesis. This is the LQG_{T1} controller. The poles (o's) are in improved locations compared to Fig. 4, but the closed loop system is still unstable. The coordinate transformation helped but not enough. Fig. 6 shows the LQG_{FW} controller when controller coefficients are implemented using only 4 bits. The system is stable, confirming for this example improved robustness to controller coefficient errors, even though the controller has been optimized only for errors in controller state computation. The performance degradation in J, listed in the column "LQG_{FW} with coefficient errors" in TABLE 3 is about 15% (compared to nominal performance in TABLE 3).

Finally, we consider errors in *both* the plant and controller coefficients (due to quantization to 4 bits). These results are summarized in TABLE 4, where the modal damping in all modes is multiplied by parameter ρ. Hence $\rho=1$ corresponds to the nominal plant in all of the prior discussion. The range for stability using the LQG_{FW} controller is $.729 \leq \rho \leq 1.23$, demonstrating improved robustness over standard LQG controllers in the presence of errors in plant and controller coefficients.

Damping Error Factor ρ	LQG Controller	LQG_{T1} Controller	LQG_{FW} Controller
1.5242e+00	unstable	unstable	unstable
1.3717e+00	unstable	unstable	unstable
1.2346e+00	unstable	unstable	STABLE
1.1111e+00	unstable	unstable	STABLE
1.0000e+00	unstable (Fig 4)	unstable (Fig 5)	STABLE (Fig 6)
9.0000e-01	unstable	unstable	STABLE
8.1000e-01	unstable	unstable	STABLE
7.2900e-01	unstable	unstable	STABLE
6.5610e-01	unstable	unstable	unstable
5.9049e-01	unstable	unstable	unstable

TABLE 4. Robustness Controllers with respect to modal damping
(4-Bit Wordlength Controllers)

VI. Conclusion

This paper solves the problem of designing an LQG controller to be optimal in the presence of finite wordlength effects (modeled as white noise sources whose variances are a function of computer wordlength). This new controller, denoted LQG_{FW}, has two computational steps. First the gains are optimized, and then a special coordinate transformation must be applied to the controller. This transformation depends on the controller gains, so the transformation cannot be performed *a priori*. (Hence, there is no separation theorem.) The new LQG_{FW} controller design algorithm reduces to the standard LQG controller when an infinite wordlength is used for the controller synthesis, so this is a natural extension of the LQG theory. It was shown both theoretically and by example that the choice of controller coordinates significantly influences the effects of computational errors on the control system and that there exists an optimal set of coordinates in which to do these computations. Since we have not obtained a closed form solution for the LQG_{FW} problem, design of the LQG_{FW} controller by this algorithm requires significant computation. Hence, the improvement of the new controller is achieved at the expense of extra computational effort in design.

Acknowledgement: The importance of this problem was pointed out to us by Darrell Williamson. We gratefully acknowledge many helpful discussions with him, and the support of this work by NASA grant NAG1-857, Technical Monitor E.S. Armstrong.

Appendix A

1. Proof of Lemma 3

a) Using the singular value decomposition of $T = U_t \Pi_t V_t^*$, then the constraint equation (15b) becomes

$$(V_t \Pi_t^{-2} V_t^*)_{ii} = s \qquad \text{for all } i \tag{32}$$

from Lemma 1, above equation is equivalent to

$$\mathrm{tr}(\Pi_t^{-2}) = s n_p \ . \tag{33}$$

Now, let us study the cost γ of (15a). Using the inequality

$$\mathrm{tr}(AA^*) \geq \frac{[\mathrm{tr}(AB^*)]^2}{\mathrm{tr}(BB^*)}$$

we have a lower bound on γ

$$\gamma = \{U_t \Pi_t^2 U_t^* P\} = \mathrm{tr}\{(\Pi_t U_t^* \sqrt{P})(\Pi_t U_t^* \sqrt{P})^*\}$$

$$\geq \frac{[\mathrm{tr}\{(\Pi_t U_t^* \sqrt{P})(U_t^*[\sqrt{P}]^{-1})^*]^2}{\mathrm{tr}\{(U_t^*[\sqrt{P}]^{-1})(U_t^*[\sqrt{P}]^{-1})^*\}} = \frac{(\mathrm{tr}\{\Pi_t\})^2}{\mathrm{tr}\{P^{-1}\}} \tag{34}$$

Now, to prove that γ is unbounded from above, we prove that for any large scalar $m > 0$, we have $\gamma(\tilde{T}) \geq m$ for some $\tilde{T}$. Let us choose a $\tilde{T}$ having the following $\tilde{\Pi}_t$:

$$\tilde{\Pi}_t = \mathrm{diag}\ (\tilde{\Pi}_i) \quad \text{such that}$$

$$\tilde{\Pi}_1 = \tilde{\Pi}_2 = \cdots = \tilde{\Pi}_{n_p-2} = \frac{1}{\sqrt{s}}$$

and

$$\tilde{\Pi}_{n_p-1} = \frac{\sqrt{m\mathrm{tr}(P^{-1})}}{\sqrt{2ms\mathrm{tr}(P^{-1}) - 1}} \ , \ \ \tilde{\Pi}_{n_p} = \sqrt{m\ \mathrm{tr}(P^{-1})}$$

where m is so chosen that

$$m > \frac{1}{2s\mathrm{tr}(P^{-1})}$$

Then

$$\mathrm{tr}(\tilde{\Pi}_t^{-2}) = \sum_{i=1}^{n_p} \frac{1}{\tilde{\Pi}_i^2} = s(n_p - 2) + \frac{2m\mathrm{tr}(P^{-1}) - 1}{m\mathrm{tr}(P^{-1})} + \frac{1}{m\mathrm{tr}(P^{-1})} = sn_p \ .$$

Hence the chosen $\tilde{T}$ satisfies the constraint (33). Now, we have:

$$\gamma \geq \frac{(\mathrm{tr}\{\tilde{\Pi}_t\})^2}{\mathrm{tr}\{P^{-1}\}} = \frac{(\sum_{i=1}^{n_p} \tilde{\Pi}_i)^2}{\mathrm{tr}\{P^{-1}\}} > \frac{(\tilde{\Pi}_{n_p})^2}{\mathrm{tr}\{P^{-1}\}} = m$$

we then conclude the proof of part a). The proof of b) and c) follows next. The lower bound and the matrix T are found by using following inequality:

$$(\mathrm{tr}R)^2 \leq \mathrm{tr}(QRQ^*)\mathrm{tr}(Q^{-*}RQ^{-1}) \tag{35}$$

the equality holds above when $Q^*Q = \lambda^2 I$.

Let us assume $T = U_t\Pi_t V_t^*$, $P = U_p\Pi_p U_p^*$, where Π_t and Π_p are diagonal, U_t, V_t, U_p are unitary matrices. Assume for the R and Q matrices in (35),

$$R = U_t^* U_p \Pi_p^{1/2} U_p^* U_t \tag{36}$$

$$Q^* Q = U_t^* U_p \Pi_p^{1/4} U_p^* U_t \Pi_t^2 U_t^* U_p \Pi_p^{1/4} U_p^* U_t \ , \tag{37}$$

then

$$(Q^* Q)^{-1} = U_t^* U_p \Pi_p^{-1/4} U_p^* U_t \Pi_t^{-2} U_t^* U_p \Pi_p^{-1/4} U_p^* U_t \ .$$

Hence, we have:

$$\mathrm{tr}(QRQ^*) = \mathrm{tr}(RQ^*Q) = \mathrm{tr}[(U_t^* U_p \Pi_p^{1/2} U_p^* U_t)(U_t^* U_p \Pi_p^{1/4} U_p^* U_t \Pi_t^2 U_t^* U_p \Pi_p^{1/4} U_p^* U_t)]$$

$$= \mathrm{tr}[U_p \Pi_p U_p^* U_t \Pi_t^2 U_t^*] = \mathrm{tr}[PTT^*] = \gamma$$

$$\mathrm{tr}(Q^{-*}RQ^{-1}) = \mathrm{tr}[R(Q^*Q)^{-1}] = \mathrm{tr}[(U_t^* U_p \Pi_p^{1/2} U_p^* U_t)(U_t^* U_p \Pi_p^{-1/4} U_p^* U_t \Pi_t^{-2} U_t^* U_p \Pi_p^{-1/4} U_p^* U_t)]$$

$$= \mathrm{tr}[U_p^* U_t \Pi_t^{-2} U_t^* U_p] = \mathrm{tr}[\Pi_t^{-2}]$$

From equation (33), and the above equation we have the following:

$$\mathrm{tr}(Q^{-*}RQ^{-1}) = \mathrm{tr}[\Pi_t^{-2}] = sn_p$$

Now, $\mathrm{tr}(R) = \mathrm{tr}(U_t^* U_p \Pi_p^{1/2} U_p^* U_t) = \mathrm{tr}(\Pi_p^{1/2}) = \mathrm{tr}(U_p \Pi_p^{1/2} U_p^*) = \mathrm{tr}\sqrt{P}$. Substitute the above equalities back into inequality (35). We then have: $[\mathrm{tr}(\sqrt{P})]^2 \le sn_p\gamma$, that is

$$\gamma \ge \frac{[\mathrm{tr}(\sqrt{P})]^2}{sn_p} \quad . \tag{38}$$

Now, suppose the matrix $\overline{T} = \overline{U}_t \overline{\Pi}_t \overline{V}_t^*$ yields the equality in (38). Since the equality in (35) holds when $Q^*Q = \lambda^2 I$, then we have:

$$\overline{U}_t^* U_p \Pi_p^{1/4} U_p^* \overline{U}_t \overline{\Pi}_t^2 \overline{U}_t^* U_p \Pi_p^{1/4} U_p^* \overline{U}_t = \lambda^2 I \ ,$$

that is

$$\overline{U}_t \overline{\Pi}_t^2 \overline{U}_t^* = \lambda^2 U_p \Pi_p^{-1/2} U_p^* \Rightarrow \overline{U}_t \overline{\Pi}_t^{-2} \overline{U}_t^* = \lambda^2 U_p \Pi_p^{1/2} U_p^* \quad . \tag{39}$$

Hence

$$\overline{\Pi}^{-2} = \frac{\overline{U}_t^* U_p \Pi_p^{1/2} U_p^* \overline{U}_t}{\lambda^2} \quad .$$

Substitute this $\overline{\Pi}^{-2}$ into equation (32) to obtain

$$(\overline{V}_t \overline{U}_t^* U_p \frac{\Pi_p^{1/2}}{\lambda^2} U_p^* \overline{U}_t \overline{V}_t^*)_{ii} = s \quad .$$

Then $\mathrm{tr}\left[\frac{\Pi\sqrt{P}}{\lambda^2}\right] = sn_p$, hence $\lambda^2 = \frac{1}{sn_p}\mathrm{tr}(\Pi_p^{1/2}) = \frac{1}{sn_p}\mathrm{tr}(\sqrt{P})$. Now, substitute the above λ^2

into (39), to obtain

$$\overline{U}_t \overline{\Pi}_t^{-2} \overline{U}_t^* = \frac{sn_p U_p \Pi_p^{1/2} U_p^*}{tr(\sqrt{P})} = \frac{sn_p \sqrt{P}}{tr(\sqrt{P})} \tag{40}$$

Hence (38) yields the lower bound in (16a), and the matrix achieving this bound, shown by (40), must satisfy (17b). (17c) can be easily deduced from (15b). This concludes the proof.

□

2. <u>Proof of Lemma 4</u>

a) Proof of (29a): Since J_{kx} does not depend on K(p,q) for $p \neq i$ or $q \neq i$, we have:

$$\nabla_k J_{rx}(p, q) = \frac{\partial}{\partial K(p, q)} J_{k,x} = 0$$

Proof of (29b): We need following equality (e.g. Page 444 of Skelton [1988]) to prove the equation:

$$\lambda_i[A] = [E^{-1}]_{ith-row} A [E]_{ith-col}$$

where λ_i is the ith eigenvalue of A, and E the eigenvector matrix of A. Now, we have by taking $A = K(i, i)X(j, j)$

$$\lambda_l[K(i, i)X(j,j)] = [E^{-1}]_{lth-row} K(i, i)X(j, j)[E]_{lth-col}$$

$$= tr\{K(i, i)X(j, j)[E]_{lth-col}[E^{-1}]_{lth-row}\}$$

Hence from the differentiation rule $\frac{\partial trAB}{\partial B} = A^T$ we get

$$\frac{\partial \lambda_l}{\partial K(i,\, i)} = [E^{-1}]^T_{l\text{th-row}} [E]^T_{l\text{th-col}} X(j,\, j)$$

Then, we have:

$$\frac{\partial J_{k,x}}{\partial K(i,\, i)} = \frac{1}{2} \sum_{l=1}^{n} \frac{\frac{\partial}{\partial K(i,\, i)} \lambda_l [K(i,\, i) X(j,\, j)]}{\sqrt{\lambda_l [K(i,\, i) X(j,\, j)]}}$$

$$= \frac{1}{2} \sum_{l=1}^{n} \frac{[E^{-1}]^T_{l\text{th-row}} [E]^T_{l\text{th-col}} X(j,\, j)}{\sqrt{\lambda_l [K(i,\, i) X(j,\, j)]}} \; .$$

The proof of part b) follows in a similar manner

□

<u>3. Proof of Theorem 2</u>: Apply Lagrangian Multipliers K_2, K_3, then (30a)-(30c) leads to minimization of

$$\tilde{J} = \mathrm{tr}\{Q([C + I_0 GM] X_1 [C + I_0 GM]^* + (I_0 G I_1)(V_p + E_z)(I_0 G I_1)^*)]\}$$
$$+ \mathrm{tr}\{K_2([A + BGM] X_1 [A + BGM]^* + D W_p D^* + (BGI_1)(V_p + E_z)(BGI_1)^* + (BI_1) E_u (BI_1)^*$$
$$- X_1)\} + \mathrm{tr}\{K_3([A + BGM]^* K_e [A + BGM] + [C + I_0 GM]^* Q$$
$$[C + I_0 GM] - K_e)\} + \frac{q}{s n_c} (\mathrm{tr} \Sigma_k)^2$$

Then

$$\frac{\partial\tilde{J}}{\partial K_2} = [A + BGM]X_1[A + BGM]^* + DW_pD^* + (BGI_1)(V_p + E_z)(BGI_1)^* + BI_1E_u(BI_1)^* - X_1 = 0$$

$$\frac{\partial\tilde{J}}{\partial K_3} = [A + BGM]^*K_e[A + BGM] + [C + I_0GM]^*Q[C + I_0GM] - K_e = 0$$

$$\frac{\partial\tilde{J}}{\partial X_1} = [C + I_0GM]^*Q[C + I_0GM] + [A + BGM]^*K_2[A + BGM] - K_2 + \nabla_{x_1} = 0$$

$$\frac{\partial\tilde{J}}{\partial K_e} = [A + BGM]K_3[A + BGM]^* - K_3 + \nabla_{k_1} = 0$$

Applying Lemma 4 on the above two equations, we can obtain ∇_{x_1} and ∇_{k_1} as stated in the theorem. This verifies (31a)-(31d). Now

$$\begin{aligned}\frac{\partial\tilde{J}}{\partial G} &= 2I_0^*QCX_1M^* + 2I_0^*QI_0GMX_1M^* + 2I_0^*QI_0GI_2(V_p + E_z)I_1^* + 2B^*K_2AX_1M^* \\ &+ 2B^*K_2BGMX_1M^* + 2B^*K_2BGI_1(V_p + E_z)I_1^* + 2B^*K_1AK_3M^* \\ &+ 2B^*K_1BGMK_3M^* + 2I_0^*QCK_3M^* + 2I_0^*QI_0GMK_3^*M^* = 0 \ ,\end{aligned}$$

but since $I_0^*QC = 0$, then,

$$\begin{aligned}\frac{\partial\tilde{J}}{\partial G} &= 2[I_0^*QI_0G(MX_1M^* + I_1(V_p + E_z)I_1^*) + B^*(K_2AX_1 + K_2AK_3)M^* \\ &+ (B^*K_1B + I_0^*QI_0)GMK_3M^* + B^*K_2BG(MX_1M^* + I_1(V_p + E_z)I_1^*)] \\ &= 2[(I_0^*QI_0 + B^*K_2B)G(MX_1M^* + I_1(V_p + E_z)I_1^*) + B^*(K_2AX_1 + K_1AK_3)M^* \\ &+ (B^*K_1B + I_0^*QI_0)GMK_3M^*] = 0 \ .\end{aligned}$$

This verifies (31e).

□

Appendix B

We now present an algorithm (originally developed by Hwang [1977]) for solving (17b) and (17c) for one set of solutions of U_t, Π_t, V_t (The solutions for U_t, Π_t, V_t are not unique). Let $\sqrt{P}$ in (17b) be written in terms of its singular value decomposition

$$\sqrt{P} = U_p \Sigma_p U_p^* \tag{41}$$

where U_p unitary, Σ_p diagonal.

Algorithm (Solving U_t, Π_t, V_t in (17b) and (17c))

I. Take:

$$U_t = U_p \tag{42a}$$

$$\Pi_t = \sqrt{\frac{tr(\Sigma_p)}{sn_p}\Sigma_p^{-1}} \tag{42b}$$

$$V_t = V_{n-1} V_{n-2} \cdots V_i \cdots V_2 V_1 \tag{42c}$$

where V_i, $i = 1, \cdots, n-1$ is computed as follows:

II. Compute V_1: Let

$$\Sigma_1 \triangleq \Pi_t^{-2} = \text{diag}\,(\cdots\ \sigma_{1j} \cdots) \tag{43a}$$

Assume σ_{11} and $\sigma_{1\beta}$ are two numbers such that one of them is bigger than s, the other is smaller than s. Then take V_1 as:

$$
V_1 = \begin{array}{r c}
 & \beta \text{ column} \\
 & \left[\begin{array}{cccccccc}
f_1 & 0 & \dots & 0 & g_1 & 0 & \dots & 0 \\
0 & 1 & \dots & 0 & 0 & 0 & \dots & 0 \\
\vdots & & \ddots & & \vdots & & & \vdots \\
0 & & & 1 & 0 & & & \\
-g_1 & 0 & \dots & 0 & f_1 & 0 & \dots & 0 \\
0 & & \dots & \dots & & 1 & & \vdots \\
\vdots & & & & & & \ddots & \\
0 & & \dots & \dots & & \dots & \dots & 1
\end{array}\right] \\
\beta \text{ row} \rightarrow \text{ (5th row)} &
\end{array} \tag{43b}
$$

where

$$
f_1 = \left[\frac{\sigma_{1\beta} - 1}{\sigma_{1\beta} - \sigma_{11}} \right]^{1/2} \tag{43c}
$$

$$
g_1 = \left[\frac{1 - \sigma_{11}}{\sigma_{1\beta} - \sigma_{11}} \right]^{1/2} \tag{43d}
$$

<u>Compute V_i</u>: Let

$$
\Sigma_i = V_{i-1}\, \Sigma_{i-1}\, V_{i-1}^* = \begin{bmatrix} \Sigma_{i1} & \Sigma_{i2} \\ \Sigma_{i2}^* & \Sigma_{i3} \end{bmatrix} \tag{44a}
$$

where $\Sigma_i \,\varepsilon\, \mathbb{R}^{(i-1)x(i-1)}$ satisfies the property $[\Sigma_{i1}]_{jj} = s$, $\Sigma_{i2} \,\varepsilon \mathbb{R}^{(i-1)x(n-i+1)}$ is a nonzero matrix, and Σ_{i3} can be written as

$$
\Sigma_{i3} = \begin{bmatrix} \sigma_{ii} & & 0 \\ & \ddots & \\ 0 & & \sigma_{nn} \end{bmatrix}
$$

Assume σ_{ii} and $\sigma_{i\alpha}$ are numbers such that one of them is bigger than s and the other is smaller than s. Then take V_i as

$$
\begin{array}{rc}
 & \begin{array}{ccccccccc} & & \text{i column} & & & & \alpha\ \text{column} & & \end{array} \\
\begin{array}{r} \\ \\ \text{i row} \rightarrow \\ \\ V_i = \\ \\ \alpha\ \text{row} \rightarrow \\ \\ \\ \end{array} &
\begin{bmatrix}
1 & & 0 & & & & 0 & & \\
 & \ddots & \vdots & & 0 & & \vdots & 0 & \\
0 & \cdots & f_i & 0 & \cdots & 0 & g_i & \cdots & 0 \\
 & & 0 & 1 & & & 0 & & \\
 & 0 & \vdots & & \ddots & & \vdots & 0 & \\
 & & 0 & & & 1 & 0 & & \\
0 & \cdots & g_i & 0 & \cdots & 0 & f_i & \cdots & 0 \\
 & 0 & \vdots & & 0 & \vdots & & \ddots & \\
 & & 0 & & & 0 & & & 1
\end{bmatrix}
\end{array}
\tag{44b}
$$

Compute f_i and g_i as:

$$f_i = \left[\frac{\sigma_{i\alpha} - 1}{\sigma_{i\alpha} - \sigma_{ii}} \right]^{1/2} \tag{44c}$$

$$g_i = \left[\frac{1 - \sigma_{ii}}{\sigma_{i\alpha} - \sigma_{ii}} \right]^{1/2} \tag{44d}$$

□

Computation of $\underline{T}_c$

$\underline{T}_c$ is formed as follows: $\underline{T}_c \triangleq U_x^* \Sigma_x^{1/2} U_K^* U_t \Pi_t V_t^*$

1) Compute the Covariance Matrix and Observability Grammian

$$K_e = [A + BGM]^* K_e [A + BGM] + [C + I_o GM]^* Q [C + I_o GM]$$

$$X_1 = [A + BGM] X_1 [A + BGM]^* + D W_p D^* + (BGI_1)(V_p + E_z)(BGI_1)^* + BI_1 E_u BI_1$$

Assume $K_e(2,2)$, $X_1(2,2)$ to be (2,2) the subblocks of K_e and X_1 (the controller subblocks).

2) Compute U_x, Σ_x, U_k.

These three matrices are computed by applying singular value decomposition on following matrices:

$$X_1(2,2) = U_x^* \Sigma_x U_x$$

$$\Sigma_x^{1/2} U_x K_2(2,2) U_x^* \Sigma_x^{1/2} = U_k^* \Sigma_k U_k$$

3) Compute U_t, Π_t, V_t.

Let us replace P matrix in the algorithm of appendix B as

$$P \triangleq \mathrm{diag}\,[\lambda_i \{K_e(2,2) X_1(2,2)\}]$$

Then we can compute U_t, Π_t, V_t by applying the algorithm on matrix P.

Appendix C

DESIGN EXAMPLE OF ROUND-OFF LQG CONTROLLER

===

Plant Model: 10th Order Euler-Bernoulli Beam

Word-Length of the Assumed Computer: 4 bits

1) The 10th Order Euler-Bernoulli Beam Model for Controller Design

$$A = \begin{bmatrix} 0.9980 & 0.0179 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -0.2196 & 0.9968 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.9687 & 0.0177 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -3.4620 & 0.9582 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.8469 & 0.0166 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -16.4457 & 0.7993 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.5594 & 0.0139 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -43.6477 & 0.4340 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1138 & 0.0095 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -72.4045 & 0.0937 \end{bmatrix}$$

$$B = \begin{bmatrix} 0.0014 & 0.0006 \\ 0.1557 & 0.0716 \\ -0.0004 & 0.0011 \\ -0.0480 & 0.1257 \\ -0.0012 & 0.0013 \\ -0.1299 & 0.1440 \\ 0.0007 & 0.0012 \\ 0.0720 & 0.1164 \\ 0.0007 & 0.0007 \\ 0.0588 & 0.0588 \end{bmatrix} \qquad D = \begin{bmatrix} 0.0014 & 0.0006 \\ 0.1557 & 0.0716 \\ -0.0004 & 0.0011 \\ -0.0480 & 0.1257 \\ -0.0012 & 0.0013 \\ -0.1299 & 0.1440 \\ 0.0007 & 0.0012 \\ 0.0720 & 0.1164 \\ 0.0007 & 0.0007 \\ 0.0588 & 0.0588 \end{bmatrix}$$

$$C = \begin{bmatrix} 0 & 7.8297 & 0 & 7.1091 & 0 & -1.3744 & 0 & -8.3569 & 0 & -6.2128 \\ 0 & 6.2128 & 0 & -8.7875 & 0 & 6.2128 & 0 & 0 & 0 & -6.2128 \end{bmatrix}$$

$$M = \begin{bmatrix} 0 & 7.8297 & 0 & 7.1091 & 0 & -1.3744 & 0 & -8.3569 & 0 \\ 0 & 6.2128 & 0 & -8.7875 & 0 & 6.2128 & 0 & 0 & 0 \end{bmatrix}$$

$$W = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad V = \begin{bmatrix} 1.0003e{-}03 & 0 \\ 0 & 1.0003e{-}03 \end{bmatrix}$$

2) Designed Regular LQG Controller in Optimal Coordinate LQG_{T1}

$$
A_c = \begin{bmatrix}
-0.4582 & -0.1633 & -0.0133 & -0.1836 & 0.1574 & -0.4386 & -0.1054 & -0.2805 & 0.2304 & -0.2815 \\
0.4144 & 0.6040 & 0.4587 & -0.4122 & -0.0201 & -0.0411 & 0.2748 & 0.1059 & -0.0786 & 0.0379 \\
0.0849 & -0.5217 & 0.5622 & -0.3257 & 0.3373 & 0.2351 & 0.0665 & 0.1975 & -0.1651 & 0.2658 \\
0.4753 & -0.3503 & 0.2226 & 0.5105 & -0.3084 & 0.0821 & 0.4446 & 0.1978 & -0.1382 & -0.0456 \\
0.3326 & 0.0383 & -0.5299 & -0.1864 & 0.4324 & 0.3391 & 0.3306 & 0.2351 & -0.1635 & -0.1155 \\
0.2946 & -0.1855 & -0.0.850 & -0.3095 & -0.2941 & -0.0605 & -0.7404 & 0.0085 & 0.1530 & 0.5389 \\
1.5034 & -0.2726 & -0.0095 & -0.2270 & -0.0416 & -0.4845 & -1.5704 & -0.3867 & -0.0236 & -0.4084 \\
0.5293 & 0.0908 & -0.0359 & -0.0617 & -0.3343 & -0.0787 & -1.0273 & -0.1971 & -0.0491 & 0.4129 \\
-0.0468 & -0.0574 & -0.0709 & -0.0716 & -0.0416 & 0.1318 & 0.5827 & -0.9215 & -0.0746 & 0.2806 \\
-0.4312 & 0.1539 & -0.0256 & 0.0559 & -0.1463 & 0.4745 & -0.0777 & -0.3449 & -0.9854 & -0.6735
\end{bmatrix}
$$

$$
B_c = \begin{bmatrix}
0.1894 & -0.2895 \\
-0.422 & 0.0230 \\
-0.0296 & 0.0941 \\
-0.0120 & -0.0024 \\
-0.0258 & 0.0940 \\
-0.0611 & 0.0609 \\
-0.2200 & 0.4919 \\
-0.0737 & 0.2522 \\
0.0252 & -0.0076 \\
0.0737 & -0.0776
\end{bmatrix}
\qquad
D_c = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}
$$

$$
C_c = \begin{bmatrix}
-1.9370 & 3.8601 & 4.1659 & 3.4458 & 1.8923 & -4.2436 & -15.7358 & -6.5380 & 3.7048 & -5.3330 \\
1.6850 & -3.2381 & -2.7357 & -2.8624 & -2.7744 & 5.2406 & 11.5365 & 3.9625 & -2.8745 & -0.1711
\end{bmatrix}
$$

3) LQG_{FW} Controller from the LQG_{FW} Algorithm of Section 4

$$A_c = \begin{bmatrix} 0.3501 & 0.4306 & -0.2223 & 0.3078 & -0.5350 & 0.1231 & 0.1595 & -0.2003 & -0.1024 & 0.1325 \\ -0.2004 & -0.2851 & -0.2294 & 0.1810 & 0.6715 & -0.4432 & 0.2756 & 0.1591 & 0.3525 & -0.3974 \\ -0.2033 & 0.2556 & -0.0197 & -0.8326 & -0.8293 & 0.0885 & -0.0605 & 0.2870 & -0.0571 & 0.1147 \\ -0.2973 & -0.3621 & 0.6480 & 0.3770 & -0.4095 & 0.4031 & -0.2736 & 0.0125 & 0.0426 & -0.0372 \\ 0.0308 & -0.2207 & -0.5168 & -0.3001 & -0.9847 & 1.1705 & -1.0703 & 0.7456 & -0.0979 & 0.1920 \\ -0.1187 & 0.4836 & 0.0470 & 0.3655 & 0.2493 & -1.0109 & 0.3516 & 0.5930 & 0.2744 & -0.3872 \\ 0.0089 & -0.3363 & 0.0664 & -0.0869 & 0.0085 & -0.0712 & -0.1936 & 0.113 & 0.3818 & -0.5248 \\ -0.3341 & 0.2935 & 0.1055 & 0.1309 & 0.2251 & -0.3631 & -0.7912 & -0.5655 & -0.2610 & 0.3180 \\ 0.0731 & -0.0312 & -0.0788 & -0.1349 & -0.4369 & 0.2594 & -0.4096 & -0.3895 & 0.7609 & 0.3237 \\ -0.1129 & 0.0070 & 0.0781 & 0.1679 & 0.4955 & -0.3354 & 0.5865 & 0.4685 & 0.2460 & 0.6396 \end{bmatrix}$$

$$B_c = \begin{bmatrix} -0.0134 & 0.0927 \\ 0.0812 & -0.1630 \\ 0.0706 & -0.3987 \\ 0.2464 & -0.6411 \\ 0.4583 & -1.0134 \\ -0.5942 & 1.0745 \\ 0.2455 & -0.2146 \\ 0.1121 & 0.0815 \\ 0.1013 & -0.2475 \\ -0.1510 & 0.3465 \end{bmatrix} \qquad D_c = \begin{bmatrix} -0.4486e{-}04 & -0.1328e{-}04 \\ -0.5913e{-}04 & -0.1567e{-}04 \end{bmatrix}$$

$$C_c = \begin{bmatrix} 0.8861 & -1.8997 & 3.8592 & -0.3107 & 5.3072 & -0.7395 & 0.6339 & -1.6517 & 0.9202 & -1.3734 \\ -1.4019 & 2.2532 & -2.6576 & 0.1575 & -3.3358 & 1.2007 & -0.5179 & 0.2062 & -0.9969 & 1.3884 \end{bmatrix}$$

References

[1] S. Hwang [1977]; "Minimum Uncorrelated Unit Noise in State-Space Diginal Filtering;" IEEE Trans, Acoust. Speech, Signal Processing; Vol-25; Aug. 1977; pp. 273-281.

[2] K. Kadiman and Williamson [1989]; "Optimal Finite Wordlength Linear Quadratic Regulation" IEEE Trans. on Automatic Contr., Vol. 34, No. 12, pp. 1218-1228, Dec. 1989.

[3] H. Kwakernaak; R. Sivan [1972]; "Linear Optimal Control Systems;" John Wiley & Sons.

[4] P. Lancaster [1969]; "The Theory of Matrix;" Academic Press.

[5] D. Luenberger [1984]; "Linear and Nonlinear Programming;" Addison-Wesley.

[6] P. Moroney; A. Willsley; P. Houpt [1983]; "Round-Off Noise and Scaling in the Digital Implementation of Control Compensators;" IEEE Trans; Acoust. Speech, Signal Processing; Vol-31; Dec. 1983; pp. 1464-1477.

[7] C. Mullis and R. Roberts [1976]; "Synthesis of Minimum Round-Off Noise Fixed Point Digital Fiters;" IEEE Trans.; Circuits and Syst.; Vol-23; Sept. 1976; pp. 551-562.

[8] R. Skelton [1988]; "Dynamical System Control;" John Wiley & Son.

[9] A. Sripad; D. Synder [1977]; "A Necessary and Sufficient Condition for Quantization Error to be Uniform and White;" IEEE Trans. Acout. Speech, Signal Processing; Vol-5; Oct. 1977; pp. 442-448.

[10] A. Sripad [1981]; "Performance Degradation in Digitally Implemented Kalman Filter;" IEEE Trans. Aerospace Electron. System; Vol-17; Sept. 1981; pp. 626-634.

[11] D. Williamson [1985]; "Finite Word Length Design of Digital Kalman Filters for State Estimation;" IEEE Trans. on Automatic Contr.; Vol-30; Oct. 1985; pp. 30-39.