Roundoff Noise and Scaling in the Digital Implementation of Control Compensators

PAUL MORONEY, MEMBER, IEEE, ALAN S. WILLSKY, SENIOR MEMBER, IEEE, AND PAUL K. HOUPT, MEMBER, IEEE

Abstract—Researchers in digital signal processing have examined at length the effects of finite wordlength in the design of digital filters. The issues that have been considered apply to any digital system. In particular, the design of digital control systems must consider these issues. In this paper we will use, adapt, and extend the ideas developed in digital signal processing to the issue of roundoff noise in digital linear-quadratic-Gaussian (LQG) compensators. We will then examine the roundoff noise effects for a particular LQG example and several different implementation structures.

I. Introduction

In the design of digital filters, it has been amply demonstrated that one must consider the effects of the finite precision inherent in the digital implementation. This finite precision leads to degradation due to quantization noise, coefficient inaccuracy, and limit cycle oscillations. These effects have been the subject of a great deal of research in digital signal processing [1]-[3].

It is also important to investigate these issues in the application of digital processing to other fields—specifically, to discrete-time control systems. In the past, complex controller designs have usually been implemented on large, expensive, floating-point computer systems. However, the number of applications that could effectively use small-scale hardware control systems that work in real time has greatly increased, especially with the advent of the inexpensive microprocessor. When implementing such compensators, we must consider the problems that arise in dealing with the fixed-point arithmetic and finite wordlengths of small-scale digital systems. As these problems are not addressed at all in the idealized mathematical design procedures that have been developed to date for control, a methodology must be established for treating the digital implementation of a compensator design.

For this methodology, we have turned to the results developed for implementing digital filters. On one level, a single-input single-output digital compensator is simply a digital filter. However, as we will show, the existence of a feedback loop around the digital compensator will frequently require us to adapt and extend the ideas developed for filter implementations. These adaptations of digital signal processing ideas will serve as the basis for techniques dealing with roundoff noise, scaling, coefficient rounding, and limit cycles in digital feedback compensators.

Some work has already been done concerning controller implementation [5]-[14]. However, these have been somewhat limited in scope, typically treating only a few points, or only one structure. Our intention is quite different—a systematic extension and adaptation of digital signal processing ideas to controller implementation. More simply, our intention is to raise the issues commonly dealt with by researchers in digital signal processing, but in the context of feedback controller implementation. We have reported results concerning the coefficient quantization issue in [15] and [16]. In this paper, we will examine the issues of scaling and roundoff quantization in fixed-point controller implementations.

The basic idea behind the examination of roundoff noise is the same for digital filters and digital compensators. Approximating the results of intermediate computations, or node signal values, with a finite number of bits will cause a degradation in the system's performance as compared to the ideal. Assuming that a given quantitive performance measure is provided, we can measure the tradeoff in the number of bits versus the degradation. Then, assuming that we specify an acceptable amount of degradation, we can determine the minimum number of node variable bits needed to meet this goal. In fixed-point digital systems, this can be done by bounding the effects of quantization [8]-[11], [17]; however, such bounds would also include limit cycle effects, and thus be very loose as regards quantization *noise*. More commonly, we assume that the roundoff operation can be modeled as additive white noise, thus allowing the use of linear system analysis techniques [3]. We will adopt the latter approach in our investigation.

In order to bring out specific issues concerning control system implementations, we will consider a specific class of control problems—linear-quadratic-Gaussian (LQG) problems. The compensator resulting from the LQG framework is optimal, in the sense that it minimizes a quadratic functional of the state and control fluctuations. This compensator will be described in detail in Section II. It is also very convenient to treat this quadratic functional as the index of performance

Manuscript received October 1, 1980; revised November 29, 1982. This work was performed in part at the M.I.T. Laboratory for Information and Decision Systems, supported by NASA Ames under Grant NGL-22-009-124, and in part at the Charles Stark Draper Laboratory.

P. Moroney is with the Linkabit Corporation, San Diego, CA 92121.

A. S. Willsky is with the Laboratory for Information and Decision Systems, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139.

P. K. Houpt is with the Laboratory for Information and Decision Systems, Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139.

0096-3518/83/1200-1464$01.00

by which we measure the relative merit of different digital implementations of a controller design. Thus, any given implementation will produce a performance index greater than the "optimal" ideal (infinite-precision) value. This increase will reflect the degradation due to some finite wordlength effect, such as roundoff quantization noise. The fact that we use this performance metric, somewhat different from but more appropriate than the l_2-norm (variance) of the output noise used in digital filtering analysis, is another reason our results differ from those reported in the digital signal processing literature. If the problem under consideration had been a Kalman or Weiner filter, then a suitable performance measure would have been the trace of the error covariance matrix. Our results extend in a straightforward manner to this case, and also to the output noise power measure. In any case, for the control problem, it is the presence of a feedback loop through some controlled system which makes our results and techniques novel. Specifically, as we will see, the concept of a structure for a digital compensator, the notation developed by Chan [18] for describing the operations in a digital filter structure (including precedence), approaches to scaling, roundoff noise analysis techniques, and the minimum roundoff noise structures of Mullis and Roberts [19], [20] and Hwang [21] all need to be adapted for the control problem. In this paper we shall deal only with single-input single-output control systems, although our results can be extended to multiple-input multiple-output systems [16].

The organization of this paper will be as follows. In Section II we will describe the LQG control problem and the resulting ideal optimal compensator–it is this ideal compensator that must be implemented with as little degradation as possible, due to finite precision effects. The notion of a compensator *structure*, somewhat different from a conventional filter structure, and an adaptation of the Chan notation for describing such structures will be presented in Section III. In Section IV we will review the digital signal processing techniques for scaling structures to satisfy the dynamic range constraints of fixed-point digital filters, and we will show how these ideas must be modified for digital control systems. The scaling issue, of course, is central to any meaningful measurement of roundoff quantization noise effects. In Section V we will review the techniques for roundoff noise analysis in digital filtering, and extend these to digital controllers. The minimum roundoff noise structures of Mullis and Roberts [19], [20] and Hwang [21] will be adapted for controllers in Section VI, along with a treatment of the more general optimization techniques introduced by Chan, also as adapted for controller implementation. Finally, using the techniques we have developed we will compare the roundoff quantization noise performance of several scaled structures for implementing a specific LQG controller example. We will also show that a "default" compensator structure, quite natural for the control designer to implement, is, in fact, *not* a very desirable structure to select.

II. LQG Controller Design

In this section we will introduce the single-input single-output LQG control problem and the optimal compensator that results. This procedure will, of course, only specify an ideal design–one that would only be possible with infinite-precision arithmetic. From this ideal we wish to select a finite-precision implementation which results in as little degradation (in the performance index) as possible. This is directly analogous to designing a digital filter using a bilinear transformation, impulse invariance, or whatever technique, and then implementing the "ideal" design in finite-precision.

Let us assume that we wish to design a digital discrete-time compensator for a continuous-time system (a "plant"), and that the control signal will be piecewise constant. We will also assume that the output of the plant is sampled at the rate $1/T$. The term linear-quadratic-Gaussian refers to the following design problem: given a linear discrete-time model of a continuous-time system subject to disturbances that can be modeled as white *Gaussian* noises, design a *linear* compensator that minimizes a *quadratic* performance index.

Consider the following discrete-time model of a continuous-time time-invariant plant:

$$\begin{aligned} x(k+1) &= \Phi x(k) + \Gamma u(k) + w_1(k) \\ y(k) &= Lx(k) + w_2(k) \end{aligned} \tag{1}$$

where x is the state n-vector, u and y are the control and output variables, Φ is an $n \times n$ state transition matrix, Γ is an $n \times 1$ input gain matrix, and L is a $1 \times n$ output gain matrix. The quantities w_1 and w_2 are the discrete Gaussian noises referred to above. These noises are zero mean, with covariance matrices Θ_1 ($n \times n$) and Θ_2 (1×1), respectively. For the steady-state LQG problem, the performance index can be written as follows:

$$J = E\left\{\lim_{i\to\infty} \frac{1}{2i} \sum_{k=-i}^{i} \{x'(k)Qx(k) + 2x'(k)Mu(k) + Ru^2(k)\}\right\}. \tag{2}$$

Thus, we see that J reflects the weighted squared deviations of the states and of the control. The weighting parameters Q, M, and R can be specified by the designer. The infinite time horizons reflect the steady-state nature of the optimization, both for the optimal state estimation and the optimal regulation [22], [23].

The determination of a linear compensator that minimizes J involves the solution of two Riccati equations involving the plant and weighting parameters. However, the resulting control $u(k)$ typically will depend on past values of the plant output up to and including $y(k)$ [22]. Unfortunately, the resulting compensator is not directly feasible for implementation, since a certain amount of time must be allowed to compute $u(k)$ from $y(k)$, $y(k-1)$, etc. Yet $u(k)$ and $y(k)$ refer to the control and plant output at *identical* times. Some delay *must* be accounted for, and thus the design, as described so far, is infeasible.

Fortunately, Kwakernaak and Sivan [23] have presented a design procedure that does account for this delay. The resulting compensator is optimal in the sense that it produces the $u(k)$ that minimizes J, but based only on a linear function of $y(k-1), y(k-2), \cdots$, and *not* on $y(k)$. Such a compensator *can* be implemented, essentially allowing one full sample

period for the computation of $u(k)$ after the $y(k-1)$ sample is generated. If, however, the computation time is much shorter than the sample interval, this implies some inefficiency; the output $u(k)$ will be available long before it is used as a control. Thus, Kwakernaak and Sivan also include a method for skewing the sample time of the plant output with respect to the rest of the compensator. The compensator output $u(k)$ will still depend on inputs up to and including $y(k-1)$, but now $y(k-1)$ is produced only one computation time before $u(k)$ is needed. This eliminates any inefficiency [16], [23]. We will return to the implications of this necessary calculation time in Section III. The sample skew method involves a change to (1) to describe the new sample y, and results in a new term proportional to $u(k)$ in this equation [16]. This added complexity was not pursued in the remainder of this paper, although it can be done with no problem. Thus, for simplicity, we assume no skewing.

The optimal compensator described above is of the following form (assuming no sample skew):

$$\hat{x}(k+1) = \Phi\hat{x}(k) + \Gamma u(k) + K(y(k) - L\hat{x}(k))$$
$$u(k+1) = -G\hat{x}(k+1) \tag{3}$$

where $\hat{x}$ is the estimated state vector and the $n \times 1$ Kalman filter matrix K and $1 \times n$ regulator matrix G result from the solution of two discrete-time algebraic Riccati equations [23]. Note that in (3), the next control $u(k+1)$ depends only on inputs $y(k), y(k-1), \cdots$. Thus, the computational delay (one full sample period) has been allowed for in this formulation.

Now, if we treat this compensator as a discrete linear system and examine its transfer function, we have[1]

$$\frac{U(z)}{Y(z)} = -G(zI - \Phi + KL + \Gamma G)^{-1}K. \tag{4}$$

In a more conventional form, this can be written

$$\frac{U(z)}{Y(z)} = \frac{a_1 z^{-1} + a_2 z^{-2} + \cdots + a_n z^{-n}}{1 + b_1 z^{-1} + b_2 z^{-2} + \cdots + b_n z^{-n}}. \tag{5}$$

Note the lack of a term a_0 in the numerator. The presence of such a term would reflect a dependence of the *present* output on the *present* input. Since (5) represents a compensator that *can* be implemented, the a_0 term must be zero.

This delay has an important implication in the way we look at structures for implementing digital compensators, as we will show in Section III.

III. Algorithms and Structures for Digital Compensators

In the nomenclature of digital signal processing [1], the term *structure* refers to the specific combination of (finite-precision) arithmetic operations by which a filter output sample is generated from intermediate values and the input. Typically, a structure can be represented by a signal-flow graph. Let us examine a simple filter structure to see whether it will be appropriate for representing the compensator of (5).

[1] Note that we have taken u to be the output of the digital network and y to be the input. This may be contrary to the expectations of many readers.

Specifically, let us examine a fourth-order ($n = 4$) direct form II [1] filter structure (see Fig. 1). Note the presence of the a_0 term. Such a structure *cannot* exactly represent an implementation, since computational delay has not been accounted for. However, such a signal-flow graph *is* taken to represent a structure in digital signal processing; basically, the extra series delay needed for computations is assumed to be present, and is ignored. In most digital filter applications, series delay is of no consequence. However, in any control system, all delays that exist *must* be adequately represented in the structure notation. If series delay exists in the compensator, and has *not* been accounted for, the entire control system may be unstable. Control system performance always deteriorates when extra delay is added to the loop. Thus, any treatment of compensator structures must include specification of *all* calculation delays. This consideration basically led to the form of (5).

Now, let us take Fig. 1 and set a_0 to zero, as in (5) (see Fig. 2). The signal-flow graph of Fig. 2 is still *not* an accurate representation of an implementation of (5). The only time available for computation is *between* sample times (ignore sample-skewing for now). Yet, Fig. 2 shows $u(k)$ depending on compensator state nodes (defined to be the outputs of delay elements), also at the same time k. Time must be allowed for the multiplications a_1-a_4. Thus, $u(k)$ cannot be in existence until *after* the state node values are calculated.

A structure appropriate for representing the compensator of (5) is depicted in Fig. 3. This can be derived from Fig. 2 by elementary signal-flow graph manipulation. For controller implementations, this will be defined as the "direct form II" structure. One clear result emerges: a unit delay must precede the $u(k)$ node. Thus, the $u(k)$ node is *always* a compensator state node. Note that this organization of the computations was only possible due to the zero value of a_0. Thus, our design procedure, allowing $u(k)$ to depend only on past y values, results in a controller which can be implemented if we are careful to include all the actual delays inherent in the structure. From this point on, all compensator structure signal-flow graphs discussed will accurately represent the computation delays that exist. (Note that sample skew would not alter the above signal-flow graph, but only our interpretation of the time index k for $y(k)$ [16]).

Another difference between filter and controller structures should be mentioned. The structure of Fig. 3 has five unit delays, while that of Fig. 1 has only four. This carries over to all types of filter and compensator structures [16]. In fact, for an nth-order transfer function, a delay-canonic filter structure has n unit delays, while a delay-canonic compensator structure has $n + 1$. This fact has an interesting consequence when we look at certain filter structures. One example is the structure composed of cascaded direct form I second-order sections (see Fig. 4). For a fourth-order transfer function, such a *filter* structure has six unit delays and is not canonic (a canonic structure would have four delays). However, such a compensator structure for a fourth-order system ($n = 4$) would have only five delays, which is canonic for compensators [16] (see Fig. 5). This again brings out some of the differences between the filter and compensator cases.

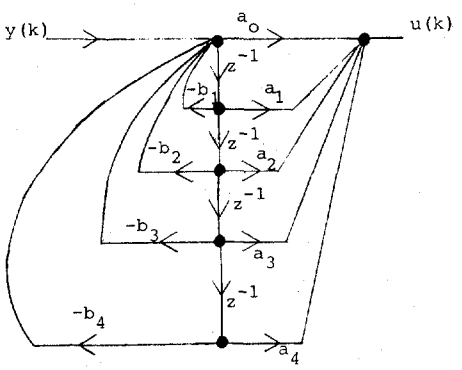


Fig. 1. Direct form II filter structure.

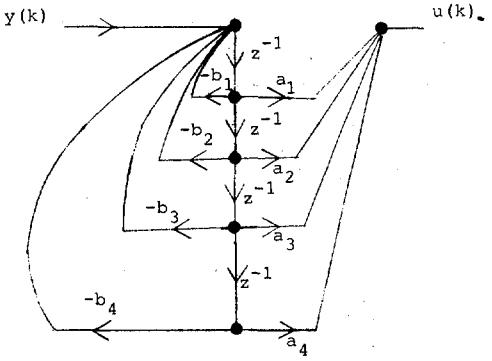


Fig. 2. Direct form II filter structure ($a_0 = 0$).

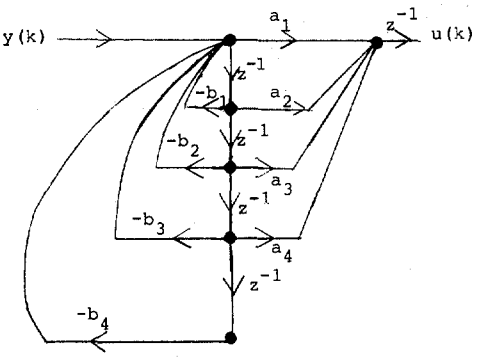


Fig. 3. Direct form II compensator structure.

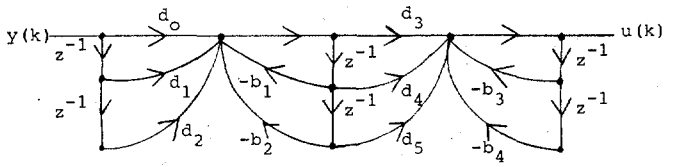


Fig. 4. Direct form I filter structure.

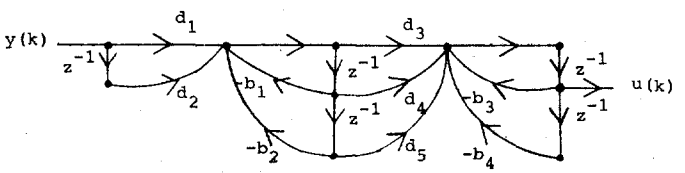


Fig. 5. Direct form I compensator structure.

In addition to representing compensator structures with the signal-flow graph, we need a mathematical notation for describing a structure. In order to accomplish this, we will adapt the filter notation developed by Chan [18] to the case of compensator structures. Chan's notation accounts for the specific multiplier coefficients in the structure, and for the exact sequence, or precedence, to the computations and quantizations involved. Using y and u to represent a filter output and input, respectively, and v the filter states (delay-element outputs), the Chan notation can be written as follows:

$$\begin{bmatrix} v(k+1) \\ y(k) \end{bmatrix} = \Psi_q \Psi_{q-1} \cdots \Psi_1 \begin{bmatrix} v(k) \\ u(k) \end{bmatrix}. \tag{6}$$

Each (finite precision) coefficient in the filter structure occurs once and only once as an entry in one of the Ψ_i matrices. The remainder of the matrix entries are ones and zeros. The precedence to the operations is shown by the ordering of the matrices. The operations involved in computing the intermediate (non-state) nodes

$$r_1(k) = \Psi_1 \begin{bmatrix} v(k) \\ u(k) \end{bmatrix}$$

are completed first, then $r_2(k) = \Psi_2\ r_1(k)$ next, and so forth. The parameter q specifies the number of such *precedence levels.*

For representing compensator structures as discussed above, several changes are necessary. First, u and y are reversed in definition: u is now the compensator output, and y the input. But more importantly, the $u(k)$ node is now a *state* of the compensator. Inclusion of these changes produces the following *modified state space* notation:

$$\begin{bmatrix} v(k+1) \\ u(k+1) \end{bmatrix} = \Psi_q \Psi_{q-1} \cdots \Psi_1 \begin{bmatrix} v(k) \\ u(k) \\ y(k) \end{bmatrix}. \tag{7}$$

Examples of the modified state space representation can be found in Section VII and in [16].

Notationally, it is also useful to define Ψ_∞ to be the infinite precision product of $\Psi_q, \Psi_{q-1}, \cdots, \Psi_1$, and to partition it as follows:

$$\Psi_\infty = [\Psi_{11} \vdots \Psi_{12}] \tag{8}$$

where Ψ_{11} is $(n+1) \times (n+1)$ and Ψ_{12} is $(n+1) \times 1$.

We can summarize the main point presented in this section as follows. An nth-order filter transfer function can be implemented with n unit delays (states). However, the nth-order transfer function of a compensator (for an nth-order system model) requires a compensator structure representation with $n+1$ unit delays. This altered form of a structure reflects the exact consideration of the computation delays that must exist in any digital implementation.

IV. Dynamic Range Constraints and Scaling

Researchers in digital signal processing have treated at length the need for scaling to reduce the dynamic range of the node

signals within a digital structure employing fixed-point arithmetic operations [1], [3], [18]-[21], [24]. The tradeoff between overflow and roundoff quantization noise must be resolved before we can measure and compare the roundoff noise performance of different structures.

Several methods of scaling digital filters exist. One seems particularly appropriate for use with stochastic systems. This stochastic (l_2) scaling procedure [19]-[21] deals with the probability of overflow at each node within the structure. Specifically, assuming uniform internal wordlength, if the filter input is zero-mean white Gaussian noise with a given variance, then scalers are selected such that the probability of overflow at each node is identical to the probability of overflow at the input. This is accomplished by equalizing the variance at each node. Chan [18] has applied l_2 scaling to digital filters, using his state space notation.

Certainly, one obvious approach to the scaling of digital compensators would be to follow the method developed by Chan, but with alterations to include the modified state space notation introduced in Section III. In fact, we have shown this in detail in [16]. However, there is one important problem that arises. Any treatment of a compensator as a stand-alone filter ignores the remainder of the feedback loop through the plant. This can have serious implications. The scaling technique is based on a white noise (compensator) input and seeks to equalize probability of overflow at each node; however, the compensator input (the system output with additive noise) is not white, due to the global feedback loop. Thus, the probabilities of overflow will not be equalized by such an open-loop scaling procedure. For example, it is possible to have an open-loop unstable compensator in a stable LQG system. In this case, the scaling equations may have no solution (blow up) depending on the pole locations.

Therefore, since the variance of each compensator node will depend on the closed loop, the closed-loop performance of the compensator must be accounted for. Thus, we have adapted the l_2 scaling procedure for digital compensators as follows. Recall the plant and compensator equations (1) and (7). To compute the system's closed-loop performance, we must combine the state and compensator equations into a single augmented state space for the overall closed-loop system:

$$\begin{bmatrix} x(k+1) \\ v(k+1) \\ u(k+1) \end{bmatrix} = A \begin{bmatrix} x(k) \\ v(k) \\ u(k) \end{bmatrix} + \begin{bmatrix} w_1(k) \\ \Psi_{12} w_2(k) \end{bmatrix}$$

where (9)

$$A = \left[\begin{array}{c|c|c} \Phi & 0_n & \Gamma \\ \hline \Psi_{12} L & \Psi_{11} & \end{array}\right]$$

and 0_n represents an all-zero $n \times n$ matrix and Ψ_{11}, Ψ_{12} represent the unscaled compensator as partitioned in (8).

Given this complete form (9), let us now follow in general the basic l_2 scaling procedure as applied by Chan [18]. Scaling will correspond to diagonal transformations of the Ψ_i matrices, analogous to the similarity transformation in linear system theory. Let the scaled compensator have the modified state space parameters $\tilde{\Psi}_q, \cdots, \tilde{\Psi}_1$. Then

$$\tilde{\Psi}_i = S_i \Psi_i (S_{i-1})^{-1} \quad \text{for } i = q, \cdots, 1$$

where (10)

$$S_0 = \begin{bmatrix} S_q & 0 \\ 0 & 1 \end{bmatrix}$$

and all S_i are diagonal. We now describe a modification of Chan's procedure for determining the elements of the S_i.

The (unscaled) state covariance matrix Z for (9) can be written (similar to [19]-[21]) as the solution of a steady-state Lyapunov equation

$$Z = AZA' + C \tag{11}$$

where Z is the $(2n+1) \times (2n+1)$ covariance matrix

$$Z = E\left\{ \begin{pmatrix} x \\ v \\ u \end{pmatrix} (x'v'u') \right\}$$

and

$$C = \begin{bmatrix} \Theta_1 & 0 \\ 0 & \Psi_{12} \Theta_2 \Psi_{12}' \end{bmatrix}.$$

At this point, we must break from the usual l_2 scaling procedure, since the plant states, of course, cannot be scaled (nor do we wish to scale them). The usefulness of (11) is that it gives us an expression for the compensator node variances. Let us partition Z as follows:

$$Z = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{12}' & Z_{22} \end{bmatrix} \tag{12}$$

where Z_{11} is $n \times n$. Since we wish to equalize the node variances with the variance of y, let us compute σ_y^2. Since $y = Lx$,

$$\sigma_y^2 = L Z_{11} L'. \tag{13}$$

From (7), we see that the node covariances will depend on the covariances of v, u, and y. Combining (1) and (12), we produce

$$E\left\{ \begin{pmatrix} v \\ u \\ y \end{pmatrix} (v'u'y') \right\} = \begin{bmatrix} Z_{22} & Z_{12}'L' \\ LZ_{12} & LZ_{11}L' \end{bmatrix}. \tag{14}$$

Using (7) and (14), the covariance matrix of the nodes r_1 can be written as

$$E(r_1 r_1') = \Psi_1 \begin{bmatrix} Z_{22} & Z_{12}'L' \\ LZ_{12} & LZ_{11}L' \end{bmatrix} \Psi_1'. \tag{15}$$

In general, for the ith intermediate nodes,

$$E(r_i r_i') = \Psi_i \Psi_{i-1} \cdots \Psi_1 \begin{bmatrix} Z_{22} & Z_{12}'L' \\ LZ_{12} & LZ_{11}L' \end{bmatrix} \Psi_1' \cdots \Psi_{i-1}' \Psi_i'. \tag{16}$$

With this information, we can again apply the methods used by Chan [18] and others to compute a set of matrices $K_i, \cdots, K_q$ [note the normalization by σ_y^2].

$$K_i = \Psi_i \Psi_{i-1} \cdots \Psi_1 \begin{bmatrix} \frac{Z_{22}}{\sigma_y^2} & \frac{Z_{12}'L'}{\sigma_y^2} \\ \frac{LZ_{12}}{\sigma_y^2} & 1 \end{bmatrix} \Psi_1' \cdots \Psi_{i-1}' \Psi_i'. \quad (17)$$

The diagonal elements of the K_i matrices represent the gains from the variance of y to the specific intermediate node variance. By this definition, K_q must then represent the gains from the y variance to the variances of the compensator state nodes v and u, and thus should equal Z_{22}/σ_y^2. This can be shown to be true [16], [18]. As in [18], the scaling matrices S_i shown in (10) are computed as

$$[S_i]_{jj} = ([K_i]_{jj})^{-1/2} \qquad i = 1, \cdots, q \quad \text{for all } j. \quad (18)$$

Thus, the *scaled* system $\tilde{K}_i$ matrices will have unit diagonals; all node variances will be equal to each other and to the variance of y. This ensures the desired equal overflow probabilities.

Two other points of difference between filter and compensator scaling arise. First, the issue of A/D and D/A scaling must be addressed. Note that the above scaling procedure actually *scales* the output node u. In order to keep the closed loop system transfer function unchanged due to this scaling, a gain factor of ρ must be included in the D/A conversion of u. From (10),

$$\rho = [S_q]^{-1}_{n+1,n+1}. \quad (19)$$

But beyond this point, how do we set the A/D and D/A scale factors in general? The scaling procedure equalized the node probability of overflow. However, the actual *value* of that probability will be determined by the way we employ the A/D converter, in other words, how we set its scale factor. This must be decided by the expected level of transients in the system (due to any external inputs) and by the closed-loop rms fluctuations (variance) of y. This issue will be less complex for digital filters, where there is no external closed loop. Again, to keep our ideal closed-loop response unchanged, the D/A scale factor must counteract any A/D gain. The D/A scale factor could be written

$$k_{da} = \rho / k_{ad}. \quad (20)$$

Whatever is selected for k_{ad}, the compensator scaling parameter matrices S_i do not change; the compensator and converter scalings are independent. Remember that in *comparing* different compensator structures for a given ideal compensator design, only the scaling of the compensator itself will be important.

The final point of difference between filter and compensator scaling concerns the nature of the control system. Most of the LQG configurations, as described in Section II, will have *set points*—in other words, reference inputs for the regulator portion of the control system. Thus, the control action will try to drive the system to some nonzero output, and also to minimize fluctuations around that value. Such set-point regulators [23] will have the same parameter values as described in Section II, independent of the set point. However, any dc offset in the plant output y (the compensator input) certainly will affect the probability of overflow at the compensator input and at all internal compensator nodes, and thus affect the scaling. The l_2 scaling procedure assumed a Gaussian *zero-mean* input to the compensator. This procedure would, unfortunately, not be valid with dc inputs.

Fig. 6 presents the set-point LQG system described in Kwakernaak and Sivan [23], where u_r is the reference input. If we wish to drive the output y to y_r, then u_r must be set to $H_c^{-1}(1)y_r$, where $H_c(z)$ is the closed-loop transfer function from u_r to y:

$$H_c(z) = L(zI - \Phi + \Gamma G)^{-1}\Gamma. \quad (21)$$

Unfortunately, this compensator has a dc input, since the steady-state value of y is nonzero. Thus, as we said, l_2 scaling is not possible.

However, we will show that we can describe the set-point system of Fig. 6 in another way. Let us define ξ, η, and γ to be the *deviations* of the states, input, and output away from their steady-state dc values x_0, u_0, and y_0. Thus, $\xi = x - x_0$, $\eta = u - u_0$ and $\gamma = y - y_0$. Using (1), the following relationship must hold between these steady-state values:

$$\begin{aligned} x_0 &= \Phi x_0 + \Gamma u_0 \\ y_0 &= L x_0. \end{aligned} \quad (22)$$

Now, we can design an LQG compensator for the system deviations, using a model for the deviations:

$$\begin{aligned} \xi(k+1) &= \Phi\xi(k) + \Gamma\eta(k) + w_1(k) \\ \gamma(k) &= L\xi(k) + w_2(k). \end{aligned} \quad (23)$$

This will, of course, produce the same parameters as the LQG design with (1). Now, if we take the resulting system, and substitute for η and γ, we produce Fig. 7.

Thus, it is possible to use an alternate LQG set-point configuration where the compensator input γ has an average value of zero, thereby allowing us to apply stochastic (1_2) scaling. The disadvantage to this alternate configuration is the necessity of having two reference inputs which maintain the precise relationship (22), typically in the presence of plant parameter uncertainty. This disadvantage will vanish whenever the plant has a series integration (at least one pole at the origin), which is a very common occurrence in control systems. Frequently, in fact, an integrator is added before an actuator (part of the plant) to provide desensitivity to constant disturbances. To see the effect of an integrator pole on the configuration of Fig. 7, let us write u_0 as $(L(I - \Phi)^{-1}\Gamma)^{-1}y_0$. However, since the dc gain $L(I - \Phi)^{-1}\Gamma$ is infinite if there are any open-loop integrator poles in the plant (poles at $z = 1$), u_0 is forced to zero. In other words, if the plant has any series integration, the LQG configuration of Fig. 7 need have only one reference input, $y_0 = y_r$, and not two. Note that the configuration of Fig. 6 does not change when the plant has integrator poles; both compensator inputs u and y will still have dc components, and the system as a whole still requires the reference input u_r.

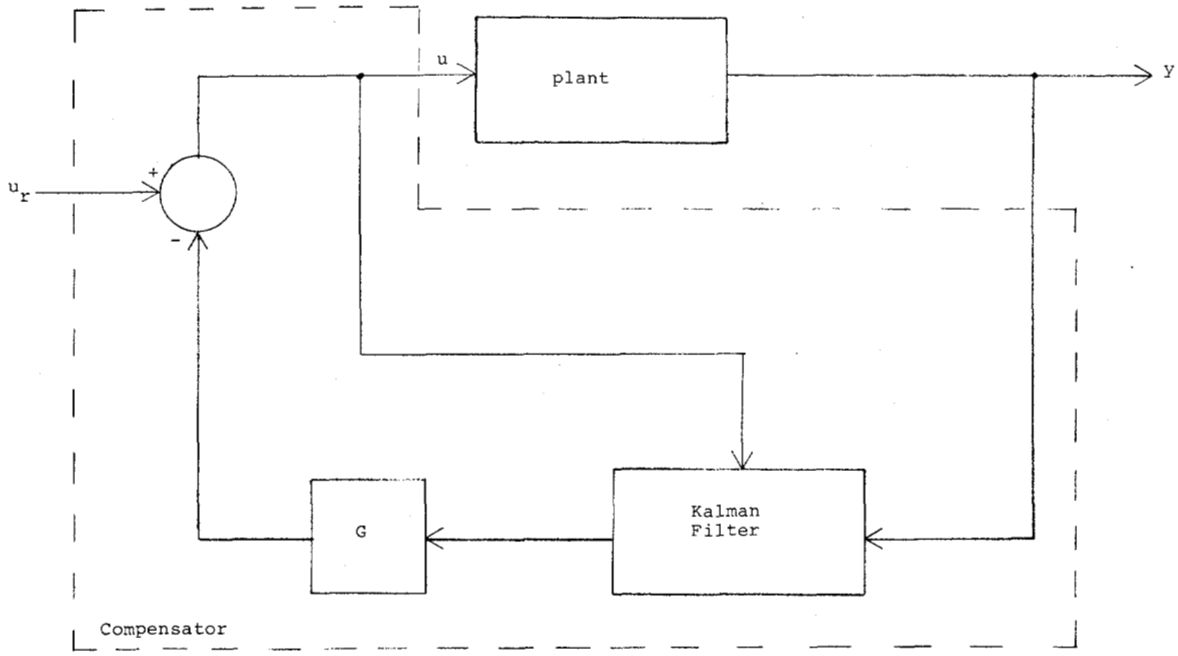


Fig. 6. Set-point compensator configuration.

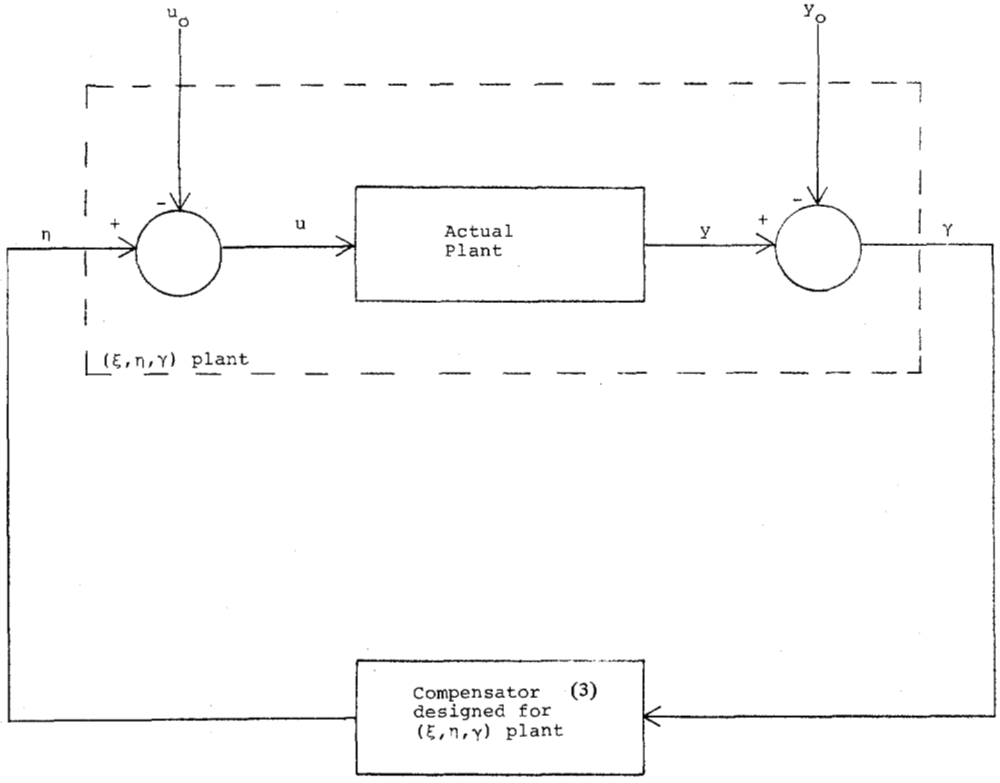


Fig. 7. Alternate LQG set-point configuration.

From this point on, the Fig. 7 configuration is assumed, so that l_2 scaling can be applied.

V. Roundoff Noise Analysis

In this section, we will develop a method for evaluating the effects of roundoff quantization noise in digital compensators. As stated in the Introduction, we will adapt a method used in evaluating digital filter roundoff effects [18]-[21]. As in most digital filtering applications, we will assume a *uniform* wordlength of the internal registers of the compensator. We will also not consider the effect of the input A/D converter quantization on system performance, since this will be independent of the structure chosen. Typically, far fewer bits are required for the converters than for the internal registers.

In most filtering applications, the filter output variance due to roundoff quantization is taken to be the measure of perfor-

mance for the structure [3], [18]-[21], [25]. The analysis of roundoff effects is based on the following assumption. We can represent each roundoff operation on some value x as x plus additive white noise, uniformly distributed between $\pm\Delta/2$ where Δ is the quantization step size. Furthermore, in a structure with many roundoff operations, all such additive noises are assumed to be independent [3]. Given this statistical description of quantization effects, one can easily apply linear system analysis techniques to compute the output node variance due to roundoff.

In examining the effects of roundoff on digital compensators, we can make the same assumptions concerning the quantizer model. However, as with scaling, we cannot blindly follow the procedures used for filters. Again, due to the feedback, the performance of the compensator will depend on the entire closed-loop system, and not on the compensator alone [16]. Furthermore, the variance of the compensator output node is not the best measure of performance for the control system. Instead, it is much more reasonable to use the original metric which figured in the ideal compensator design—the quantity J shown in (2).

Some work has been done previously along related lines. Curry [7] has considered the second moment of the system output error due to rounding for a specific sampled-data control system with a direct form II compensator structure. Knowles and Edwards [6] have also used the additive white noise model for generating a bound on the quantization noise effects of direct form II, cascade, and parallel compensator structures. Sripad [12] has considered the increase in the performance index J due to roundoff, using the additive white noise model, but has not addressed either the scaling issue or the general concepts of compensator structures and representations. The results we present in this section will be more general, since we can consider any type of compensator structure, using the modified state space notation, and since we have accounted for the necessary scaling operation.

The following roundoff analysis procedure results if we consider J to be the performance measure, and if we consider the closed-loop nature of the control system. Let us model the roundoff errors after each compensator multiplication as additive noise. (The structure-independent A/D contribution will be ignored.) The scaled, augmented system of plant *and* compensator, including all the internal roundoff sources, can be written [see (9) and (10)]

$$\begin{bmatrix} x(k+1) \\ \tilde{v}(k+1) \\ \tilde{u}(k+1) \end{bmatrix} = \tilde{A} \begin{bmatrix} x(k) \\ \tilde{v}(k) \\ \tilde{u}(k) \end{bmatrix} + \begin{bmatrix} w_1(k) \\ \epsilon_q(k) + \left(\sum_{i=2}^{q} \tilde{\Psi}_q \cdots \tilde{\Psi}_i \epsilon_{i-1}(k) + \tilde{\Psi}_{12} k_{ad} w_2(k) \right) \end{bmatrix} \tag{24}$$

where

$$\tilde{A} = \left[\begin{array}{c|c|c} \Phi & 0 & \Gamma k_{da} \\ \hline \tilde{\Psi}_{12} L k_{ad} & & \tilde{\Psi}_{11} \end{array}\right].$$

The tilde will refer to scaled quantities, $\epsilon_i(k)$ will represent the noise vector due to the product quantizations associated with the precedence level matrix $\tilde{\Psi}_i$, and q will be the number of precedence levels. Since this is a linear system, superposition can be applied. The noises $w_1(k)$ and $w_2(k)$ are present even in the idealized infinite precision system—they are the uncertainties that produce the ideal design value of J. Thus, if we treat the $\epsilon_i(k)$ noises alone, our analysis will yield the *increase* in J due to internal roundoff quantization. Thus, our system model for roundoff analysis is

$$\begin{bmatrix} x(k+1) \\ \tilde{v}(k+1) \\ \tilde{u}(k+1) \end{bmatrix} = \tilde{A} \begin{bmatrix} x(k) \\ \tilde{v}(k) \\ \tilde{u}(k) \end{bmatrix} + \begin{bmatrix} 0 \\ \epsilon_q(k) + \left(\sum_{i=2}^{q} \tilde{\Psi}_q \cdots \tilde{\Psi}_i \epsilon_{i-1}(k) \right) \end{bmatrix}. \tag{25}$$

Given this model, the resulting (scaled) state covariance matrix can be computed as in Section IV, by solving a Lyapunov equation:

$$\tilde{Z} = \tilde{A} \tilde{Z} \tilde{A}' + \begin{bmatrix} 0 & 0 \\ 0 & \Omega \end{bmatrix} \tag{26}$$

where

$$\Omega = \frac{\Delta^2}{12} \{ \Lambda_q + \tilde{\Psi}_q \Lambda_{q-1} \tilde{\Psi}_q' + \tilde{\Psi}_q \tilde{\Psi}_{q-1} \Lambda_{q-2} \tilde{\Psi}_{q-1}' \tilde{\Psi}_q' + \cdots + \tilde{\Psi}_q \cdots \tilde{\Psi}_2 \Lambda_1 \tilde{\Psi}_2' \cdots \tilde{\Psi}_q' \}.$$

The matrices Λ_i are diagonal matrices whose (j, j) th entry equals the number of noninteger coefficients in the jth row of $\tilde{\Psi}_i$, that is, the number of roundoff error sources associated with the jth component of r_i. This expression assumes that roundoff occurs after every nontrivial product. It would also be possible to produce double-precision products and do double-precision addition. A single roundoff quantization would then be needed to generate the new node value, in which case all the nonzero entries of Λ_i would be ones. This method requires more hardware, but results in a reduced roundoff noise effect.

To relate the state covariance matrix to the performance index J in (2), we must rewrite J as described in [26]. This yields the following expression in terms of the unscaled quantities:

$$\begin{aligned} J &= \text{trace } (Q\overline{xx'}) + 2 \text{ trace } (M\overline{ux'}) + \text{trace } (R\overline{uu'}) \\ &= \text{trace } \Upsilon Z \end{aligned} \tag{27}$$

where

$$\Upsilon = \begin{bmatrix} Q & 0_n & M \\ 0_n & 0_n & 0 \\ M' & 0 & R \end{bmatrix}$$

and Z is defined in (11).

Now let us apply this to the increase dJ in J due to roundoff noise, and then relate dJ to the *scaled* covariance $\tilde{Z}$. Note that

$x(k)$ is not scaled and $\tilde{v}(k)$ does not figure into dJ directly. The only scaled quantity that does enter into the computation is $\tilde{u}$. However, $u = k_{da}\tilde{u}$, or $u = \rho k_{ad}^{-1}\tilde{u}$. Thus, substituting into (27) and considering dJ

$$dJ = \text{trace } \tilde{\Upsilon}\tilde{Z} \tag{28}$$

where

$$\tilde{\Upsilon} = \begin{bmatrix} Q & 0_n & k_{da}M \\ 0_n & 0_n & 0 \\ k_{da}M' & 0 & k_{da}^2 R \end{bmatrix}.$$

Thus, our analysis procedure for the increase dJ due to roundoff involves solving (26) for $\tilde{Z}$, and evaluating (28).

VI. Minimum Roundoff Noise Structures

Mullis and Roberts [19], [20] and Hwang [21] have developed techniques for creating structures with minimum roundoff noise effects. The most practical of these structures is the *block optimal* form. In this filter structure, one assumes the overall structure to be composed of a parallel or series combination of one-precedence level ($q = 1$) second-order sections, each of which is optimized for minimum roundoff noise. The resulting structure will have about $4n$ coefficients, where n is the filter order, which is about twice the number for a direct form II cascade or parallel structure.

Certainly we could treat a compensator as a stand-alone filter and follow the above-mentioned procedure to generate a minimum roundoff noise structure. However, as explained in Sections IV and V, in order to obtain an accurate picture of roundoff effects, we must consider the *closed-loop* system. Thus, certain changes will be required in the procedure.

For filters, the procedure involves the computation of a block transformation matrix, which, when applied to the original unscaled filter structure, generates a *new* structure with minimum roundoff noise. The keys to this transformation are the matrices K_1 and W_1. K_1 is related to the filter scaling procedure discussed in Section IV, and W_1 is a "noise gain" matrix. It reflects the gain (in variance) from each noise source to the output node. Recall that K_1 reflects the gain from the input node to each internal node. Thus, we can apply the Mullis-Roberts and Hwang techniques to digital compensator roundoff noise minimization if we can compute K_1 and W_1 matrices that account for the closed-loop control system. We have already specified the technique for finding K_1 in the section on scaling [see (17)]. In this section, we will develop an expression for W_1.

Since W_1 represents the gain from roundoff noise sources to output variance (for filters), we need a matrix which reflects the gain from these sources to the performance index increase dJ. The following will be adapted from Mullis and Roberts and from Hwang. Let us rewrite (26) in terms of the unscaled compensator parameters Ψ_{11} and Ψ_{12} (as in [17], this assumes $q = 1$).

$$\tilde{Z} = TAT^{-1}\tilde{Z}T^{-1}A'T + \frac{\Delta^2}{12}\begin{bmatrix} 0 & 0 \\ 0 & \Lambda_1 \end{bmatrix} \tag{29}$$

where T includes the scaling matrix S_1,

$$T = \begin{bmatrix} I_n & 0 \\ 0 & S_1 \end{bmatrix},$$

I_n is the $n \times n$ identity matrix, and

$$A = \left[\begin{array}{c|c|c} \Phi & 0_n & \Gamma \\ \hline \Psi_{12}L & & \Psi_{11} \end{array}\right]$$

as before. (We have assumed k_{ad} to be 1; k_{ad} would, at any rate, not affect the optimal structure.) By manipulating (10) and using the definition of Z and T, we can recognize that the matrix Z just equals $T^{-1}\tilde{Z}T^{-1}$. Substituting in (29), we have

$$Z = AZA' + \frac{\Delta^2}{12}\begin{bmatrix} 0 & 0 \\ 0 & \Lambda_1 S_1^{-2} \end{bmatrix}. \tag{30}$$

The expression for the increase in performance index due to roundoff noise for the scaled system can also be written in terms of the unscaled covariance matrix Z [see (27)].

$$\begin{aligned} dJ &= \text{trace}\{\tilde{\Upsilon}\tilde{Z}\} = \text{trace}\{\tilde{\Upsilon}T^{-1}ZT^{-1}\} \\ &= \text{trace}\{T^{-1}\tilde{\Upsilon}T^{-1}Z\} \\ &= \text{trace}\{\Upsilon Z\}. \end{aligned} \tag{31}$$

Using an adjoint Lyapunov equation (see the Appendix), (30) and (31) can be replaced by

$$dJ = \frac{\Delta^2}{12}\text{ trace}\left\{\begin{bmatrix} 0 & 0 \\ 0 & \{\Lambda_1 S_1^{-2}\} \end{bmatrix} W\right\} \tag{32}$$

where

$$W = A'WA + \Upsilon. \tag{33}$$

The trace expression in (32) can be simplified if we define W_1 to be the lower right-hand $(n + 1) \times (n + 1)$ portion of W:

$$dJ = \frac{\Delta^2}{12}\text{ trace}\{\Lambda_1 S_1^{-2} W_1\}. \tag{34}$$

W_1 is the matrix needed to apply the Mullis-Roberts and Hwang techniques for generating the optimal transformation matrix. Using K_1 and W_1, as presented above and in Section IV, we can follow the remainder of their technique to generate a one-level minimum roundoff noise compensator structure. It may, of course, be quite different from that resulting from a treatment of the compensator as a stand-alone filter.

Conceptually, the technique described above could be extended to multiple levels. However, the iterative structure optimization procedure developed by Chan [18] for filters is far more useful for minimizing roundoff noise for general structures. In Chan's method, an initial structure is subjected to continuous transformations, each of which reduces some overall objective function, given constraints on certain coefficient values. For filters, an equation similar to (34) (but without the closed-loop) is used as the objective function. Almost any of the coefficients of the initial structure can be held fixed during the optimization process. Thus, we can use this method to trade off an increase in the number of coefficients

versus performance. The extension of this very useful procedure to low-noise compensator design is presented in [16].

VII. An LQG Example

A specific sixth-order LQG example was chosen so that we could examine the roundoff noise performance of several structures using the techniques developed above [16]. This example was adapted from the longitudinal control system design done for the F8 digital fly-by-wire fighter [27]. The continuous-time plant parameters and performance index parameters are given below.

Continuous Time System Parameters:

$$A = \begin{bmatrix} -0.6696 & 5.7 \times 10^{-4} & -9.01 & 0 & -15.77 & 0 \\ 0 & -0.01357 & -14.11 & -32.2 & -0.433 & 0 \\ 1 & -1.2 \times 10^{-4} & -1.214 & 0 & -0.1394 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -12 & 12 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$B = [0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 1]$$

$$C = [1 \;\; 0.003091 \;\; 31.28 \;\; 1 \;\; 3.592 \;\; 0].$$

Continuous-Time Performance Index Parameters:

$$\hat{Q} = \begin{bmatrix} 6.637 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2.6554 \times 10^{-7} & 2.686 \times 10^{-3} & 0 & 3.085 \times 10^{-4} & 0 \\ 0 & 2.686 \times 10^{-3} & 27.174 & 0 & 3.121 & 0 \\ 0 & 0 & 0 & 27.174 & 0 & 0 \\ 0 & 3.085 \times 10^{-4} & 3.121 & 0 & 0.3585 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\hat{R} = 5.252.$$

Continuous-Time Noise Covariances:

$$\Xi_1 = \text{diag}\,[0 \;\; 0 \;\; 0 \;\; 0 \;\; 10^{-6} \;\; 10^{-6}]$$

$$\Xi_2 = 0.0018441.$$

This continuous-time system was discretized at a sample rate of 10 Hz and the optimal regulator and Kalman filter designed. The double-precision parameters Φ, Γ, L, Q, M, R, Θ_1, Θ_2, G, and K can be found in [16].

Before discussing the different structures tested, it will be helpful to mention the A/D noise contribution for this example (independent of structure). If we allow a 5 percent increase in J due to this single noise source, then a procedure similar to that outlined in Section V requires a 4.98 bit A/D wordlength. Including a sign bit, and selecting the next largest integer value, our actual wordlength would be 6 bits. As will be shown, this bears out the need for shorter A/D wordlengths as compared to internal wordlengths.

Five types of structures for implementing the ideal compensator transfer function (4) were examined. These were the digital filtering-based direct form II structure for compensators (see Section III), several cascade structures consisting of direct form II or direct form I second-order sections, several parallel structures composed of direct form II sections, a block-optimal minimum roundoff noise structure, and a "default" compensator structure which we will call the *simple* structure. In all five cases, we will indicate the initial design coefficient locations *before* implementing the l_2 scaling procedure of Section IV. This will generally alter the initial values, and will frequently create a few extra scaling coefficients. In any case, where a unity entry in the unscaled structure has become a multiplier coefficient (nonunity, nonpower-of-two) when scaled, we have indicated this with an asterisk.

The first structure we will examine is the direct form II compensator structure. Fig. 3 presents a signal-flow graph for the fourth-order case ($n = 4$). Note the presence of the delay preceding the output node. For the sixth-order case, the 12 coefficients of the direct form II structure come directly from the unfactored transfer function (35). Its modified state space representation (two precedence levels) is shown in (36).

$$H(z) = \frac{a_1 z^{-1} + a_2 z^{-2} + a_3 z^{-3} + a_4 z^{-4} + a_5 z^{-5} + a_6 z^{-6}}{1 + b_1 z^{-1} + b_2 z^{-2} + b_3 z^{-3} + b_4 z^{-4} + b_5 z^{-5} + b_6 z^{-6}} \tag{35}$$

$$\Psi_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ a_6 & a_5 & a_4 & a_3 & a_2 & a_1 \end{bmatrix}$$

$$\Psi_1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ -b_6 & -b_5 & -b_4 & -b_3 & -b_2 & -b_1 & 0 & 1^* \end{bmatrix} \tag{36}$$

When this structure is l_2 scaled, the unity value marked with an asterisk will become a 13th nonunity coefficient.

The second type of structure, the cascade, derives its coefficients from a multiplicative factorization into three series second-order sections. The factored transfer function, assuming direct form II sections, is

$$H(z) = \frac{(d_1 z^{-1} + d_2 z^{-2})(1 + d_3 z^{-1} + d_4 z^{-2})(1 + d_5 z^{-1} + d_6 z^{-2})}{(1 + c_1 z^{-1} + c_2 z^{-2})(1 + c_3 z^{-1} + c_4 z^{-2})(1 + c_5 z^{-1} + c_6 z^{-2})}. \tag{37}$$

Such a structure will have 12 coefficients, four precedence levels, and will require three additional scaling multipliers when l_2-scaled [see (38)].

$$\Psi_4 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & d_6 & d_5 & 1^* \end{bmatrix}$$

$$\Psi_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & d_4 & d_3 & -c_6 & -c_5 & 1^* \end{bmatrix}$$

$$\Psi_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ d_2 & -c_4 & -c_3 & 0 & 0 & d_1 \end{bmatrix}$$

$$\Psi_1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ -c_2 & -c_1 & 0 & 0 & 0 & 0 & 0 & 1^* \end{bmatrix}. \tag{38}$$

Details are available in [16]. Note that several cascade structures can be formed simply by grouping the poles and zeros differently, or by ordering the sections differently, or even by implementing each section differently [16]. Unlike digital filters, the typical presence of more than one real pole in a digital compensator further complicates the choices. In each second-order section, we will also have to *pair* different real poles together. Two cascade structures will be considered, both using direct form II sections, but with a different pairing and ordering. In addition, a direct form I structure will be tested. (See Fig. 5 for a fourth-order example.) Its modified state space is given in [16].

The third type of structure, the parallel form, corresponds to a partial-fraction expansion of (35), which allows the use of parallel first- and/or second-order sections. For the case of five *parallel* direct form II sections, four first-order and one second order, the expanded transfer function (also having 12 coefficients before scaling) is shown in (39), and its modified state space is given in (40).

$$H(z) = \frac{e_1 z^{-1} + e_2 z^{-2}}{1 + c_1 z^{-1} + c_2 z^{-2}} + \frac{e_3 z^{-1}}{1 + g_3 z^{-1}} + \frac{e_4 z^{-1}}{1 + g_4 z^{-1}} + \frac{e_5 z^{-1}}{1 + g_5 z^{-1}} + \frac{e_6 z^{-1}}{1 + g_6 z^{-1}}. \tag{39}$$

$$\Psi_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ e_2 & e_1 & e_3 & e_4 & e_5 & e_6 \end{bmatrix}$$

$$\Psi_1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -c_2 & -c_1 & 0 & 0 & 0 & 0 & 0 & 1^* \\ 0 & 0 & -g_3 & 0 & 0 & 0 & 0 & 1^* \\ 0 & 0 & 0 & -g_4 & 0 & 0 & 0 & 1^* \\ 0 & 0 & 0 & 0 & -g_5 & 0 & 0 & 1^* \\ 0 & 0 & 0 & 0 & 0 & -g_6 & 0 & 1^* \end{bmatrix}. \tag{40}$$

To scale this structure, five additional scalers (one per section) are required. Of course, there are many other parallel forms, depending on how one groups the poles and zeros, and how one implements each section [16]. We will also examine two parallel structures having only second-order direct form II sections. The two structures differ only in the pairing of the four real poles of the compensator.

The fourth type of structure tested is a parallel block optimal minimum roundoff noise structure, as described in Section VI. This structure will have 25 coefficients, many more than the previous three types. Its modified state space is given in (41),

and shows three second-order sections and only one precedence level [16].

$$\Psi_1 = \begin{bmatrix} f_1 & f_2 & 0 & 0 & 0 & 0 & 0 & f_3 \\ f_4 & f_5 & 0 & 0 & 0 & 0 & 0 & f_6 \\ 0 & 0 & f_7 & f_8 & 0 & 0 & 0 & f_9 \\ 0 & 0 & f_{10} & f_{11} & 0 & 0 & 0 & f_{12} \\ 0 & 0 & 0 & 0 & f_{13} & f_{14} & 0 & f_{15} \\ 0 & 0 & 0 & 0 & f_{16} & f_{17} & 0 & f_{18} \\ f_{19} & f_{20} & f_{21} & f_{22} & f_{23} & f_{24} & f_0 & f_{25} \end{bmatrix}. \quad (41)$$

The last structure we will consider has no analog in the filtering literature. This structure would result if we tried to directly implement the compensator equations shown in (3). It is important to analyze this type of "simple" structure, because a naive implementation might employ it more or less by default, as it is the natural result of the ideal design. Taking the equations shown in (3), we must first rewrite them in an implementable form. (Note that $u(k+1)$ cannot be computed from $\hat{x}(k+1)$, since some computation delay must be allowed for the multiplication by G.) Substituting for $\hat{x}(k+1)$, we get

$$\hat{x}(k+1) = \Phi\hat{x}(k) + \Gamma u(k) + K(y(k) - L\hat{x}(k))$$
$$u(k+1) = -G\Phi\hat{x}(k) - G\Gamma u(k) - GK(y(k) - L\hat{x}(k)). \quad (42)$$

These equations do represent a feasible compensator structure. Its modified state space representation would have three precedence levels:

$$\Psi_3\Psi_2\Psi_1 = \left[\begin{array}{c} I_6 \\ \hline -G \end{array}\right] \left[\begin{array}{c|c|c} \Phi & \Gamma & K \end{array}\right] \left[\begin{array}{c|c|c} I_6 & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline -L & 0 & 1 \end{array}\right]. \quad (43)$$

The important disadvantage of this type of structure can be easily seen from (43). For a sixth-order example, there will be up to 60 coefficients.

The actual scaled coefficient values for these nine structures are presented in [16]. Table I summarizes the roundoff noise results for these structures. As mentioned previously, the A/D noise contribution is the same for all structures and is not included.

The "levels" column lists the number of precedence levels, and the "N" column lists the number of coefficients including scalers in the structure. The roundoff noise results are presented in terms of the number of signal bits (wordlength) that are required to hold the increase in J due to product roundoff noise to 5 percent of the ideal value. Again, these numbers do not include the sign bit. Two wordlengths are presented for each structure. The left-hand column (larger) corresponds to the case of roundoff after every nontrivial multiplication and then the use of single-precision adders, while the right-hand column corresponds to the case of double-precision adders and quantization *after* addition.

From Table I, we can see that the different pole pairings associated with parallel structures c) and d) produced results

TABLE I
ROUNDOFF NOISE RESULTS

Structure	Levels	N	Wordlength spa	Wordlength dpa
a) direct form II	2	13	19.65	18.25
b) parallel direct form II	2	17	8.05	7.45
c) parallel direct form II	2	15	10.18	9.39
d) parallel direct form II	2	15	14.74	13.94
e) block optimal parallel	1	25	7.88	7.06
f) cascade direct form II	4	15	15.69	14.68
g) cascade direct form II	4	15	10.51	9.47
h) cascade direct form I	3	14	15.52	14.36
i) simple, default structure	3	50	9.01	7.54

that differed by 4.5 bits. Placing the near-unit magnitude real poles of the compensator in different sections was significantly superior. Of the two similar cascades f) and g), the one with these same two poles in different sections required 5.2 fewer bits. As with filters, the pairing/ordering issue is clearly *not* a trivial question. Also, note that the cascade of direct form I sections h), which used the same pairing/ordering as g), has nearly identical performance, but one less precedence level and one less coefficient. Unlike digital filter applications, it is worth considering for feedback compensator implementation.

Structure b), the combination of first- and second-order parallel sections, with its 17 coefficients, outperformed every other structure except the block optimal. Even so, the extra eight coefficients of the block optimal structure with second-order sections only gained 0.2 bits of performance over this structure. Thus, when evaluating different structures, it is important to know the block optimal result (for various pairings) so that we can judge whether a suboptimal structure like b) is effective enough. In this case, it clearly is.

As expected from the literature on digital filters, the direct form II has very poor noise performance. It is also very important to note that even though the simple structure i) performed fairly well [but still one bit worse than the structure b)], it has (comparatively) *far* too many multiplier coefficients. Were such a structure to be built, this surfeit of coefficients would require either a slower sampling rate or a more expensive compensator than b) or e), either of which would *outperform* it.

The second wordlength column in Table I shows the gain possible when using double precision adders and fewer quantizers. Depending on the structure tested, a savings of from 0.6 to 1.47 bits was realized. Whether this small savings is enough to justify the higher-precision adders will depend on the particular application.

Calculation time can be derived from Table I in a general way. If multiplications are done in software, serially, then the calculation time will be roughly proportional to the number of multiplies. If as many parallel multiply modules are available as could be used, then the calculation time will be proportional to the number of precedence levels, since all multiplies in each precedence level could be done in parallel. A more detailed discussion of this issue can be found in [16].

VIII. SUMMARY

We have shown that the implementation of digital compensators can benefit greatly from the concepts developed for

dealing with different computational structures and finite wordlength effects in digital filters. However, due to the nature of the control problem, new issues arise that must be considered. Thus, we have had to adapt and extend these digital signal processing concepts and techniques.

First, the importance of unaccounted delay in a control system required that we rethink the notion of a compensator structure. Any computational delay that will be present must be included in our model, so that the closed-loop system can perform as close to the ideal design as possible. When dealing with the LQG control problem, this led to a specific ideal design procedure. Any structure for implementing the resulting compensator can be shown to have its output node as a *state*, that is, the output of a unit delay element. Thus, the structures used commonly for filters must be modified when applied to compensators.

The concept of scaling for fixed-point digital compensators also had to be reconsidered. Since the entire closed-loop system will affect the internal compensator node variance, any scaling technique must take into account the overall closed-loop response. That is, it is inappropriate to treat the compensator as a stand-alone filter. Thus, the l_2 scaling technique was adapted for LQG compensators. In addition, the form of a set-point regulator control system was shown to be important in determining the type of scaling that would be effective. An alternate set-point configuration that would allow the use of l_2 scaling was proposed.

Once scaling has been accomplished, we can begin to analyze the effects of quantization noise on control system performance. As with scaling, the compensator cannot be treated as a stand-alone filter. Thus, we adapted the filter analysis techniques to include the plant and feedback loop, and to use the natural LQG performance index as our figure of merit rather than output noise variance due to roundoff. Minimum roundoff noise structures, such as those developed for filters by Mullis and Roberts, Hwang, and Chan, were treated next. These structural design techniques also needed modification to include the effects of the closed-loop on system performance.

Finally, several example structures for a single LQG compensator were scaled and their roundoff performances compared. The pairing and ordering issues involved with parallel and cascade type structures were shown to be even more complex for compensators, due to the number of real poles that are common in control system compensators. It was also shown that the default type of structure for LQG controllers is a poor choice of structure for the LQG compensator. Its extremely large number of multiplier coefficients would impose unnecessary speed limitations or complexity on the compensator and the control system in which it is embedded. In fact, this result points out the need for considering the issues we have dealt with in this paper. The specific results we have presented (for the F8 example) are not intended to be representative of *all* such controller results. The main intention is to show the importance of a consideration of the issues arising from the digital implementation of such controllers.

Appendix

If we take the trace of the product of two matrices to be an inner product on the space of matrices, and π to be a matrix operator, then

$$\text{trace}\,(\pi(X)\,U) = \text{trace}\,(X\pi^*(U)) \tag{A1}$$

where π^* is the adjoint operator of π. For $\pi(X) = X - AXA'$, the operator π^* can be derived from (A1):

$$\begin{aligned}\text{trace}\,((X - AXA')U) &= \text{trace}\,(XU) - \text{trace}\,(AXA'U)\\ &= \text{trace}\,(XU) - \text{trace}\,(XA'UA)\\ &= \text{trace}\,(X(U - A'UA)).\end{aligned} \tag{A2}$$

Thus, $\pi^*(U) = U - A'UA$.

As used in Section VI, the Lyapunov equation (30) and the trace (31) were replaced by the equivalent equations (32) and (33). Relating this to the derivation above,

$$X = Z$$

$$U = W$$

$$\pi(X) = \frac{\Delta^2}{12}\begin{bmatrix}0 & 0\\ 0 & \Lambda_1 S_1^{-2}\end{bmatrix}$$

$$\pi^*(U) = \Upsilon.$$

References

[1] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

[2] T.A.C.M. Claasen, W.F.G. Mecklenbräuker, and J.B.H. Peek, "Effects of quantization and overflow in recursive digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 517–529, Dec. 1976.

[3] A. V. Oppenheim and C. J. Weinstein, "Effects of finite register length in digital filtering and the fast Fourier transform," *Proc. IEEE*, vol. 60, pp. 957–976, Aug. 1972.

[4] A. S. Willsky, *Digital Signal Processing and Control and Estimation Theory–Points of Tangency, Areas of Intersection, and Parallel Directions*. Cambridge, MA: M.I.T. Press, 1979.

[5] A. Fettweis and K. Meerkötter, "On parasitic oscillation in digital filters under looped conditions," *IEEE Trans. Circuits Syst.*, vol. CAS-24, pp. 475–481, Sep. 1977.

[6] J. B. Knowles and R. Edwards, "Effect of a finite-word-length computer in a sampled-data feedback system," *Proc. IEE*, vol. 112, pp. 1197–1207, June 1965.

[7] E. E. Curry, "The analysis of round-off and truncation errors in a hybrid control system," *IEEE Trans. Automat. Contr.*, vol. AC-13, pp. 601–604, Oct. 1967.

[8] J. E. Bertram, "The effect of quantization in sampled-feedback systems," *Trans. AIEE*, vol. 77, Part 2, pp. 177–182, Sep. 1958.

[9] J. B. Slaughter, "Quantization errors in digital control systems," *IEEE Trans. Automat. Contr.*, vol. AC-9, pp. 70–74, Jan. 1964.

[10] G. W. Johnson, "Upper bound on dynamic quantization error in digital control systems via the direct method of Lyapunov," *IEEE Trans. Automat. Contr.*, vol. AC-10, pp. 439–448, Oct. 1965.

[11] G.N.T. Lack and G. W. Johnson, "Comments on 'Upper bound on dynamic quantization error in digital control systems via the direct method of Lyapunov,' " *IEEE Trans. Automat. Contr.*, vol. AC-11, pp. 331–334, Apr. 1966.

[12] A. B. Sripad, "Models for finite precision arithmetic, with application to the digital implementation of Kalman filters," Sc.D. dissertation, Sever Inst., Washington Univ., St. Louis, MO, Jan. 1978.

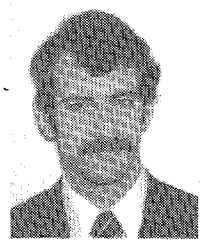
[13] R. E. Rink and H. Y. Chong, "Performance of state regulator systems with floating-point computation," *IEEE Trans. Automat. Contr.*, vol. AC-24, pp. 411–421, June 1979.

[14] F. A. Farrar, "Microprocessor implementation of advanced control modes," in *Proc. Summer Comput. Simulation Conf.*, Chicago, IL, July 1977, pp. 339–342.

[15] P. Moroney, A. S. Willsky, and P. K. Houpt, "The digital implementation of control compensators: The coefficient wordlength issue," *IEEE Trans. on Automat. Contr.*, vol. AC-25, pp. 621–630, Aug. 1980.

[16] P. Moroney, "Issues in the digital implementation of control compensators," Ph.D. dissertation, Dep. Elec. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, Sept. 1979.

[17] S. R. Parker and S. F. Hess, "Limit cycle oscillations in digital filters," *IEEE Trans. Circuit Theory*, vol. CT-18, pp. 687–697, Nov. 1971.
[18] D.S.K. Chan, "Theory and implementation of multidimensional discrete systems for signal processing," Ph.D. dissertation, Dep. Elec. Eng. Comput. Sci., Massachusetts Inst. Technol., Cambridge, May 1978.
[19] C. T. Mullis and R. A. Roberts, "Synthesis of minimum roundoff noise fixed-point digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-23, pp. 551–562, Sept. 1976.
[20] ——, "Filter structures which minimize roundoff noise in fixed-point digital filters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Philadelphia, PA, Apr. 1976, pp. 505–508.
[21] S. Y. Hwang, "Minimum uncorrelated unit noise in state-space digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 273–281, Aug. 1977.
[22] A. P. Sage, *Optimal Systems Control.* Englewood Cliffs, NJ: Prentice-Hall, 1968.
[23] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems.* New York: Wiley, 1972.
[24] L. B. Jackson, "On the interaction of roundoff noise and dynamic range in digital filters," *Bell Syst. Tech. J.*, vol. 49, pp. 159–184, Feb. 1970.
[25] S. Y. Hwang, "Roundoff noise in state-space digital filtering: A general analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 256–262, June 1976.
[26] G. K. Roberts, "Consideration of computer limitations in implementing on-line controls," Massachusetts Inst. Technol., Cambridge, MA, ESL Rep. ESL-R-665, June 1976.
[27] A. E. Bryson, Jr., Guest Ed., *IEEE Trans. Automat. Contr.*, Mini-Issue on the F-8 DFWB, vol. AC-22, pp. 752–806, Oct. 1977.



Paul Moroney (M'79) was born in Auburn, NY, on August 27, 1952. He received the B.S., M.S., Engineer's, and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, MA in 1974, 1977, 1977, and 1979, respectively.

From September 1974 until May 1976 he held positions as a Teaching Assistant and then a Research Assistant at M.I.T. From September 1977 until September 1979 he was a Draper Fellow at the Charles Stark Draper Laboratory, Cambridge. He has held summer positions with Varian Associates (1973), NERComP, Inc. (1974), Watkins-Johnson (1976), and the Draper Laboratory (1977). Since September 1979, he has been with the Linkabit Corporation, San Diego, CA, working in the area of digital communications.

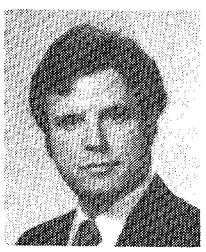
Dr. Moroney is a past member of Eta Kappa Nu, Tau Beta Pi, and Sigma Xi.

Alan S. Willsky (S'70–M'73–SM'82) was born in Newark, NJ, on March 16, 1948. He received the S.B. degree in aeronautics and astronautics and the Ph.D. degree in instrumentation and control from the Massachusetts Institute of Technology, Cambridge, in 1969 and 1978, respectively.

From 1969 through 1973 he held a Fannie and John Hertz Foundation Fellowship. He joined the faculty of the M.I.T. Department of Electrical Engineering and Computer Science in 1973 and is presently an Associate Professor of Electrical Engineering. From 1974 to 1981

he served as Assistant Director of the M.I.T. Laboratory for Information and Decision Systems. He is also a founder and member of the board of directors of Alphatech, Inc. From February through June of 1977 he was a Science Research Council Senior Visiting Fellow at Imperial College, London, England, and from September 1980 through January 1981 he was a Professeur Associé at L'Université de Paris-Sud, Orsay, France. He is Editor of the M.I.T. Press series of books in signal processing, optimization, and control, was the program chairman for the 17th IEEE Conference on Decision and Control held in San Diego, CA, in January 1979, is Associate Editor for Estimation of the IEEE TRANSACTIONS ON AUTOMATIC CONTROL, is an Associate Editor of the journals *Stochastics* and *Control* and *System Letters*, is a member of the Administrative Committee of the IEEE Control Systems Society, and was the Control Systems Society Program Chairman for the 1981 Bilateral Seminar on Control Systems sponsored by the IEEE Control Systems Society and the Chinese Association for Automation. Also, he gave the opening plenary lecture at the 20th IEEE Conference on Decision and Control. He is the author of *Digital Signal Processing and Control and Estimation Theory: Points of Tangency, Areas of Intersection, and Parallel Directions*, (M.I.T. Press, May 1979) and an undergraduate text, *Signals and Systems*, co-authored with Prof. A. V. Oppenheim and Dr. I. T. Young (Prentice-Hall, 1982). He has written numerous papers in algebraic system theory, nonlinear filtering, failure detection, stochastic processes, and biomedical signal processing. His present research interests are in problems involving abrupt changes in signals and systems and the related problems of detection and reliability, the modeling and processing of spatially-distributed random data, large-scale, decision-directed signal processing, and the asymptotic analysis of control and estimation systems.

In 1975 Dr. Willsky received an award from the M.I.T. Graduate Student Council for outstanding teaching, and in August 1975 he received the Donald P. Eckman Award from the American Automatic Control Council. He is a member of SIAM, AAAS, Sigma Xi, Sigma Gamma Tau and Tau Beta Pi.



Paul K. Houpt (S'64–M'66–S'73–M'75) was born in Washington, DC, in 1944. He received the B.S. degree from Syracuse University, Syracuse, NY, in 1966, the M.S. degree from New York University, New York, NY, in 1968, and the Ph.D. degree from the Massachusetts Institute of Technology, Cambridge, in 1974, all in electrical engineering.

From 1966 to 1970, he was with the Electronic Power Systems Division of Bell Laboratories, Whippany, NJ. In 1974 he became a research associate on the staff of the M.I.T. Laboratory for Information and Decision Systems, where his research focused on the application of optimal control and estimation theory to urban vehicular traffic control and automated transit systems. Since 1979, Dr. Houpt has been with the M.I.T. Department of Mechanical Engineering, where he is currently an Associate Professor. His current research interests include computer-based diagnostics and control of internal combustion engines, computer aided control systems design, and fundamental studies of algorithms for digital control. He has served as a consultant on various industrial applications of digital control including automotive engines, HVAC systems, and wind energy systems.