Performance-Rate Functions for Dynamically Quantized Feedback Systems

Michael D. Lemmon and Rong Sun

***Abstract*— This paper studies the performance of dynamically quantized feedback systems. In particular, we examine the relationship between the minimum summed squared quantization error and the rate at which feedback measurements are quantized. The closed loop system's performance (as measured by the summed squared bit rate) can vary greatly for a given bit rate, depending on how the quantization bits are allocated. This paper derives the bit assignment policy that minimizes the summed squared quantization level achievable under a constant bit rate. The proof of the bit assignment's optimality allows us to identify a "performance-rate" function that expresses the best achievable performance as a function of the bit rate.**

I. Introduction

In many computer-controlled systems, the plant's output must be quantized before it can be fed back to the controller. Quantization of the feedback signal can have a dramatic effect on the closed loop system's behavior and in recent years there has been considerable interest in understanding the fundamental limitations that such quantization places on feedback control systems. This paper derives an *optimal* bit assignment for quantized feedback systems using the *uncertainty set* method introduced by Brockett and Liberzon [1]. The bit assignment is optimal in the sense of minimizing the summed squared quantization error. We then identify **performance-rate** functions that characterize the optimal achievable performance as a function of the bit rate. The proposed bit assignment policy is similar to those proposed earlier for scalar [2] and two-dimensional [3] systems. This paper extends the two-dimensional results of [3] to n-dimensional systems with bounded noise.

Static or so-called *memoryless* quantization policies use static codebooks to map real-valued signals onto one of a discrete set of quantization symbols. Delchamps [4] demonstrated that static quantization policies required an infinite number of quantization levels to assure closed loop asymptotic stability. Elia and Mitter [5] later derived the lowest "density" static quantizer assuring asymptotic stability. This quantizer, however, still required an infinite number of bits. With only a finite number bits, it has been repeatedly shown that the best we can hope for using static quantization is the uniform ultimate boundedness of the state [6] [7] [8].

Brockett and Liberzon showed [1] that *dynamic* quantization or so-called quantization with "memory" can achieve asymptotic stability with a finite number of bits [1]. These policies generate a sequence of sets, $\{P[k]\}$ such that each element of the sequence is encoded with a finite number of bits and it can be guaranteed that the system state $x[k]$ lies in $P[k]$ for all k. If the quantization policy guarantees that the "size" of $P[k]$ goes to zero as k goes to infinity, then the quantized system is asymptotically stable. Moreover, since each $P[k]$ is encoded with a finite number of bits, we achieve asymptotic stability with only a finite number of quantization levels. Sufficient conditions for asymptotic stability with a finite number of bits were established in [1] with later extensions in [9]. Tatikonda [10] [11] established necessary and sufficient conditions for asymptotic stability for general linear systems with a finite number of bits. This bound requires that the number of bits Q used in the feedback path must satisfy the following bound

$$Q \geq \sum_{i=1}^{n} \max\left[0, \log_2 \lambda_i\right] \tag{1}$$

to assure asymptotic stability. In equation 1, n is the system's dimension and λ_i is the ith eigenvalue of the system matrix. Similar bounds were established by Nair and Evans [12] for general linear systems in the stochastic sense.

Most of the aforecited work only focuses on the impact of quantization on closed loop stability. Since many computer-controlled systems quantize feedback signals at rates much greater than the rate suggested in equation 1, it may be more important to study the effect that quantization has on system *performance*. Early work in this direction focused on static quantization policies in which the quantization error was treated as a noise term. This noise analysis becomes less valid as a system approaches its equilibrium point and subsequent work used describing-function methods to study limit cycle effects in quantized sampled data systems. Astrom and Wittenmark [13] provide a high-level survey of this earlier work with appropriate references. More recent work has studied the performance of *dynamically* quantized feedback systems for either scalar systems [2] or two-dimensional diagonalizable systems [3]. The primary emphasis of this later work was to characterize the optimal performance achieved by the dynamic quantization policies of Brocekt/Liberzon. This paper extends the earlier work in [2] and [3] to n-dimensional diagonalizable systems with bounded process noise.

II. Problem Statement

Consider the quantized feedback system shown in figure 1. The discrete-time system has a state $x[k] \in \Re^n$ that satisfies

The authors are with the department of Electrical Engineering, Univ. of Notre Dame, Notre Dame, IN 46556; e-mail: lemmon,rsun@nd.edu. The authors gratefully acknowledge the partial financial support of the National Science Foundation NSF-ECS0400479

the state equations,

$$x[k+1] = Ax[k] + Bu[k] + w[k]$$

for $k = 0, \ldots, \infty$. $w[k] \in \Re^n$ is an exogenous disturbance such that $\|w[k]\|_\infty \leq M$ where M is a finite constant. $u[k] \in \Re^m$ is the feedback control signal which is generated by a state feedback control law of the form

$$u[k] = Fx^q[k]$$

where $x^q[k]$ is a *quantized* approximation of the state at time k (to be described below). We assume that (A, B) is controllable, A is diagonalizable with unstable eigenvalues ($|\lambda_i| > 1$ for $i = 1, \ldots, n$), and F is a stabilizing state feedback gain matrix of appropriate dimensions.

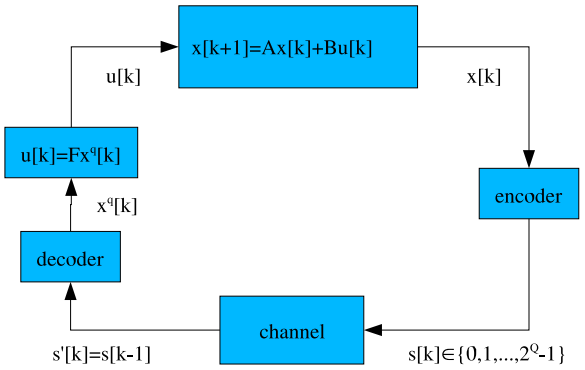

Fig. 1. Quantized Feedback Control System

Figure 1 shows how the system state, $x[k]$, is quantized and transmitted over the feedback channel. In this figure, the system state $x[k]$ is measured at time k by the *encoder* and that measurement is mapped onto a symbol $s[k]$ that is drawn from the discrete set $\{0, 1, \cdots, 2^Q - 1\}$. The symbol $s[k]$ is therefore represented by Q bits. This symbol is transmitted across a lossless communication channel with a single step delay. The *decoder* receives a symbol $s'[k] = s[k-1]$ that is a one-step delayed version of the transmitted symbol. The decoder then uses $s'[k]$ to construct the quantized approximation, $x^q[k]$, of the system state.

The quantization method used to construct $x^q[k]$ from the received symbol originates in the uncertainty set evolution method introduced in [1]. This approach presumes at the beginning of the kth time interval, the encoder and decoder agree that the state lies within the set

$$x[k] \in x^q[k] + U[k].$$

$U[k]$ is a rectangular set of the form

$$U[k] = \prod_{i=1}^{n} [-L_i[k], L_i[k]]$$

where $x^q[k]$ is the quantized state and $L_i[k]$ is the half length of the ith side of the rectangle $U[k]$ at time k. We sometimes refer to $U[k]$ as the *uncertainty set*. The *quantization error* between the true state and the quantized state is denoted as $e[k] = x[k] - x^q[k]$.

Immediately after the start of the kth time interval, the encoder measures the system's current state $x[k]$. The encoder then uses this measurement to determine that

$$x[k] \in x^q_{s[k]} + U_{s[k]}$$

where $x^q_{s[k]}$ is the center of a smaller subset and

$$U_{s[k]} = \prod_{i=1}^{n} \left[-\frac{L_i[k]}{2^{b_i[k]}}, \frac{L_i[k]}{2^{b_i[k]}} \right].$$

$b_i[k]$ represents the number of bits used to quantize the ith component of the state vector at time k. The new center and smaller uncertainty set are indexed by the symbol $s[k]$ which is drawn from the set $\{1, 2, \cdots, 2^Q - 1\}$. This symbol is then transmitted across the channel with a one step delay.

The decoder receives the symbol $s[k]$ at time $k+1$. As soon as it receives this symbol it knows that the system state at time k lies in the set $x^q_{s[k]} + U_{s[k]}$. However, time has now marched ahead from k to $k+1$, so the decoder must propagate the uncertainty set through the state dynamics to determine the quantized state at time $k+1$. This is done through the following equations,

$$\begin{aligned} x[k+1] &\in x^q[k+1] + U[k+1] \\ U[k+1] &= \prod_{i=1}^{n} [-L_i[k+1], L_i[k+1]] \end{aligned}$$

where

$$x^q[k+1] = Ax^q_{s[k]} + BFx^q_{s[k]}$$

and

$$L_i[k+1] = \frac{\lambda_i}{2^{b_i[k]}} L_i[k] + M \quad (i = 1, \ldots, n). \quad (2)$$

Throughout this paper we impose a *constant bit rate* constraint which requires $b_i[k] \geq 0$ and

$$\sum_{i=1}^{n} b_i[k] = Q \quad (3)$$

where Q is a fixed positive integer representing the number of bits used to encode the state.

This paper is interested in determining bit assignments that are *optimal* in the sense of minimizing the worst-case summed quantization error over a finite horizon of length N. In other words we're interested in determining $\{b_i[k]\}$ for fixed Q that minimize

$$P = \sup_{x[0]} \sum_{k=1}^{N} \sum_{i=1}^{n} (x[k] - x^q[k])^2$$

Note that by definition $|x[k] - x^q[k]| \leq L_i[k]$ for all i and any $x[0]$. This inequality becomes equality for a specific initial condition which means that P may be rewritten as

$$P = \sum_{k=1}^{N} \sum_{i=1}^{n} L_i^2[k]$$

For a given number of quantization bits, Q, the objective is to find $\{b_i[k]\}$ for $k = 1, \ldots, N-1$ and $i = 1, \ldots, n$ that minimize P subject to the constraint that $b_i[k] \geq 0$,

$\sum_{i=1}^{n} b_i[k] = Q$ and that $L_i[k]$ satisfies equation 2. This optimization problem may be formally stated as follows,

$$\begin{array}{ll} \text{minimize:} & \sum_{k=1}^{N}\sum_{i=1}^{n} L_i^2[k] \\ \text{with respect to:} & b_i[k] \in \Re \\ \text{subject to:} & b_i[k] \geq 0 \\ & Q = \sum_{i=1}^{n} b_i[k] \\ & L_i[k+1] = \frac{\lambda_i}{2^{b_i[k]}} L_i[k] + M \end{array} \tag{4}$$

This problem will be solved below using dynamic programming [14]. Note that the optimization is done with respect to $b_i[k]$ over the set of real, rather than the set of non-negative integers. The above problem may therefore be seen as a relaxation of the true bit assignment problem which would require the bit assignments to be integers.

III. Optimal Bit Assginments

This section states and proves the optimal bit assignment solving the problem in equation 4. The following mathematical notation will be used throughout this section.

$$\begin{aligned} \Lambda_i[k] &= \lambda_i L_i[k] \\ \overline{\Lambda}[k] &= \sqrt[n]{\prod_{i=1}^{n} \Lambda_i[k]} \\ \overline{\lambda} &= \sqrt[n]{\prod_{i=1}^{n} \lambda_i} \end{aligned}$$

The paper's main proposition is stated below.

Proposition 3.1: If

$$Q > \sum_{i=1}^{n} \log_2(\lambda_i) \tag{5}$$

$$\frac{Q}{n} = \frac{\overline{\Lambda}[0]}{\Lambda_i[0]} \tag{6}$$

for $i = 1, \ldots, n$, then the bit assignment

$$b_i[k] = \frac{Q}{n} - \log_2\left(\frac{\overline{\Lambda}[k]}{\Lambda_i[k]}\right) \tag{7}$$

is a local minimizer of the bit assignment problem in equation 4.

Remark: Note that the first constraint is the stabilizing bit rate constraint in equation 1 for systems in which $|\lambda_i| > 1$ for all i. The second constraint requires that the initial uncertainty set be *balanced*. Note that the bit assignment in equation 7 has an intuitive interpretation which requires us to equally distribute the Q available bits between all channels and then to adjust that "average" bit assignment to the ith side based on the balance between that side's uncertainty set and the geometric average of all sides of the uncertainty set. This means that optimal bit assignments seek to balance the uncertainty almost all components of the state vector, a principle that lies at the heart of the optimal bit assignment described in [3].

Proposition 3.1 will be established using dynamic programming. That method requires the solution of an intermediate single step optimization problem. The solution of that intermediate problem is given in the following lemma.

Lemma 3.2: Let $Q > 0$, $R > 0$, $M > 0$, $\Lambda_i > 0$ and $L_i > 0$ be known real constants for $i = 1, \ldots, n$. Consider the problem of minimizing

$$J = R + \sum_{i=1}^{n} L_i^2 + \sum_{i=1}^{n} \left(M + \frac{\Lambda_i}{2^{b_i}}\right)^2 \tag{8}$$

subject to $b_i \geq 0$ and $\sum_{i=1}^{n} b_i = Q$. There exists a nonempty set $\mathcal{M} \subset \{1, \ldots, n\}$ with cardinality m such that

$$b_i = \begin{cases} \frac{Q}{m} - \log_2\left(\frac{\overline{\Lambda}}{\Lambda_i}\right) & \text{if } i \in \mathcal{M} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

is a local minimizer of J.

Proof: This result is proven using the Karush-Kuhn-Tucker (KKT) conditions [15]. The augmented Lagrangian for this problem is

$$\begin{aligned} L(\mathbf{b}, \mathbf{p}) &= R + \sum_{i=1}^{n} L_i^2 + \sum_{i=1}^{n} \left(M + \frac{\Lambda_i}{2^{b_i}}\right)^2 \\ &\quad + p_0 \left(\sum_{i=1}^{n} b_i[k] - Q\right) + \sum_{i=1}^{n} p_i b_i \end{aligned}$$

where $\mathbf{b} = \{b_1, \cdots, b_n\}$ is the decision variable and $\mathbf{p} = \{p_0, \cdots, p_n\}$ are Lagrange multipliers.

The first KKT condition is that

$$0 = \frac{\partial L}{\partial b_i} = 2\ln 2 \frac{\Lambda_i^2}{2^{2b_i}} + 2M \ln 2 \frac{\Lambda_i}{2^{b_i}} - p_0 - p_i$$

for $i = 1, \ldots, n$. The second KKT condition is the constant bit rate constraint,

$$0 = \frac{\partial L}{\partial p_0} = \sum_{i=1}^{n} b_i - Q$$

The third KKT condition for the inequality constraints requires that $p_i b_i = 0$ for $i = 1, \ldots, n$, where $p_i \geq 0$ and $b_i \geq 0$.

¿From the third KKT condition, we know that if $b_i > 0$, then $p_i = 0$. So let $\mathcal{M}$ denote the set of indices for which $b_i > 0$. We know this set is nonempty because $Q > 0$. For $i \in \mathcal{M}$ the first and third KKT conditions can be used to infer that

$$p_0 = 2\ln 2 \frac{\Lambda_i^2}{2^{2b_i}} + 2M \ln 2 \frac{\Lambda_i}{2^{b_i}} \tag{10}$$

for $i = 1, \ldots, n$.

We may solve equation 10 for b_i as follows. First let $X_i = \frac{\Lambda_i}{2^{b_i}}$ and $\overline{p}_0 = p_0/(2\ln 2)$. With this variable substution equation 10 becomes

$$0 = X_i^2 + M X_i - \overline{p}_0$$

which has a non-negative solution

$$X_i = -M/2 + \sqrt{M^2/4 + \overline{p}_0} \tag{11}$$

which means X_i is independent of the index i.

Note that

$$\overline{X} = \sqrt[n]{\prod_{i=1}^{n} X_i} = \frac{\overline{\Lambda}}{2^{Q/n}} = -\frac{M}{2} + \sqrt{\frac{M^2}{4} - \overline{p}_0} \tag{12}$$

The denominator in equation 12 follows from the constant bit rate constraint constraint (eq. 3) and the righthand side follows from equation 11. We can therefore solve for $\overline{p}_0$ to see that

$$\begin{aligned}\overline{p}_0 &= \left(\frac{\overline{\Lambda}}{2^{Q/n}}+\frac{M}{2}\right)^2-\frac{M^2}{4}\\ &= \frac{\overline{\Lambda}^2}{2^{2Q/n}}+\frac{M\overline{\Lambda}}{2^{Q/n}}\end{aligned} \tag{13}$$

Inserting our expression for $\overline{p}_0$ into equation 11 yields,

$$\begin{aligned}X_i &= \frac{\Lambda_i}{2^{b_i}}\\ &= -\frac{M}{2}+\sqrt{\frac{M^2}{4}+\frac{\overline{\Lambda}^2}{2^{2Q/n}}+\frac{M\overline{\Lambda}}{2^{Q/n}}}\\ &= -\frac{M}{2}+\left(\frac{\overline{\Lambda}}{2^{Q/n}}+\frac{M}{2}\right)\\ &= \frac{\overline{\Lambda}}{2^{Q/n}}\end{aligned}$$

which can be solved for b_i to obtain equation 9 for those $i\in\mathcal{M}$. $\diamondsuit$

Proof of Proposition 3.1: This proof uses dynamic programming [14]. In particular, let's consider a value function with the following form,

$$V(\mathbf{L}[k])=R[k]+\sum_{i=1}^{n}L_i^2[k] \tag{14}$$

where $\mathbf{L}[k]=\{L_i[k]\}_{i=1}^n$ is the "system state" and $R[k]$ is a non-negative sequence of parameters ($k=1,\ldots,N$) in which $R[N]=0$. For notational convenience we let $V_k=V(\mathbf{L}[k])$. Solving this problem involves finding $\{R[k]\}$ and the bit assignemnts $\{\mathbf{b}[k]\}$ that satisfy the dynamic programming recursion,

$$V_k=\min_{\mathbf{b}[k]}\left(\sum_{i=1}^{n}L_i^2[k]+V_{k+1}\right)$$

for $k=0,\ldots,N-1$. The dynamic programming recursion consists of a sequence of single-step optimization problems. The cost functional that must be optimized over this single step can be rewritten as

$$\begin{aligned}&\sum_{i=1}^{n}L_i^2[k]+V_{k+1}\\ =\;&\sum_{i=1}^{n}L_i^2[k]+R[k+1]+\sum_{i=1}^{n}L_i^2[k+1]\\ =\;&R[k+1]+\sum_{i=1}^{n}L_i^2[k]+\sum_{i=1}^{n}\left(\frac{\Lambda_i[k]}{2^{b_i[k]}}+M\right)^2\end{aligned}$$

This cost must be minimized subject to the $b_i[k]\geq 0$ and $\sum_{i=1}^n b_i[k]=Q$. Note that this is precisely the problem we solved in lemma 3.2. So we know that if the ith state component is assigned a nonzero number of bits that the bit assignment must be

$$b_i[k]=\frac{Q}{m[k]}-\log_2\frac{\overline{\Lambda}[k]}{\Lambda_i[k]} \tag{15}$$

where $m[k]$ is the number of state components with nonzero bit assignments at time k. Note that if

$$\frac{Q}{n}>\log_2\frac{\overline{\Lambda}[k]}{\Lambda_i[k]} \tag{16}$$

for all i and k , then all state components can have a nonzero state assignment and equation 15 holds for all i with $m[k]=n$.

We now show that the proposition's assumptions in equations 5 and 6 imply that equation 16 is true for all k. We'll prove this by mathematical induction. First note that equation 6 requires that equation 16 holds for $k=0$. So let's now assume that the inequality holds for k and try to show it also holds for $k+1$.

Note that

$$\begin{aligned}\frac{\overline{\Lambda}[k+1]}{\Lambda_i[k+1]} &= \frac{\sqrt[n]{\prod_{i=1}^n L_j[k+1]\lambda_j}}{L_i[k+1]\lambda_i}\\ &= \frac{\overline{\lambda}}{\lambda_i}\frac{\sqrt[n]{\prod_{j=1}^n\left(\Lambda_j 2^{-b_j[k]}+M\right)}}{\overline{\Lambda}_i[k]2^{-b_i[k]}+M}\end{aligned}$$

Substituting our expression for the optimal $b_i[k]$ into the above equation shows that

$$\frac{\overline{\Lambda}[k+1]}{\Lambda_i[k+1]}=\frac{\overline{\lambda}}{\lambda_i} \tag{17}$$

By the assumption in equation 5, however, we know that

$$\begin{aligned}\frac{Q}{n} &> \frac{1}{n}\left(\sum_{j=1}^{n}\log_2\lambda_j\right)=\log_2\overline{\lambda}\\ &> \log_2\frac{\overline{\lambda}}{\lambda_i}\end{aligned}$$

Under the optimal bit assignment we can insert equation 17 into the above expression to conclude that

$$\frac{Q}{n}>\frac{\overline{\Lambda}[k+1]}{\Lambda_i[k+1]}$$

for any i. We've just shown that if inequality 16 holds for k, then it must also hold for $k+1$. So by the principle of mathematical induction, we know that inequality 16 holds for all k, which implies that under the optimal bit assignment all state components get assigned a nonzero number of bits for all $k\geq 0$.

To complete this proof we need to verify that we can construct the sequence $R[k]\geq 0$ such that the value function in equation 14 satisfies the dynamic programming recursion for all k. Inserting this canadidate form of the value function into the dynamic programming recursion yields,

$$\begin{aligned}V_k &= R[k]+\sum_{i=1}^{n}L_i^2[k]\\ &= R[k+1]+\sum_{i=1}^{n}L_i^2[k]+\sum_{i=1}^{n}\left(M+\frac{\Lambda_i[k]}{2^{b_i^*[k]}}\right)^2\end{aligned} \tag{18}$$

where $b_i^*[k]$ is the optimal bit assignment,

$$b_i^*[k]=\frac{Q}{n}-\log_2\frac{\overline{\Lambda}[k]}{\Lambda_i[k]}. \tag{19}$$

Inserting our expression for b_i^* (equation 19) into the expression for V_k (equation 18) yields,

$$\begin{aligned} & R[k] + \sum_{i=1}^{n} L_i^2[k] \\ = \quad & R[k+1] + \sum_{i=1}^{n} \left(M + \frac{\overline{\Lambda}[k]}{2^{Q/n}} \right)^2 + \sum_{i=1}^{n} L_i^2[k] \end{aligned}$$

So the sequence $R[k]$ is generated by the following recursion,

$$\begin{aligned} R[N] &= 0 \\ R[k] &= R[k+1] + n \left(M + \frac{\overline{\Lambda}[k]}{2^{Q/n}} \right)^2 \end{aligned} \tag{20}$$

for $k = 0, \cdots, N-1$.

To solve the recursion in equation 20, we need to know $\overline{\Lambda}[k]$. Note, however, that if we let $\rho = \overline{\lambda}/2^{Q/n}$ then

$$\begin{aligned} \overline{\Lambda}[1] &= \rho \overline{\Lambda}[0] + \overline{\lambda} M \\ \overline{\Lambda}[2] &= \rho^2 \overline{\Lambda}[0] + \rho \overline{\lambda} M + \overline{\lambda} M \\ & \cdots \end{aligned} \tag{21}$$

$$\overline{\Lambda}[k] = \rho^k \overline{\Lambda}[0] + \overline{\lambda} M \frac{1-\rho^k}{1-\rho} \tag{22}$$

This means that $\overline{\Lambda}[k]$ can be computed as a function of the initial $\overline{\Lambda}[0]$ and we can therefore use equation 22 to compute the required sequence of $R[k]$. $\diamondsuit$

IV. Performance-Rate Functions

One interesting aspect of of the proof in proposition 3.1 is that the value function V_0 gives the optimal cost achieved by the bit assignment. This is generally a function of the bit rate, Q, the horizon length, N, and the noise level, M. We may therefore think of $V_0(Q, M, N)$ as a **performance-rate** function that characterizes the optimal achievable performance (as measured by the summed squared quantization error) as a function of the bit rate. Such performance-rate functions represent fundamental limitations on the system's achievable performance. They are similar in spirit to the Rate-Distortion functions found in source coding theory [16]. This section presents several closed form expressions for the performance-rate function of a dynamically quantized linear system under various assumptions.

The following result establishes a general closed form expression for the optimal performance achievable under the assumption that the system is noise-free.

Corollary 4.1: Under the assumptions of proposition 3.1 and assuming that $M = 0$, the optimal cost achieved under the optimal bit assignment is

$$P^* = \sum_{i=1}^{n} L_i^2[0] + n \frac{\overline{\Lambda}^2[0]}{2^{2Q/n}} \frac{\rho^{2N} - 1}{\rho^2 - 1}$$

where $\rho = \frac{\overline{\lambda}}{2^{Q/n}}$.

Proof: The optimal cost is $\sum_{i=1}^{n} L_i^2[0] + R[0]$. So we need to determine $R[0]$. From equation 20 with $M = 0$ we know that

$$R[k] = R[k+1] + n \frac{\overline{\Lambda}^2[k]}{2^{2Q/n}}$$

From equation 22 with $M = 0$ we know that

$$\overline{\Lambda}[k] = \rho^k \overline{\Lambda}[0]$$

Combining both expressions above yields,

$$R[k] = R[k+1] + n \rho^{2k} \frac{\overline{\Lambda}^2[0]}{2^{2Q/n}}$$

for $k = 0, \ldots, N-1$. We know $R[N] = 0$ so that

$$\begin{aligned} R[N-1] &= n \rho^{2(N-1)} \frac{\overline{\Lambda}^2[0]}{2^{2Q/n}} \\ R[N-2] &= n \rho^{2(N-1)} \frac{\overline{\Lambda}^2[0]}{2^{2Q/n}} + n \rho^{2(N-2)} \frac{\overline{\Lambda}^2[0]}{2^{2Q/n}} \\ &= \frac{n \overline{\Lambda}^2[0]}{2^{2Q/n}} \left(\rho^{2(N-1)} + \rho^{2(N-2)} \right) \\ & \cdots \\ R[0] &= \frac{n \overline{\Lambda}^2[0]}{2^{2Q/n}} \sum_{i=0}^{N-1} \rho^{2i} \\ &= \frac{n \overline{\Lambda}^2[0]}{2^{2Q/n}} \frac{1-\rho^{2N}}{1-\rho^2} \end{aligned}$$

which completes the proof. $\diamondsuit$

The following corollary states a closed form expression for the summed squared quantization error in the noisy case ($M \neq 0$).

Corollary 4.2: Under the assumptions of proposition 3.1, the optimal cost achieved by the optimmal bit assignment is

$$V_0(Q, N) = \sum_{i=1}^{n} L_i^2[0] + A + B \frac{1-\rho^N}{1-\rho} + C \frac{1-\rho^{2N}}{1-\rho}$$

where

$$\begin{aligned} A &= nNM^2 + 2nNMY + nY^2 \\ B &= n(2M+Y)(X-Y) \\ C &= n(X-Y)^2 \end{aligned}$$

and $X = \frac{\rho \overline{\Lambda}[0]}{\overline{\lambda}}$, $Y = \frac{\rho M}{1-\rho}$, and $\rho = \frac{\overline{\lambda}}{2^{Q/n}}$.

Proof: From equation 20, we can readily show that

$$\begin{aligned} R[0] &= n \left(\sum_{k=0}^{N-1} \left(M + \frac{\rho \overline{\Lambda}[k]}{\overline{\lambda}} \right)^2 \right) \\ &= nNM^2 + 2Mn \sum_{k=0}^{N-1} \frac{\rho \overline{\Lambda}[k]}{\overline{\lambda}} + n \sum_{k=1}^{N-1} \left(\frac{\rho \overline{\Lambda}[k]}{\overline{\lambda}} \right)^2 \end{aligned} \tag{23}$$

Using equation 22, the first summation above can be written (after some algebra) as

$$\begin{aligned} \sum_{k=0}^{N-1} \frac{\rho \overline{\Lambda}[k]}{\overline{\lambda}} &= \left(\frac{\rho}{\overline{\Lambda}[0]} \overline{\lambda} - \frac{\rho M}{1-\rho} \right) \sum_{k=0}^{N-1} \rho^k + \frac{\rho M}{1-\rho} N \\ &= (X-Y) \sum_{k=0}^{N-1} \rho^k + YN \end{aligned} \tag{24}$$

and the second summation may be written as

$$\sum_{k=0}^{N-1} \left(\frac{\rho \overline{\Lambda}[k]}{\overline{\lambda}} \right)^2 = (X-Y)^2 \sum_{k=0}^{N-1} \rho^{2k} \tag{25}$$

$$+ Y(X-Y) \sum_{k=1}^{N-1} \rho^k + Y^2 \tag{26}$$

Inserting equations 24-26 into equation 23 and simplifying yields the expression in the corollary's statement. $\diamondsuit$

In the limit as $N \to \infty$, we can expect the summed squared quantization error to become unbounded if $M \neq 0$. In this case it makes more sense to use the following performance measure,

$$P_\infty = \lim_{N\to\infty} \frac{1}{N} V_0(Q, N)$$

The following corollary provides a closed form expression for P_∞.

Corollary 4.3: Under the assumptions of proposition 3.1 then under optimal bit assignment

$$P_\infty = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \sum_{i=1}^{n} L_i^2[k] = nM^2 \frac{1+\rho}{1-\rho}$$

Proof: This is obtained by simply letting N go to infinity in the result from corollary 4.2. $\diamondsuit$

V. Bit Assignment Algorithm Performance

The performance rate curves shown in the preceding section represent the minimum summed quantization error assuming bit assignments are real-valued. In reality, these bit assignments are integer valued and the natural question to ask is how far a "real" algorithm making integer assignments may deviate from the theoretical performance-rate curve.

To answer this question, we developed a simple recursive algorithm to compute the optimal bit assignments and compared the resulting quantization error against the theoretical performance-rate function. The following algorithm was used to make the integral assignments. In this algorithm n is the dimension of the system, L_i is the uncertainty on the i component of the state, λ_i is the eigenvalue associated with the ith subsystem, and Q are the number of bits that need to be assigned.

001 intialize: $i = n$

002 $\Lambda_j = L_j \lambda_j \quad (j = 1, \ldots, n)$

003 $\overline{\Lambda} = \prod_{j=1}^{n} (L_j \lambda_j)^{1/n}$

004 $Q_m = Q$

005 LOOP: do while $i > 1$

006 $k = \arg\min_j \left(\frac{\overline{\Lambda}}{\Lambda_j} \right)$

007 if $i == 1$, then $b_k = Q_m$

008 else $b_k = \mathrm{F}\left(\frac{Q_m}{i} - \log_2 \left(\frac{\overline{\Lambda}}{\Lambda_k} \right) \right)$

009 $Q_m = Q_m - b_k$

010 $L_k = 0.0$

011 $i = i - 1$

The actual bits assigned to the kth state component is given above in variable b_k. The integer assignment is done in step 008, where the function $F(\cdot)$ may be taken as either a rounding, ceiling, or floor function.

Figure 2 compares the results obtained using this algorithm against the theoretical bound. In this case we assumed $M = 0$, so a log-log plot of the performance-rate function is simply a straight line. The blue dots in the figure represent the performance levels that were achieved by an algorithm with integer bit assignments. This figure shows results for integer assignments made by both the ceiling function and floor function. The figure shows that the integer assignment algorithm performs very close to the theoretical performance-rate function.

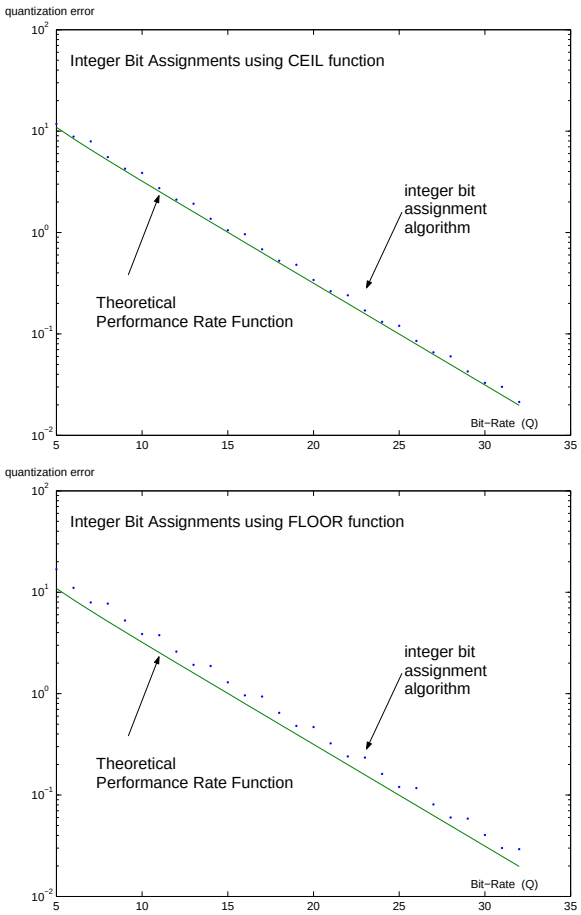

Fig. 2. Integer versus Real Bit Assignments: noise free case

Similar results were generated for the bounded noise case when $M = 1$. Figure 3 plots the power of the quantization error signal P_∞ as a function of Q for the infinite horizon case ($N = \infty$). The results show close agreement between the integer bit assignment and the predictions made by the performance-rate function.

VI. Summary

This paper did three things. It first used dynamic programming to determine an optimal way of assigning bits in a dynamically quantized feedback system, where optimality refers to minimizing the summed squared quantization error. The proof of optimality used a dynamic programming argument that allowed us to obtain closed form expressions for the optimal achievable performance as a function of the bit-rate. These expressions can be thought of as rate-distortion functions for feedback control loops. This paper referred to these functions as "performance-rate" functions.

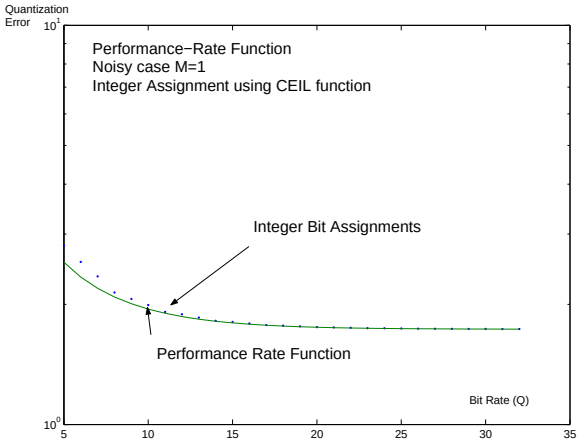

Fig. 3. Integer veruss Real Bit Assignments: noisy case

Since theoretical performance-rate functions assume real-valued bit assignments, we experimentally compared of the performance achieved using integer-valued bit assignments to the performance predicted by the performance-rate function. The resulting comparisons were very close to each other.

Prior results on the optimal performance of dynamically quantized feedback systems were obtained in [3] for noise-free 2-dimensional systems. This paper extends that prior work to multi-dimensional diagonalizable systems with bounded noise. The analysis method used in this paper may be extended to non-diagonalizable system by simply using a larger noise term M.

REFERENCES

[1] R. Brockett and D. Liberzon, "Quantized feedback stabilization of linear systems," *IEEE Transactions on Automatic Control*, vol. 45(7), pp. 1279 –1289, 2000.

[2] M. Lemmon and Q. Ling, "Control system performance under dynamic quantization: the scalar case," in *IEEE Conference on Decision and Control*, Atlantis, Paradise Island, Bahamas, 2004.

[3] Q. Ling and M. Lemmon, "Optimal dynamic bit assignment in second-order noise-free quantized linear control systems," in *IEEE Conference on Decision and Control*, Seville, Spain, 2005.

[4] D. Delchamps, "Stabilizing a linear system with quantized state feedback," *IEEE Transactions on Automatic Control*, vol. 35(8), pp. 916–924, 1990.

[5] N. Elia and S. Mitter, "Stabilization of linear systems with limited information," *IEEE Transactions on Automatic Control*, vol. 46(9), pp. 1384 –1400, 2001.

[6] W. Wong and R. Brockett, "Systems with finite communication bandwidth constraints- part ii: stabilization with limited information feedback," *IEEE Transactions on Automatic Control*, vol. 44(5), pp. 1049–1053, 1999.

[7] J. Baillieul, "Feedback designs in information-based control," in *Stochastic Theory and Control, Lecture Notes in Control and Information Sciences, B. Pasik-Duncan (ed.),Springer-Veralg LNCIS 280*, 2002, pp. 35–37.

[8] F. Fagnani and S. Zampieri, "Stability analysis and synthesis for scalar linear systems with a quantized feedback," *IEEE Transactions on Automatic Control*, vol. 48(9), pp. 1569–1584, 2003.

[9] D. Liberzon, "On stabilization of linear systems with limited information," *IEEE Transactions on Automatic Control*, vol. 48(2), pp. 304–307, 2003.

[10] S. Tatikonda, "Control under communication constraints," Ph.D. dissertation, M.I.T., 2000.

[11] S. Tatikonda and S. Mitter, "Control under communication constraints," *IEEE Transactions on Automatic Control*, vol. 49(7), pp. 1056–1068, 2004.

[12] G. Nair and R. Evans, "Exponential stabilisability of finite-dimensional linear systems with limited data rates," *Automatica*, vol. 39, pp. 585–593, 2003.

[13] K. Astrom and B. Wittenmark, *Computer-Controlled Systems: theory and design*, 2nd ed. Prentice-Hall, 1990.

[14] D. Bertsekas, *Dynamic Programming and Optimal Control, volume 1*, 2nd ed. Athena Press, 2000.

[15] M. Bazaraa, H. Sherali, and C. Shetty, *Nonlinear Programming: theory and algorithms*, 2nd ed. John Wiley and Sons, 1993.

[16] R. Gray, *Source Coding Theory*. Kluwer Academic Press, 1990.