Submitted, 2006 American Control Conference
http://www.cds.caltech.edu/~murray/papers/2005q_gdmh06-acc.html

On the Effect of Quantization on Performance

Vijay Gupta, Amir F. Dana, Richard M Murray and Babak Hassibi

***Abstract*— We study the effect of quantization on the performance of a scalar dynamical system. We provide an expression for calculation of the LQR cost of a dynamical system for a general quantizer. Using the high-rate approximation, we evaluate it for two commonly used quantizers: uniform and logarithmic. We also provide a lower bound on performance of the optimal quantizer based on entropy arguments and consider the case when the channel drops data packets stochastically.**

I. INTRODUCTION

In recent years analysis of linear systems in which the process and the controller communicate using a digital communication channel has gained increasing attention. This situation can be expected to model various situations in which the controller and the process are physically separated and communication between the two is not assumed to be over an infinite bandwidth channel, such as in networked or distributed control applications. Moreover, this area lies at the intersection of control theory and communications and as such can provide useful insights into both the areas. For example, from the view-point of communication theory, this problem, and the more general question of control under information constraints, can provide a handle on the question of usage based measure of information raised by Witsenhausen [24] and others. From a control-theoretic viewpoint, this problem attacks basic questions such as the value of information in achieving performance objectives [17]. In short, the interaction of control and communication is a very important problem to be tackled for gaining a deeper understanding of both the fields.

As a result, this problem has been analyzed with increasing regularity since the seminal paper [6]. The problem of stabilization with finite communication bandwidth was considered by Wong and Brockett [25], [26]. Baillieul [1] also reported a tight bound on the data rate requirement for stabilizing a scalar system. Nair et al [19], [18] considered the stabilization of stochastic linear systems and Markov jump linear systems with finite data rates. Tatikonda [22] studied stabilization of finite-dimensional discrete-time noiseless linear processes and also presented results about the optimal LQG control of linear systems across noisy feedback links (see also [2]). Elia and Mitter [7] considered the question of optimal quantizer for stabilization. Various quantization and coding schemes for stabilization have also been studied in the literature, (see, e.g., [20], [3], [16], [11], [12], [8]).

However, most of the work reported so far has been focused on the effect of quantization and coding on stability. It is worthwhile to also consider the question of performance of the system in the presence of quantization. The chief works in this direction seem to be [23], [15].The performance of scalar statically quantized system with delays was considered in [23]. Lemmon and Ling [15] presented an upper bound for the quantization noise for the case when dynamic uniform quantization is done over a channel that drops packets. They defined the performance in terms of signal to quantization ratio and presented some interesting trade-offs between the number of bits, locations of the system poles and the performance.

In this paper we study the effect of quantization on the LQR performance of the system. We consider a linear time-invariant scalar system with a control law in place and see how the performance degrades as less and less data is allowed to pass from the process to the controller. We come up with some interesting bounds for specific quantizers and some entropy-based general bounds on any general quantization and encoding scheme. We also consider extensions to dynamic quantizations schemes and packet-dropping channels.

The paper is organized as follows. In the next section, we set up the problem and define some notation. We study some specific static quantizers and give a bound on the performance achievable by any scheme. Then we consider dynamic quantizers and the extension to packet-dropping channels. We end with conclusions and outline some scope for future work.

II. PROBLEM FORMULATION

Consider the situation described in figure 1. We have a linear time-invariant process evolving in discrete time according to the relation

$$x_{k+1} = ax_k + u_k + w_k \tag{1}$$

with x_k as the process state, u_k as the control input and w_k as zero-mean white noise. In this paper, we will only consider the case of a scalar process, thus $x_k \in \mathbf{R}$. The noise w_k is assumed to be bounded in an interval $[-M, M]$ with variance σ^2. The initial condition x_0 is assumed to be uniformly distributed in a symmetric interval $[-u, u]$ and is assumed to be independent of w_k. We also assume that w_k has a symmetric distribution about the origin.

The process state is observed in a noiseless fashion by the encoder. The encoder denotes a mapping from the state x_k to a stream of bits b_k. The encoder has access to all the previous states $\{x_j\}_{j=0}^{k}$ and the previous control signals $\{u_j\}_{j=0}^{k-1}$ when it encodes x_k. We restrict our attention to encoders that merely perform the action of quantization and ignore the possibility of other source coding.

Division of Engineering and Applied Science, California Institute of Technology, Pasadena, CA 91125, USA {gupta,amirf,murray,hassibi }@caltech.edu

The bit-stream $\{b_k\}$ is transmitted over a channel. A channel is specified by a set of conditional probabilities between the input and the output. For every bit transmitted by the encoder, the channel outputs another bit to the decoder. In general, these bits can differ if, e.g., the channel introduces noise or bit flips or the bit is erased. We ignore all such possible channel effects. In other words, the channel we consider is a noiseless digital channel.

After passing through the channel, the bits are received at the decoder. The decoder looks at the bit stream and outputs an estimate of the state $\hat{x}_k$. The decoder has access to all the previous bit streams $\{b_j\}_{j=0}^{k}$, the previous decoded estimates $\{\hat{x}_j\}_{j=0}^{k-1}$ and the previous control signals $\{u_j\}_{j=0}^{k-1}$ when it decodes $\{b_k\}$. We assume a linear control law of the form

$$u_k = f\hat{x}_k. \tag{2}$$

This control signal u_k is then used in the further evolution of the process described by (1). We assume that there is no channel present between the controller and the process.

In the absence of any encoder, channel and decoder, we have $\hat{x}_k = x_k$. In general the two quantities would not be equal. For this process we consider the finite-time horizon

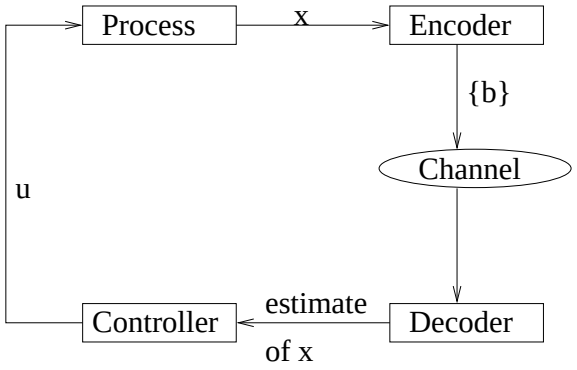

Fig. 1. Representation of the system considered.

LQ cost given by

$$J_K = E\left[\sum_{k=0}^{K} x_k^2 q + u_k^2 r\right], \tag{3}$$

where, as usual, q is positive and r is non-negative. For the infinite-time horizon case, we consider

$$J_\infty = \lim_{K\to\infty} \frac{1}{K} J_K,$$

if the limit exists. We assume that system has been sufficiently well-designed so that the system remains stable and inside the range of operation of the quantizer at all times to avoid quantizer over-flow issues. For this system, we wish to consider the effect of various quantizers on J_K and J_∞.

We will denote the probability density function of a continuous random variable X by $f_X(x)$ and its expectation by $E[X]$. The differential entropy of X is denoted by $h(X)$ and defined according to (see, e.g., [5])

$$h(X) = -\int_{-\infty}^{\infty} f_X(x) \log f_X(x) dx,$$

where $0 \log 0$ is interpreted as 0 and the $\log$ is taken to the base 2. A scalar quantizer Q of size N is a mapping from an interval on the real number line into a finite set $\mathcal{C}$ containing N reproduction points called codewords. The interval is partitioned into N cells with the i-th cell defined by

$$R_i = \{x \in \mathbf{R} : Q(x) = y_i\},$$

where y_i is the i-th codeword. We call a quantizer static or fixed if the mapping does not change with time, otherwise we call it dynamic. The rate of the quantizer is defined as $r = \log_2(N)$. A distortion measure d is an assignment of a non-negative cost $d(x, \hat{x})$ associated with quantizing the vector x with a reproduction vector $\hat{x}$. The average distortion or quantization error is thus defined by $D = E\left[d(X, \hat{X})\right]$ where the expectation is taken over the probability density function of X. In particular, the mean squared distortion is defined as $D = E\left[|X - \hat{X}|^2\right]$.

III. Analysis

We begin by considering the case when the quantizer is fixed. In this paper, we shall assume that $\hat{x}_k$ is the midpoint of the quantization cell corresponding to the bits $\{b_k\}$. We will also assume that the quantizer is symmetric about the origin; because of symmetry of the problem, this is a reasonable assumption. Let $\hat{x}_k = x_k + \Delta_k$, where we have denoted the estimation (or the quantization) error by Δ_k. By symmetry, we have for all k

$$\begin{aligned}
E[\Delta_k \hat{x}_k] &= \int_t E[\Delta_k \hat{x}_k | \Delta_k = t] f_{\Delta_k}(t) dt \\
&= \int_{t>0} \Big(E[\Delta_k \hat{x}_k | \Delta_k = t] f_{\Delta_k}(t) dt \\
&\quad + E[\Delta_k \hat{x}_k | \Delta_k = -t] f_{\Delta_k}(-t) dt \Big) \\
&\overset{(1)}{=} \int_{t>0} f_{\Delta_k}(t) dt \Big(E[\Delta_k \hat{x}_k | \Delta_k = t] \\
&\quad + E[\Delta_k \hat{x}_k | \Delta_k = -t] \Big) \\
&= \sum_{t>0} f_{\Delta_k}(t) dt \times 0 = 0,
\end{aligned}$$

where (1) follows from the fact that $f_{\Delta_k}(t) = f_{\Delta_k}(-t)$ because the quantizer is symmetric. Thus Δ_k and $\hat{x}_k$ are uncorrelated. This allows us to write

$$E[x_{k+1}^2] = (a+f)^2 E[x_k^2] - (f^2 + 2af) E[\Delta_k^2] + \sigma^2.$$

Thus the cost can be evaluated as

$$\begin{aligned}
J_K &= \sum_{k=0}^{K} \left[E\left[x_k^2\right] q + E\left[u_k^2\right] r\right] \\
&= \sum_{k=0}^{K} \left[(q+rf^2)E[x_k^2] - rf^2 E[\Delta_k^2]\right] \\
&= \sum_{k=0}^{K} \Big[(q+rf^2)\big[(a+f)^2 E[x_{k-1}^2] \\
&\quad -(f^2+2af)E[\Delta_{k-1}^2+\sigma^2]\big] - rf^2 E[\Delta_k^2]\Big] \\
&\vdots \\
&= \sum_{k=0}^{K} \Big[-rf^2 E[\Delta_k^2] + (q+rf^2)(a+f)^{2k} E[x_0^2] \\
&\quad +\sigma^2(q+rf^2)\sum_{j=0}^{k-1}(a+f)^{2j} - (f^2+2af)\times \\
&\quad (q+rf^2)\sum_{j=0}^{k-1}(a+f)^{2j} E[\Delta_{k-1-j}^2]\Big].
\end{aligned}$$

A. Evaluation for Specific Quantizers

In general, the cost function is not easy to calculate analytically. This is because the quantization error depends on the probability density function of x_k which is not easy to calculate as time k evolves. To obtain a handle on the performance of different quantizers, we make the high rate approximation, which says that the rate of the quantizer is high (and hence the distortion is low). The results we obtain can thus be treated as approximations which become better as the rate of the code increases. We first consider a uniform quantizer, which is a very simple and commonly used quantizer. A standard result from scalar quantization [9] says that for a uniform quantizer with step size δ, the mean squared distortion when a random variable with uniform probability distribution is quantized is given by $\frac{\delta^2}{12}$. To evaluate the terms $E\left[\Delta_k^2\right]$ that appear in the cost function, we proceed as follows. By assumption, x_0 is uniformly distributed. Thus $E\left[\Delta_0^2\right] = \frac{\delta^2}{12}$. Since x_1 is given by $(a+f)x_0 + f\Delta_0 + w_0$, we see that conditioned on the values of Δ_0 and w_0, x_1 is uniformly distributed. Thus,

$$E[\Delta_1^2 | \Delta_0 = \alpha, w_0 = \beta] = \frac{\delta^2}{12},$$

where α and β are arbitrary values. This in turn yields

$$E[\Delta_1^2] = E[E[\Delta_1^2|\Delta_0, w_0]] = \frac{\delta^2}{12}.$$

By a similar procedure, we obtain that $E[\Delta_k^2] = \frac{\delta^2}{12}, \forall k$. Here we have assumed that there is no quantization cell such that x_k can assume only a part of the values corresponding to it. This is obviously the high-rate approximation.

Thus we can now evaluate the cost function to be

$$\begin{aligned}
J_K = \sum_{k=0}^{K} \big[&-rf^2\frac{\delta^2}{12} + (q+rf^2)(a+f)^{2k}E[x_0^2] \\
&+ (q+rf^2)(\sigma^2 - (f^2+2af))\frac{\delta^2}{12}\sum_{j=0}^{k-1}(a+f)^{2j}\big]
\end{aligned}$$

For calculation of J_∞ we need to find conditions such that J_K/K does not diverge. To this end, we need to assume that $(a+f)^2 < 1$ and that $E[x_{k+1}^2] < E[x_k^2]$. The first condition means f is stabilizing while the second condition places a limit on the size of the quantization cell. The condition implies that

$$(a+f)^2 E[x_k^2] - (f^2+2af)\frac{\delta^2}{12} + \sigma^2 \leq E[x_k^2].$$

Assuming that there are N quantization cells, this yields

$$N^2 \geq \frac{u^2\left(a^2 - (a+f)^2\right)}{3\left((1-(a+f)^2)E[x_k^2] - \sigma^2\right)}.$$

In particular for $k=0$, this condition implies

$$N^2 \geq \frac{u^2\left(a^2 - (a+f)^2\right)}{u^2(1-(a+f)^2) - 3\sigma^2}.$$

Note that for the case when there is no noise and the control law $f = -a$, this reduces to the results derived in [22]. With these assumptions, the infinite-horizon cost can be evaluated to be

$$J_\infty = -rf^2\frac{\delta^2}{12} - \frac{(q+rf^2)}{1-(a+f)^2}\left((f^2+2af)\frac{\delta^2}{12} - \sigma^2\right).$$

We now calculate the cost for a logarithmic quantizer that has been shown to be the most optimal quantizer for stabilization [7]. For studying its performance, we use the following result from source coding theory [10].

Theorem 1: Given a scalar quantizer with a mean square based distortion measure $d(x,y) = ||x-y||^2$, the expected distortion of the random variable X being quantized can be bounded as follows

$$\bar{d} \geq \bar{d}_L = \frac{1}{12N^2} E\left[\lambda(X)^{-2}\right],$$

where $\lambda(X)$ is the asymptotic quantizer density normalized to unit integral, obtained as we keep on increasing the number of quantization levels while N refers to the total number of quantization cells. Further, the lower bound becomes tighter as the rate of the code gets high.

When we apply this theorem to the uniform quantizer, we obtain the cost function calculated above. To apply this result for a logarithmic quantizer that is operating over the union of the regions $[-a,-\epsilon]$ and $[\epsilon,a]$, we note that the asymptotic quantizer density is given by

$$\lambda(x) = \frac{1}{|x|}\Big|_{\text{normalized to unit integral}} = \frac{1}{2|x|\ln(\frac{a}{\epsilon})}.$$

Thus the distortion measure approximately evaluates to

$$\bar{d} = \frac{1}{3}\left(\frac{\ln(\frac{a}{\epsilon})}{N}\right)^2 E\left[x^2\right].$$

Now consider a logarithmic quantizer with ratio g. Thus the quantization cells are given for the positive axis by the intervals $[\epsilon, g\epsilon]$, $[g\epsilon, g^2\epsilon]$, $\cdots$, $[g^{p-1}\epsilon, g^p\epsilon]$, where p and N are related by $2p = N$. Since $g^p\epsilon = a$, we obtain

$$N = \frac{2\ln(\frac{a}{\epsilon})}{\ln(g)}.$$

Thus the distortion measure is given by

$$\bar{d} = \frac{1}{3}\left(\frac{\ln(g)}{2}\right)^2 E\left[x_k^2\right] = \frac{(\ln g)^2}{12} E\left[x_k^2\right] \approx E\left[\Delta_k^2\right].$$

If we define

$$\begin{aligned} r_1 &= q + rf^2 - rf^2\frac{(\ln g)^2}{12} \\ r_2 &= (a+f)^2 - (f^2+2af)\frac{(\ln g)^2}{12}, \end{aligned}$$

we can calculate the cost function to be

$$\begin{aligned} J_K &= \frac{r_1(1-r_2^K)}{1-r_2}E[x_0^2] + \sum_{k=0}^{K}\frac{r_1\sigma^2(1-r_2^{k-1})}{1-r_2} \\ J_\infty &= \frac{r_1\sigma^2}{1-r_2}. \end{aligned}$$

The condition for existence of J_∞ is that $r_2 < 1$.

B. Performance Bounds

In this subsection, we try to lower bound the cost function for the choice of any quantizer and for any variable to be quantized and sent. We use entropy arguments that do not require high rate approximations used above. We note the following [14], [5]

- Given n bits to describe a random variable X with differential entropy $h(X)$, the error can have differential entropy no less than $h(X) - n$.
- Given a random variable X with differential entropy $h(X)$, the lowest possible variance of X is $\frac{1}{2\pi e}2^{2h(X)}$.
- *The Entropy-Power Inequality*: Given two independent random variables X and Y with differential entropy $h(X)$ and $h(Y)$ respectively,
$$2^{2h(X+Y)} \geq 2^{2h(X)} + 2^{2h(Y)}.$$
- Entropy of a random variable X is no less than the entropy of X given additional information about another random variable Y.

At time step $k = 0$, the entropy is simply $h(x_0)$, thus the entropy of Δ_0 is at least $h(x_0) - n$. At time step $k = 1$, we have

$$\begin{aligned} h(x_1) &\geq h\left(x_1|\hat{x}_0\right) \\ &= h\left(ax_0 + f\hat{x}_0 + w_0|\hat{x}_0\right) \\ &= h\left(ax_0 + w_0|\hat{x}_0\right) \end{aligned}$$

.

Now x_0 and w_0 are independent (even given $\hat{x}_0$). Thus the entropy-power inequality yields

$$\begin{aligned} 2^{2h(x_1)} &\geq 2^{2h(ax_0|\hat{x}_0)} + 2^{2h(w|\hat{x}_0)} \\ &= 2^{2\log(a)+2h(x_0|\hat{x}_0)} + 2^{2h(w)} \\ &\geq 2^{2\log(a)}2^{2h(x_0)-2n} + 2^{2h(w)} \\ &= c2^{2h(x_0)} + 2^{2h(w)}, \end{aligned}$$

where $h(w)$ is the entropy of the noise and $c = a^2 2^{-2n}$. Thus we obtain

$$h(x_1) \geq \frac{1}{2}\log\left[c2^{2h(x_0)} + 2^{2h(w)}\right],$$

or that

$$h(\Delta_1) \geq \frac{1}{2}\log\left[c2^{2h(x_0)} + 2^{2h(w)}\right] - n.$$

Similarly we may obtain

$$\begin{aligned} h(x_2) &\geq \frac{1}{2}\log\left[c2^{2h(x_1)} + 2^{2h(w)}\right] \\ &\geq \frac{1}{2}\log\left[c\left(c2^{2h(x_0)} + 2^{2h(w)}\right) + 2^{2h(w)}\right] \\ &= \frac{1}{2}\log\left[c^2 2^{2h(x_0)} + c2^{2h(w)} + 2^{2h(w)}\right]. \end{aligned}$$

This yields

$$h(\Delta_2) \geq \frac{1}{2}\log\left[c^2 2^{2h(x_0)} + c2^{2h(w)} + 2^{2h(w)}\right] - n.$$

In general, we thus have

$$h(x_k) \geq \frac{1}{2}\log\left[c^k 2^{2h(x_0)} + \sum_{j=0}^{k-1} c^j 2^{2h(w)}\right],$$

and thus we can bound the entropy of the error as

$$h(\Delta_k) \geq \frac{1}{2}\log\left[c^k 2^{2h(x_0)} + \sum_{j=0}^{k-1} c^j 2^{2h(w)}\right] - n.$$

Finally the error variance at time step k is bounded by

$$E[\Delta_k^2] \geq \frac{1}{2\pi e}2^{-2n}\left[c^k 2^{2h(x_0)} + \sum_{j=0}^{k-1} c^j 2^{2h(w)}\right].$$

Thus we can go ahead and evaluate the lower bound on cost function as

$$\begin{aligned} &\sum_{k=0}^{K}\Bigg[-rf^2\frac{1}{2\pi e}2^{-2n}\left(c^k 2^{2h(x_0)} + \sum_{j=0}^{k-1} c^j 2^{2h(w)}\right) \\ &+ (q+rf^2)(a+f)^{2k}E[x_0^2] + \sigma^2(q+rf^2)\sum_{j=0}^{k-1}(a+f)^{2j} \\ &- (f^2+2af)(q+rf^2)\frac{1}{2\pi e}2^{-2n}\sum_{j=0}^{k-1}(a+f)^{2j} \\ &\left(2^{2h(x_0)}c^{k-j-1} + \sum_{i=0}^{k-j-2} 2^{2h(w)}c^i\right)\Bigg]. \end{aligned}$$

Further, if we assume

$$(a+f)^2 \leq 1, \qquad a^2 2^{-2n} \leq 1, \tag{4}$$

we obtain

$$J_\infty = -rf^2 \frac{1}{2\pi e} 2^{-2n} 2^{2h(w)} \frac{1}{1-c} + \frac{\sigma^2(q+rf^2)}{1-(a+f)^2}$$
$$-(f^2+2af)(q+rf^2)\frac{1}{2\pi e} 2^{2(h(w)-n)} \frac{1}{1-(a+f)^2}\frac{1}{1-c}.$$

Note that the condition given in (4) is similar to the condition obtained for stability of a scalar unstable system in, e.g., [1]. We do not yet have an analytic expression for the tightness of the bound. Some numerical simulations are presented later in the paper.

Remarks:

- It is known that the optimal quantizer minimizing the mean-square distortion error depends on the probability density function of the variable being quantized. Since the density function of x_k depends on the densities of all previous quantization errors, it is difficult to compute a priori and the optimal quantizer would be obtained at every step through an iterative algorithm such as the Lloyd-Max algorithm [9], [10] or through a dynamic programming based algorithm [4].
- So far we have assumed that the quantization is not followed by any noiseless coding. Moreover we have concentrated on the case of fixed rate quantization. Thus we defined the rate of the quantizer as $\log(N)$, where N is the number of quantization levels. If we assume that noiseless coding is permitted, it makes more sense to consider the entropy of the output vector as a measure of the rate. In such a case, we note the following result [10]

  Theorem 2: The constrained entropy high rate quantizer bound is given by

  $$\bar{d}_L \geq \frac{1}{12} e^{-2\left(H(q(X))-h(X)\right)},$$

  where $h(X)$ is the differential entropy of the random variable X while $H(q(X))$ is the entropy of quantized variable $q(X)$. Furthermore, equality is achieved if and only if the asymptotic quantizer density $\lambda(x)$ is a constant, that is, the quantizer reproduction vectors are uniformly distributed over some set having probability 1. Thus the bound is achieved by high rate lattice vector quantizers since they have a uniform density of quantization levels.

  If we define rate as $R = H(q(X))$ (which gives the average codeword length achievable using noiseless coding and hence the average rate), we obtain the above bound. For a fixed R, it can be proved that this distortion is lower than the one achieved for a given code rate. However, actually achieving this rate might require long codewords and hence might not be practical in a real-time system.
- There also have been some works in the information theory literature, e.g. [21], that provide a way to bound the output entropy of a quantizer given the number of levels of the quantizer. Such works may provide an interesting way to achieve a trade-off between the two notions of rate that we have presented.

C. Dynamic Quantization: Uniform Quantizer

It is apparent that only the region corresponding to the uncertainty that the decoder has about x_k needs to be quantized and the information sent. We now consider this case of dynamic quantization in which the number of quantization levels N remains fixed; however the range over which quantization is being done varies with time. Thus the meaning of the bits $\{b_k\}$ changes as time index k progresses. This is similar to schemes like prediction based encoding outlined in [11] and yields better performance, at the cost of added complexity due to a time-varying quantizer. Moreover it assumes some level of synchronization between the encoder and the decoder so that both agree on the specific quantizer to which the bits at time k pertain.

Denote the length of the interval that is quantized at time step k by l_k. Also suppose that x_k falls in the i-th quantization cell that has length c_i. Then we have

$$l_0 = 2u \qquad\qquad l_{k+1} = ac_i + 2M.$$

We assume that the quantizer used at time k is symmetric with respect to the midpoint of the region being quantized at time step k (this region can be identified from $\hat{x}_{k-1}$, u_{k-1} and M). Thus by symmetry, we would once again have $E[\Delta_k \hat{x}_k] = 0$. The cost function retains the same general form as the one for static quantizer.

For simplicity, we consider only the infinite-time horizon cost function J_∞. For the case of a uniform quantizer with N levels, the quantization step size at time k is given by

$$N\delta_k = l_k = a\delta_{k-1} + 2M.$$

The variance of the quantization error at time k, $E[\Delta_k^2]$ can be evaluated as before to be $\frac{\delta_k^2}{12}$. Thus the cost function evaluates as

$$J_\infty = -rf^2 \frac{M^2}{3(N^2-a^2)} + \frac{\sigma^2(q+rf^2)}{1-(a+f)^2}$$
$$-(q+rf^2)(f^2+2af)\left(\frac{M^2}{3(N^2-a^2)(1-(a+f)^2)}\right).$$

The conditions for existence of J_∞ are

$$(a+f)^2 \leq 1, \qquad\qquad \frac{a}{N} < 1.$$

Also note that this cost is equivalent to that of a static uniform quantizer with step size

$$\delta = \frac{4M^2}{N^2-a^2}.$$

Since the cost function for a static uniform quantizer is an increasing function in the step size δ, this gives us a relation between the parameters M, N and a for determining which of the two quantizers, static or dynamic, is better.

D. Stochastic Packet Drops

So far we had assumed a perfect channel model, in that the bits $\{b_k\}$ were transmitted to the decoder without fail. A more realistic channel model is one that suffers from stochastic data loss. We model the data loss in this paper using the random packet loss model. At each time step, the channel can either be in a 'good' or a 'bad' state. In the good state, it transmits the bits $\{b_k\}$ while in the bad state the packet containing the bits is dropped and no data is transmitted to the decoder. For simplicity we shall consider only the case when the channel transitions between these two states in an i.i.d. fashion although the results may readily be extended to the case when the transition occurs according to a Markov chain (the classical Gilbert-Elliot channel model). Let the probability of packet drop at every time step be p. Note that the expectation in the cost function is now also over the probability of packet drops at each time step. For simplicity, we consider only J_∞.

- Uniform Quantizer: The state evolves according to the equation

$$x_{k+1} = \begin{cases} (a+f)x_k + f\Delta_k + w_k & \text{with prob } 1-p \\ ax_k + w_k & \text{with prob } p. \end{cases} \tag{5}$$

For a uniform quantizer, $E[\Delta_k^2] = \frac{\delta^2}{12}$ while Δ_k and w_k are independent of each other. Thus we can evaluate the steady-state covariance as

$$\begin{aligned} P_\infty &= (1-p)\Big((a+f)^2 P_\infty + f^2\frac{\delta^2}{12} \\ &\quad +\sigma^2 - f(a+f)\frac{\delta^2}{12}\Big) + p\left(a^2 P_\infty + \sigma^2\right) \\ &= \frac{(1-p)\left((f^2 - f(a+f))\frac{\delta^2}{12} + \sigma^2\right) + p\sigma^2}{1-((a+f)^2(1-p) + a^2 p)}. \end{aligned}$$

Since the cost function is given by

$$J_\infty = E\left[(q + rf^2)P_\infty - rf^2\frac{\delta^2}{12}\right],$$

it can be easily evaluated. For convergence, we have the additional condition

$$(1-p)(a+f)^2 + pa^2 < 1.$$

- Logarithmic Quantizer: Once again, the state evolves according to (5). For a logarithmic quantizer, $E[\Delta_k^2] = \frac{(\ln(g))^2}{12}E[x_k^2]$. Thus by a similar exercise as above, the steady state covariance is given by

$$P_\infty = \frac{\sigma^2}{1-(1-p)\left((a+f)^2 - fa(\frac{\ln(g)^4}{144})\right) - pa^2}.$$

Thus the cost function can again be evaluated.

E. Example

In this subsection, we consider a simple example to illustrate the above results. We consider the system described by

$$\begin{aligned} x_{k+1} &= 2x_k + u_k + w_k \\ u_k &= f\hat{x}_k. \end{aligned}$$

The initial condition x_0 is assumed to be uniformly distributed in the range $[-20, 20]$ while the white noise w_k is assumed to be uniformly distributed in the range $[-1, 1]$. The cost function we consider is

$$J = \lim_{k\to\infty} E\left[x_k^2 + u_k^2\right].$$

For this cost function, the optimal control law without quantization turns out to be $f = -1.618$. We use this control law to consider the performance of various quantization schemes considered above. For the quantizers that operate on a fixed range, the minimum region to be quantized is $[-20, 20]$. We will assume that the control law does not allow the system to go outside this range, thus avoiding issues like quantizer overflow.

Figure 2 shows the performance of the system when uniform and logarithmic quantizers are used. It can be seen that even for this simple system, logarithmic quantizer yields much better performance for the same number of bits. However, for convergence, the uniform quantizer requires 2 or more bits while for $\epsilon = 0.1$, the logarithmic quantizer requires at least 3 bits. Also, it may be noted that the plots provide merely a qualitative comparison since the expressions provided in the analysis are approximations. Figures 3 and 4 show a comparison of our theoretical approximations with simulation results for uniform and logarithmic quantizers respectively. The simulation results refer to the cost in steady state averaged over 10000 runs for a system using the particular quantizer. The initial condition and the noise driving the system were chosen randomly for each run. It can be seen that the approximations are quite good, at least in this example. Figure 5 shows a comparison of the performance achieved by the dynamic uniform quantizer with the performance bound derived using entropy arguments. Again we see that the bound is reasonably tight in this example. Of course, static quantizers perform much worse than the dynamic quantizers, especially when a small number of bits are used. Figure 6 shows the performance of the system as a function of the packet loss probability across a channel that drops packets in an i.i.d. fashion. A uniform quantizer with 6 bits is used. The system becomes unstable at the theoretical value of $p = 0.22$.

F. Extensions to Vector Processes

So far we have only considered the case of scalar processes. A more general case is when the process state x_k is a vector. The process evolution and measurement are

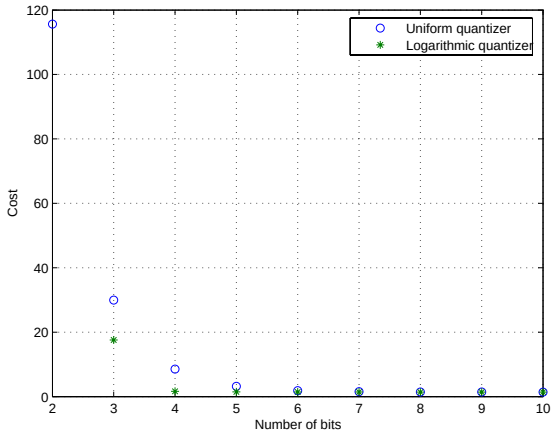

Fig. 2. Performance of static uniform and logarithmic quantizers.

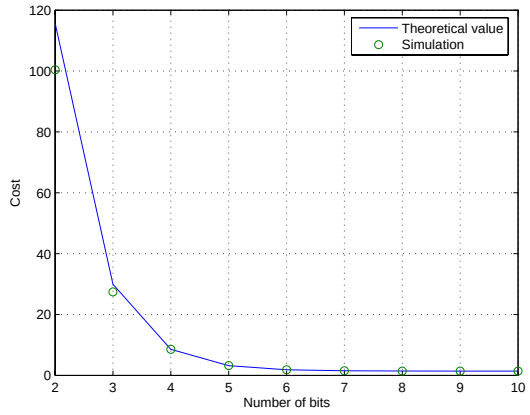

Fig. 3. Performance of approximations presented in the paper for uniform quantizer with simulation results.

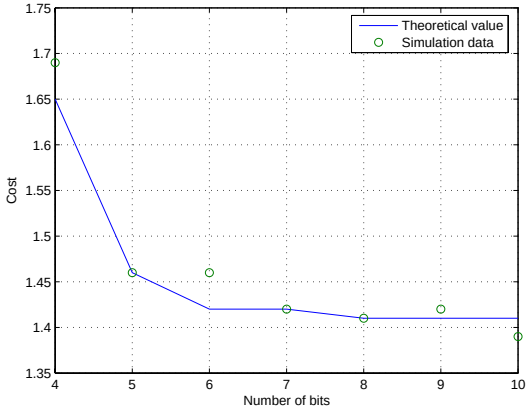

Fig. 4. Performance of approximations presented in the paper for logarithmic quantizer with simulation results.

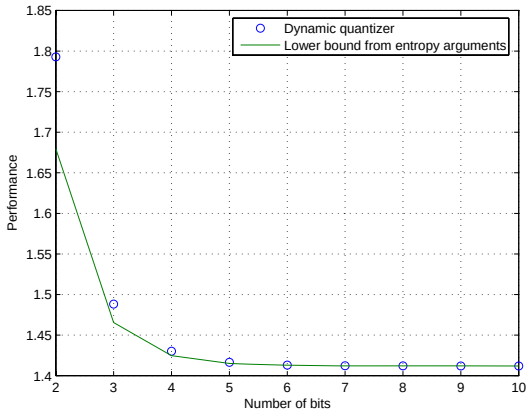

Fig. 5. Comparison of the dynamic quantizer performance with the lower bound derived in the paper.

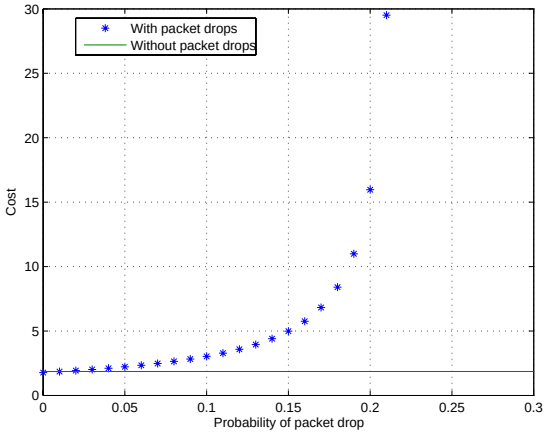

Fig. 6. Performance of the system across a packet dropping channel.

thus described by

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + w_k \\ y_k &= Cx_k + v_k, \end{aligned}$$

and the measurement vector y_k is quantized and sent over the channel. Studying quantization issues for such plants takes us into the realm of vector quantization theory, which is less well-developed than its scalar counterpart and hence the extension is not trivial. The basic difficulty is that each component of the vector y_k carries information about other components and hence it is extremely wasteful to do scalar quantization on each component separately. If the system matrix A is diagonal (or diagonalizable) and the matrix C is invertible, and hence this dependence is not present, the results from scalar quantization that we derive above can be used on each component. We are currently working on extending our results to the case when these assumptions do not hold.

IV. CONCLUSIONS AND FUTURE WORKS

In this work, we looked at a scalar system in which the state is being quantized prior to its transmission to the controller. We looked at the problem of evaluating the performance of the controller in minimizing a quadratic cost for uniform and logarithmic quantizers. We saw that the problem is difficult to solve exactly in most cases; however through approximations like high-rate assumption, we were able to evaluate the performance. We also provided a lower bound on the performance under any quantization. We also considered some simple extensions such as dynamic quantization and packet dropping channels.

There are many interesting directions in which this work can be extended. As discussed above, we are currently working on extending these results to the case of non-scalar processes. Another important question is identification of the quantity that should be quantized and sent across the link. This question is even more important in output feedback systems which will be the case for most vector processes. Some initial results are known in the area through the work of Ishwar et al. [13], but more research needs to be done in the area.

ACKNOWLEDGMENTS

Research supported in part by AFOSR grant F49620-01-1-0460 and by NSF grant CCR-0326554 for the first author and in part by the National Science Foundation under grant no. CCR-0133818, by David and Lucille Packard Foundation, and by Caltech's Lee Center for Advanced Networking for the second author.

REFERENCES

[1] J. Baillieul. Feedback designs for controlling device arrays with communication channel bandwidth constraints. In *ARO workshop on smart structures*, 1999.

[2] V. S. Borkar, S. K. Mitter, and S. Tatikonda. Markov control problems with communication constraints. *Comm. in Information and Systems*, 1(1):16–33, 2001.

[3] R. W. Brockett and D. Liberzon. Quantized feedback stabilization of linear systems. *IEEE Transactions on Automatic Control*, 45(7):1279–89, 2000.

[4] J. D. Bruce. Optimum quantization. Technical Report 429, Research Laboratory of electronics, Massachusetts Institute of Technology, Cambridge, 1965.

[5] T. M. Cover and J. A. Thomas. *Information Theory*. Wiley and Sons, New York, 1991.

[6] D. F. Delchamps. Stabilizing a linear system with quantized state feedback. *IEEE Transactions on Automatic Control*, 35:916–924, 1990.

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

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

[9] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1991.

[10] R. M. Gray. *Source Coding Theory*. Kluwer Academic Publishers, 1990.

[11] J. Hespanha, A. Ortega, and L. Vasudevan. Towards the control of linear systems with minimum bit-rate. In *Proc. of the 15th Int. Symp. Math. The. Netw. Sys.*, 2002.

[12] H. Ishii and B. A. Francis. Quadratic stabilization of sampled-data systems with quantization. *Automatica*, 39:1793–1800, 2003.

[13] P. Ishwar, R. Puri, K. Ramchandran, and S. S. Pradhan. On rate-constrained distributed estimation in unreliable sensor networks. *IEEE Journal on Selected Areas in Communications: Special issue on self-organizing distributed collaborative sensor networks*, 23(4):765–775, April 2005.

[14] R.L. Lagendijk. Information theoretic background of compression. http://www-it.et.tudelft.nl/html/education/courses/et4_089/pdf/dscdepsignals.pdf.

[15] M. D. Lemmon and Q. Ling. Control system performance under dynamic quantization: The scalar case. In *Proceedings of the IEEE Conference on Decision and Control*, 2004.

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

[17] S. Mitter. Control with limited information: the role of systems theory and information theory. *European Journal of Control*, 7:122–131, 2000.

[18] G. N. Nair, S. Dey, and R. J. Evans. Infimum data rates for stabilising Markov jump linear systems. In *Proc. of the 42nd IEEE Conference on Decision and Control*, pages 1176–81, 2003.

[19] G. N. Nair and R. J. Evans. Stabilizability of stochastic linear systems with finite feedback data rates. *SIAM Journal on Control and Optimization*, 2004. accepted.

[20] I. R. Petersen and A. V. Savkin. Multi-rate stabilization of multivariable discrete-time linear systems via a limited capacity communication channel. In *Proc. of the 40th IEEE Conference on Decision and Control*, pages 304–309, 2001.

[21] D. K. Sharma. Design of absolutely optimal quantizers for a wide class of distortion measures. *IEEE Transactions on Information Theory*, IT-24(6):693–702, November 1978.

[22] S. Tatikonda. *Control under Communication Constraints*. PhD thesis, MIT, Cambridge, MA, 2000.

[23] E. Verriest and M. Egerstedt. Control with delayed and limited information: A first look. In *Proceedings of the IEEE Conference on Decision and Control*, 2002.

[24] H. S. Witsenhausen. Separation of estimation and control for discrete time systems. *Proceedings of the IEEE*, 59(11):1557–1566, 1971.

[25] W. S. Wong and R. W. Brockett. Systems with finite communication bandwidth-part I: State estimation problems. *IEEE Transactions on Automatic Control*, 42(9):1294–1298, 1997.

[26] W. S. Wong and R. W. Brockett. Systems with finite communication bandwidth-part II: Stabilization with limited information feedback. *IEEE Transactions on Automatic Control*, 44(5), 1999.