State Estimation for Linear Discrete-Time Systems Using Quantized Measurements

Minyue Fu\textsuperscript{a}, Carlos E. de Souza\textsuperscript{b}

\textsuperscript{a}School of Electrical Engineering and Computer Science, The University of Newcastle, NSW 2308, Australia.

\textsuperscript{b}Department of Systems and Control,Laboratório Nacional de Computação Científica (LNCC/MCT),
Av. Getúlio Vargas 333, Petrópolis, RJ 25651-075, Brazil.

Abstract

In this paper, we consider the problem of state estimation for linear discrete-time dynamic systems using quantized measurements. This problem arises when state estimation needs to be done using information transmitted over a digital communication channel. We investigate how to design the quantizer and the estimator jointly. We consider the use of a logarithmic quantizer, which is motivated by the fact that the resulting quantization error acts as a multiplicative noise, an important feature in many applications. Both static and dynamic quantization schemes are studied. The results in the paper allow us to understand the tradeoff between performance degradation due to quantization and quantization density (in the infinite-level quantization case) or number of quantization levels (in the finite-level quantization case).

Key words: Quantized estimation, state estimation, discrete-time systems, networked control.

1 Introduction

Control and estimation using quantized information can be traced back to early days of control research. In particular, research into the so-called quantized linear quadratic Gaussian (LQG) control problem started in 1960’s; see, e.g., Lewis & Tou (1965) and Tou (1963). More broad attempts on quantized feedback control can be traced back further to the works of Kalman (1956) and Widrow (1961) on the effects of quantization errors to sampled-data feedback systems. The overwhelming success of networked control systems, especially for industrial control and automation, has brought a resurgence interest in quantized feedback control. Examples of works include Wong & Brockett (1997), Brockett & Liberzon (2000), Baillieul (2001), Elia & Mitter (2001), Nair & Evans (2000), Tatikonda & Mitter (2004), and Fu & Xie (2005, 2006). Recent attempts on the quan-
tized LQG problem include Tatikonda, Sahai & Mitter (2004), Matveev & Savkin (2004), and Fu (2008).

Similar to the classical control theory where state estimation plays an essential role, estimation based on quantized information is also critical to quantized feedback control. This has been well recognized in most of the references above. In addition, quantized estimation has a broad range of applications beyond feedback control. Examples include sensor network-based estimation and tracking (Epstein et. al., 2008; Tiwari et. al., 2005) and consensus networks (Carli et al., 2007, 2008). In addition, quantized estimation is a part of the solution to a more broad problem of network-based estimation where transmitted information suffers also from transmission delays and packet dropouts (Xiao, Xie & Fu, 2009; Epstein et. al., 2008; Tiwari et. al., 2005).

Traditional quantizers employ linear (or uniform) quantization. While they preserve information well when the input signal falls into the dynamic range of the quantizer, the number of quantization levels required for a given quantization step-size increases linearly as the dynamic range increases. This paper considers logarithmic quantizers where the quantization step-size grows exponentially as the input increases. The use of logarithmic quantizers is motivated by the fact they are shown to outperform linear quantizers in control problems, as demonstrated by Elia & Mitter (2001), and Fu & Xie (2005, 2006). When used for state estimation, logarithmic quantization leads to a multiplicative noise, rather than additive noise as in the case of linear quantiza-
tion. This allows us to have accurate estimation when the state is small and less accurate estimation when the state is large. That is, logarithmic quantization guarantees the relative error due to quantization to be roughly constant. This is a very important feature in many applications. Imagine the situation of a pilot looking for a runway: it is not necessary to have very accurate positioning of the runway when the plane is far away, but the positioning must be accurate when the runway gets close. Another major advantage of logarithmic quantization is that many physical measurements inherently carry multiplicative noises (i.e., the sensors are designed with a specified relative error). Optical sensors, infrared sensors and hall-effect sensors are among sensing devices with natural multiplicative noises. When a measured signal as such is further quantized by a logarithmic quantizer, the overall noise is still multiplicative.

In this paper, we study how to design a state estimator for a single-output linear discrete-time system when the measurements are subject to logarithmic quantization. The problem setting is the same as in the standard Kalman filtering problem, except that now we need to design the state estimator and quantizer jointly. Both infinite-level and finite-level quantizers are considered. For an infinite-level quantizer with a given quantization density, we propose a design method which can deliver good estimation performance and at the same time guarantees the stability of the state estimation error dynamics. For finite-level quantization, design methods are offered for both static quantizer and dynamic quantizer. The first case uses a truncated infinite-level quantizer, suitable for stable systems. The latter involves a dynamic scaling parameter which acts as a zoom-in/zoom-out function (similar to Brockett & Liberzon (2000)), allowing us to deal with unstable systems. We also demonstrate via examples how the proposed methods work. Simulation results suggest that near optimal performance can be achieved with a relatively low bit-rate quantizer.

2 Problem Formulation

Consider the following linear system:

\[
\begin{align*}
    x(k+1) &= Ax(k) + Bu(k), \quad x(0) = x_0 \\
    y(k) &= Cx(k) + v(k)
\end{align*}
\]

where \( x(k) \in \mathbb{R}^n \) is the state, \( w(k) \in \mathbb{R}^m \) is the process noise, \( y(k) \in \mathbb{R} \) is the measurement, \( v(k) \in \mathbb{R} \) is the measurement noise, and \( A, B \) and \( C \) are known matrices of appropriate dimensions. It is assumed that \( x_0 \in \mathbb{R}^n \) is a random variable with mean \( x_0 \) and covariance matrix \( \Sigma_0 \), and \( w(k) \) and \( v(k) \) are uncorrelated widely stationary white noises with zero mean and covariance matrices \( \Sigma_w \) and \( \Sigma_v \), respectively, and they are assumed to be uncorrelated with \( x_0 \) for all integers \( k \geq 0 \). It is further assumed that \( x(0) - x_0 \), \( w(k) \) and \( v(k) \) have even probability density functions.

Our quantized estimator consists of three parts: a quantizer, a digital communication channel and an estimator, as shown in Fig. 1. The channel is assumed to be free of transmission errors and time delay. Instead of quantizing the measured signal directly, we quantize the prediction error of the estimator. The estimator is chosen to be

\[
\begin{align*}
    \hat{x}(k+1) &= A\hat{x}(k) + LQ(y(k) - \hat{y}(k)), \quad \hat{x}(0) = \bar{x}_0 \\
    \hat{y}(k) &= C\hat{x}(k)
\end{align*}
\]

where \( \hat{x}(k) \in \mathbb{R}^n \) is the estimate of \( x(k) \), \( \hat{y}(k) \in \mathbb{R} \) is the estimate of \( y(k) \) based on \( \hat{x}(k) \), \( Q(\cdot) \) is the quantizer, and \( L \) is the estimation gain. Define the prediction error as

\[
\varepsilon(k) = y(k) - \hat{y}(k)
\]

and denote the quantization error by

\[
\varepsilon_q(k) = \varepsilon(k) - Q(\varepsilon(k)).
\]

Since the state estimate is constructed only using the quantized prediction error and communication is assumed to be noiseless, both sides of the channel can construct the same estimate using the quantized prediction error. In particular, the construction of \( \hat{x}(k) \) on the transmitter side does not require the estimated state to be transmitted back from the receiver side. Quantizing \( \varepsilon(k) \) is known to be better than quantizing \( \hat{y}(k) \) directly; see Tatikonda, Sahai & Mitter (2004) and Fu (2008).

We consider two types of quantizers, static ones and dynamic ones. A static quantizer takes one input sample and produces one output sample without referring back to the previous input samples. This is what is explicitly assumed in (2) and (4). A dynamic quantizer is more complex, allowing quantization to be done using the current and all past samples of the input.

Our objective of quantized state estimation is similar to that of stationary Kalman filter, namely to minimize the asymptotic variance of the estimation error defined by

\[
\lim_{k \to \infty} \mathcal{E}\{ (x(k) - \hat{x}(k))^T (x(k) - \hat{x}(k)) \}
\]

subject to certain constraints on the structure and information flow of the quantizer to be specified in later sections, where in the above \( \mathcal{E}\{ \cdot \} \) denotes expectation.
In this section, we employ a static logarithmic quantizer which is depicted in Figure 2 and described by

\[
Q(\varepsilon) = \begin{cases} 
\rho^i \mu_0, & \text{if } \frac{1}{1+\rho} \rho^i \mu_0 < \varepsilon \leq \frac{1}{1-\rho} \rho^i \mu_0, \\
0, & \text{if } \varepsilon = 0, \\
-Q(-\varepsilon), & \text{if } \varepsilon < 0 
\end{cases} 
\]

where \(\rho \in (0, 1)\) represents the quantization density and

\[
\delta = (1 - \rho)/(1 + \rho). 
\]

A small \(\rho\) (or large \(\delta\)) implies coarse quantization, and a large \(\rho\) (or small \(\delta\)) means dense quantization.

![Fig. 2. Logarithmic quantizer.](image)

### 3.1 Basic Properties

Defining the estimation error

\[
e(k) := x(k) - \hat{x}(k)
\]

the estimation error dynamics can be described by the following state-space model:

\[
\begin{cases} 
\varepsilon(k+1) = A\varepsilon(k) + Bw(k) - LQ(\varepsilon(k)) \\
\varepsilon(k) = Ce(k) + v(k) 
\end{cases}
\]

(8)

Our task is to choose \(\rho\) and \(L\) so that the asymptotic estimation error variance (5) is minimized.

As observed in Fu & Xie (2005), a logarithmic quantizer is easily bounded by a sector bound, namely

\[
|Q(\varepsilon) - \varepsilon| \leq \delta |\varepsilon|. 
\]

(9)

Using the above, we may rewrite (8) as

\[
e(k+1) = Ae(k) - L\varepsilon(k) + Bw(k) + L\Delta(k)\varepsilon(k) 
\]

(10)

where

\[
\Delta(k) = \begin{cases} 
\varepsilon_q(k)/\varepsilon(k), & \text{if } \varepsilon(k) \neq 0, \\
0, & \text{otherwise} 
\end{cases} 
\]

(11)

with the property that \(|\Delta(k)| \leq \delta\) for all \(k\).

Given the sector bound for \(\Delta(k)\) as above, we consider an auxiliary uncertain system defined by

\[
z(k+1) = (A - LC)z(k) - Lv(k) + Bw(k) + L\Delta_k(Cz(k) + v(k)), \quad |\Delta_k| \leq \delta. 
\]

(12)

Note that (12) differs from (10) in the sense that \(\Delta_k\) is an arbitrary function, whereas \(\Delta(k)\) in (10) is due to the quantizer \(Q(\cdot)\). It turns out that \(\Delta(k)\) in (10) can be viewed as a special instance of \(\Delta_k\).

We first present some key properties for the auxiliary system (12).

**Theorem 3.1** The estimation error dynamics (10) has the following properties:

(a) The estimation error \(e(k)\), the prediction error \(\varepsilon(k)\), and the quantization error \(\varepsilon_q(k)\) have zero-mean and an even probability density function for all \(k \geq 0\);

(b) The estimation error dynamics (8) is quadratically stable if and only if the auxiliary system (12) is quadratically stable, i.e., there exists a matrix \(X = X^T > 0\) such that

\[
e^T X e > (Ae - LQ(Ce))^T X (Ae - LQ(Ce)) \quad (13)
\]

for all nonzero \(e \in \mathbb{R}^n\) if and only if there exists a matrix \(P = P^T > 0\) such that

\[
P > (A - L(1 - \Delta)C)^T P(A - L(1 - \Delta)C), \quad \forall |\Delta| \leq \delta \quad (14)
\]

(c) If the auxiliary system (12) is quadratically stable, then the covariance matrix of \(e(k)\) is bounded and asymptotically invariant;

(d) The minimum quantization density \(\rho_{\inf}(L)\) for the auxiliary system (12) to be quadratically stable for a given \(L\) is given by

\[
\rho_{\inf}(L) = \frac{1 - \delta_{\sup}(L)}{1 + \delta_{\sup}(L)} 
\]

(15)

where

\[
\delta_{\sup}(L) = 1/\|C(zI - A + LC)^{-1}L\|_{\infty}. 
\]

(16)

**Proof.** The statement (a) can be easily shown by induction. Since \(\hat{x}(0) = \bar{x}_0\), then \(e(0)\) is zero-mean with an even probability density. Note that \(Q(\cdot)\) is an odd function. Suppose \(e(k)\) is so too for some \(k\). Then, it follows from (8) that \(e(k+1)\) is also zero-mean with an even probability density. Hence, by induction, \(e(k)\) is zero-mean with an even probability density for all \(k \geq 0\). In
addition, in view of (4) and since \( \varepsilon(k) = Ce(k) + v(k) \), it follows that \( \varepsilon(k) \) and \( \varepsilon_q(k) \) have zero-mean and an even probability density function for all \( k \geq 0 \).

The statement (b) is proved in Fu & Xie (2005). To show the statement (c), we assume that (14) holds for some matrix \( P = P^T > 0 \). It follows that

\[
(1-2\eta)P > (A-L(1-2\Delta)C)^TP(A-L(1-\Delta)C), \quad \forall \, 1 \leq \Delta \leq \delta
\]

for some sufficiently small scalar \( \eta > 0 \). Next, define the Lyapunov function \( V(e) = e^TPe \) for system (10). Considering (10) and denoting \( \tilde{\Delta}(k) = 1 - \Delta(k) \), we get

\[
V(e(k+1)) = [Ae(k) - L\tilde{\Delta}(k)e(k) + Bw(k)]^TP[Ae(k) - L\tilde{\Delta}(k)e(k) + Bw(k)]
\]

\[
+ [-L\tilde{\Delta}(k)e(k) + Bw(k)]^TP[-L\tilde{\Delta}(k)e(k) + Bw(k)]
\]

\[
- 2e^TP[A - L\tilde{\Delta}(k)C]^TP[L\tilde{\Delta}(k)e(k) + Bw(k) - e(k)]
\]

\[
\leq (1+\tau)e^TP(A - L\tilde{\Delta}(k)C)^TP[A - L\tilde{\Delta}(k)C]e(k)
\]

\[
+ (1+\tau^{-1})[-L\tilde{\Delta}(k)e(k) + Bw(k)]^TP[-L\tilde{\Delta}(k)e(k) + Bw(k)]
\]

for any scalar \( \tau > 0 \). To obtain the inequality above, we have used the well-known triangular inequality that \( 2a^Tb \leq \tau a^Tb + \tau^{-1}b^Tb \) for any column vectors \( a \) and \( b \) of the same dimension. In particular, we may choose \( \tau \) such that \((1-2\eta)(1+\tau) = 1-\eta \). Then, it follows that

\[
V(e(k+1)) \leq (1-\eta)\rho V(e(k)) + m_1\rho^2 + m_2\rho^T\rho
\]

for some sufficiently large \( m_1 \) and \( m_2 \) independent of \( k \). Applying the result above recursively, we obtain

\[
V(e(k)) \leq (1-\eta)^k V(e(0)) + \sum_{i=1}^{k} (1-\eta)^{k-i}(m_1\rho^2 + m_2\rho^T\rho)
\]

which implies

\[
\text{Tr}(e(k)e^T(k)) \leq (1/\lambda_{\text{min}}(P)) \left\{ (1-\eta)^k V(e(0)) + \sum_{i=1}^{k} (1-\eta)^{k-i}(m_1\rho^2 + m_2\rho^T\rho) \right\}
\]

where \( \text{Tr}\{\cdot\} \) denotes matrix trace and \( \lambda_{\text{min}}(P) \) is the minimum eigenvalue of \( P \). Defining \( R_{ee}(k) := \mathcal{E}\{e(k)e^T(k)\} \), the latter inequality leads to

\[
\text{Tr}(R_{ee}(k)) \leq (1/\lambda_{\text{min}}(P))(1-\eta)^k \mathcal{E}\{V(e(0))\}
\]

\[
\quad + \left( m_1\Sigma_v + m_2\text{Tr}(\Sigma_w) \right) \sum_{i=1}^{k} (1-\eta)^{k-i}
\]

\[
\leq \tilde{m}_0 \text{Tr}(R_{ee}(0)) + \tilde{m}_1\Sigma_v + \tilde{m}_2\text{Tr}(\Sigma_w)
\]

for some constants \( \tilde{m}_0, \tilde{m}_1 \) and \( \tilde{m}_2 \). Hence, \( R_{ee}(k) \) is bounded, which implies the boundedness of the covariance matrix of \( e(k) \).

The proof of the asymptotic invariance of the covariance matrix of \( e(k) \) is obtained by noting that

\[
\mathcal{E}\{V(e(k+1)) - V(e(k))\} \to -\infty \text{ as } \|e(k)\| \to \infty
\]

and using arguments similar to those in Kushner (1971, Chapter 8) related to the probability measure of \( e(k) \).

The statement (d) follows from the known fact in robust stability analysis that \((12) \leq \tau \mathcal{E}\{C(zI-A-LC)^{-1}L\} \leq \delta^{-1} \) see Packard & Doyle (1990). Therefore, the largest \( \delta \) to maintain quadratic stability is given by \((16) \) and the minimum quantization density \( \rho_{\text{int}}(L) \) is related to \( \delta_{\sup}(L) \) by \((15) \).

3.2 Asymptotic Covariance Matrix of Estimation Error

We now proceed to quantify the asymptotic covariance matrix of \( e(k) \). Denote by \( E(k) \) the covariance matrix of \( e(k) \) and its asymptotic version by \( E(k) = \lim_{k \to \infty} E(k) \). We assume that \( \rho > \rho_{\text{int}}(L) \) so that \( E(k) \) is bounded (by Theorem 3.1 (c)). In the sequel it is assumed that \( x_0 - \bar{x}_0, w(k) \) and \( v(k) \) are Gaussian distributed. Note that in view of Theorem 3.1 (a) the latter assumption implies that the estimation error, the prediction error and the quantization error have zero mean and an even probability density function. Moreover, we will denote by \( \sigma^2_q \) and \( \sigma^2_e \) the asymptotic variances of \( e(k) \) and \( e_q(k) \), respectively, and define

\[
\tilde{e}_q^2 = \sigma^2_q / \sigma^2_e
\]

(17)

to be the normalized quantization error variance.

The computation of \( E(k) \) is complicated by the fact that \( Q(\cdot) \) is a nonlinear function. But when the number of quantization levels is not too small, the following conditions hold very well in numerical simulations.

C1. The quantization error \( e_q(k) \) is uncorrelated with \( \tilde{e}(k+1) := e(k) - L\tilde{\Delta}(k)e(k) + Bw(k) \) (note that the latter is the predicted state estimation error without quantization error);

C2. Asymptotically, the prediction error \( \varepsilon(k) \) is approximately Gaussian distributed with zero mean and variance \( \sigma^2_e \).

Remark 3.1 Some remarks on the above conditions are in order. We first note the well known fact (Anderson & Moore, 1979) that, if there had been no quantization before time \( k \), then \( \tilde{e}(k+1) \) would be uncorrelated with \( \varepsilon(k) \), thus independent of \( \varepsilon(k) \) because both would be Gaussian distributed. Hence, if \( \varepsilon(k) \) were quantized, its quantization error \( e_q(k) \) would be uncorrelated with \( \tilde{e}(k+1) \). Now because quantization happened before time \( k \), \( \varepsilon(k+1) \) and \( \varepsilon(k) \) (thus \( e_q(k) \)) are correlated in general. However, the correlation is typically weak. In
particular, if the quantization density is relatively high, the effect of past quantization errors should be negligible and it is thus fair to assume $\delta(k+1)$ and $\varepsilon_q(k)$ to be uncorrelated. For the same reason, the Gaussian distribution assumption is also approximately valid when the quantization density is relatively high.

Under Condition C2, we may relate the variance of the quantization error $\varepsilon_q(k)$ to that of the prediction error $\varepsilon(k)$. We observe that $\varepsilon_q(k)$ is influenced by the choice of $\mu_0$ in (6). However, two simple properties are easily observed from (6):

P1. The quantization error $\varepsilon_q(k)$ is periodic in $\mu_0$ in a logarithmic scale, i.e., if $\mu_0$ is multiplied by $\rho^j$ for any integer $j$, $\varepsilon_q(k)$ remains the same;

P2. A logarithmic quantizer is linearly scalable in the sense that if $\varepsilon(k)$ is multiplied by $\rho^j$ for any integer $j$, $\varepsilon_q(k)$ is multiplied by the same factor.

In fact, the influence of $\mu_0$ to $\varepsilon_q(k)$ is negligible for small values of $\delta$. This means that $\sigma_q^2$ is approximately proportional to $\sigma^2$ for a given $\delta$. That is, for a given $\delta$, the value of $\sigma_q^2$ in (17) is approximately constant. For a Gaussian distributed prediction error, the assertion above is demonstrated in Fig. 3 which is produced by Monte-Carlo simulations. In the figure, the upper and lower bounds are the maximum and minimum values of $\sigma_q^2$ with respect to $\mu_0$, and they are indeed very close, especially for small $\delta$ (up to 0.3). It turns out that the actual $\sigma_q^2$ can be well approximated by

$$\sigma_q^2 \approx \tilde{\delta}^2 := \frac{1 + 0.45\delta^2}{3} \delta^2$$

which is also shown in Fig. 3.

![Normalized quantization error variance](image)

**Fig. 3.** Estimates of normalized quantization error variance.

We now provide an estimate for the asymptotic covariance matrix of $\varepsilon(k)$. Consider the quantized estimation error dynamics (8). We suppose that Conditions C1 and C2 and the approximation (18) for $\sigma_q^2$ hold and $k \to \infty$.

Using Condition C1, it follows from (8) that the asymptotic covariance matrix of $\varepsilon(k)$, denoted by $E$, satisfies

$$E = (A-LC)E(A-LC)^T + B\Sigma_w B^T + L\Sigma_v L^T + \sigma_q^2 LL^T.$$  

Using (17), (18) and considering that $\sigma_q^2 = CEC^T + \Sigma_v$, we can approximate $E$ by the solution $\tilde{E}$ by the following generalized Lyapunov equation:

$$\tilde{E} = (A-LC)\tilde{E}(A-LC)^T + B\Sigma_w B^T + L\Sigma_v L^T + \tilde{\delta}^2 L(CEC^T + \Sigma_v)L^T.$$  

(19)

Note that if a solution $\tilde{E}$ to (19) exists, it is unique and positive semidefinite. In connection with (19), consider the generalized Lyapunov difference equation as follows:

$$\tilde{E}(k+1) = (A-LC)\tilde{E}(k)(A-LC)^T + B\Sigma_w B^T + L\Sigma_v L^T + \tilde{\delta}^2 L(C\tilde{E}C^T + \Sigma_v)L^T, \quad \tilde{E}(0) = \Sigma_0.$$  

(20)

**Theorem 3.2** Suppose the system (8) is quadratically stable. Then $\tilde{E} = \lim_{k \to \infty} \tilde{E}(k)$ exists and is finite, and is also the positive semidefinite solution to (19).

**Proof.** Suppose that system (8) is quadratically stable. By Theorem 3.1, $\|\delta C(zI - A + LC)^{-1}\|_\infty < 1$. Using the discrete-time bounded-real lemma (de Souza & Xie, 1992), there exists a matrix $\Omega = \Omega^T > 0$ such that

$$1 - \delta^2 C\Omega C^T > 0,$$

$$\Omega - LL^T - A_c(\Omega^{-1} - \delta^2 C^T C)^{-1} A_c^T > 0$$

where $A_c = A - LC$. Since $(\Omega^{-1} - \delta^2 C^T C)^{-1} \geq \Omega$, the above inequalities imply

$$\Omega - A_c \Omega A_c^T > \delta^2 L C \Omega C^T L^T.$$  

(21)

We denote

$$\tilde{E} = \alpha \Omega, \quad \Upsilon = (1 + \tilde{\delta}^2) L \Sigma_v L^T + B\Sigma_w B^T$$

where $\alpha > 0$ is a scaling parameter. Since (21) is linear in $\Omega$ and is a strict inequality, it follows that there exists a sufficiently large $\alpha > 0$ such that $\tilde{E} \geq \Sigma_0$ and

$$\tilde{E} - A_c \tilde{E} A_c^T > \delta^2 L C \tilde{E} C^T L^T + \Upsilon.$$  

(22)

We will show that $\tilde{E}(k) \leq \tilde{E}$ for all $k \geq 0$. This can be proved by induction. Note that $\tilde{E}(0) \leq \tilde{E}$ and $\delta < \tilde{\delta}$. Suppose $\tilde{E}(k) \leq \tilde{E}$ for some $k$. Then, from (20) and (22),

$$\tilde{E}(k+1) \leq A_c \tilde{E}(k) A_c^T + \delta^2 L C \tilde{E}(k) C^T L^T + \Upsilon \leq \tilde{E}.$$  

Hence, $\tilde{E}$ is indeed an upper bound of $\tilde{E}(k)$ for all $k$.

Now we use the result above to show the convergence of $\tilde{E}(k)$. Note that (20) is a linear difference equation in $\tilde{E}(k)$. Using standard properties of Kronecker product, (20) can be rewritten as
where $\hat{E}_v(k)$ and $\Upsilon_v$ are the vector forms of $\hat{E}(k)$ and $\Upsilon$, respectively, and $\hat{A}$ is a matrix that depends on $A$, $L$, $C$ and $\delta$. Since $\hat{E}$ is bounded for any bounded input $\Upsilon$ and initial state $\hat{E}_v(0)$, (23) has bounded-input, bounded-output stability. This in turn implies that (23) has asymptotic stability, i.e., $\hat{A}$ is Schur stable. It follows that $\hat{E}_v$ converges to a constant vector as $k \to \infty$ (because the input $\Upsilon_v$ is constant). Therefore, $\hat{E} = \lim_{k \to \infty} \hat{E}(k)$ exists and is finite, and is also the (unique) positive semidefinite solution to (19).

Remark 3.2 Notice that (20) and (19) can be viewed as the equations defining respectively the covariance matrix and the stationary covariance matrix of the signal $\tilde{e}(k)$ given by the following system with multiplicative noise:

$$
\tilde{e}(k + 1) = (A - LC)\tilde{e}(k) + Bw(k) - (1 + \tilde{\delta}^2)\frac{\tilde{\delta}}{2}\nu(k) - \tilde{\delta}\xi(k)LC\tilde{e}(k), \quad \tilde{e}(0) = \tilde{e}_0
$$

where $\tilde{e}(k) \in \mathbb{R}^n$, $w(k)$ and $\nu(k)$ are the same white noise signals as in system (1), $\tilde{e}_0 \in \mathbb{R}^n$ is a zero-mean random variable with covariance matrix $\Sigma_0$, and $\xi(k)$ is a scalar widely stationary zero-mean white noise sequence with unitary variance and uncorrelated with $w(k)$, $\nu(k)$ and $\tilde{e}_0$. In the absence of quantization noise, i.e., $\tilde{\delta} = 0$, (24) becomes the state equation of the estimation error for the quantization-free state estimation problem for system (1) and filter (2). Note that $(1 + \tilde{\delta}^2)\frac{\tilde{\delta}}{2} = 1$ when $\delta$ is relatively small. Hence, as far as the variance of the estimation error is concerned, the main effect of quantization amounts to a multiplicative noise $\tilde{\delta}\xi(k)LC\tilde{e}(k)$.

3.3 Design of Estimation Gain

So far, we have assumed that the estimation gain $L$ is given. We now discuss how to design $L$. From (19), it is natural to choose $L$ to minimize $\hat{E}$. If $\delta$ (and thus $\tilde{\delta}$) is small and the Kalman gain $1 - LK$, which is the optimal $L$ when $\delta = 0$, is not large, it is typically sufficient to choose $L = L_K$. In general, the following result can be used:

Theorem 3.3 The optimal $L$ that minimizes $\text{Tr}(\hat{E})$ in (19) can be found by solving the following generalized discrete-time algebraic Ricatti equation for a symmetric positive-definite matrix $\hat{E}$:

$$
\hat{E} = A\hat{E}A^T + B\Sigma_wB^T - A\hat{E}C^TC\hat{E}A^T S^{-1},
$$

and the optimal estimation gain $L$ is given by

$$
L = A\hat{E}C^TS^{-1}.
$$

Equivalently, we can obtain $\hat{E}$ in (25) by solving the following convex optimization problem:

$$
\begin{aligned}
\min_{\hat{E}} \text{Tr}(Q), \text{ subject to } \\
\begin{bmatrix}
P & PA \\
I & P
\end{bmatrix} \succeq 0 \\
\begin{bmatrix}
A^TP & (1 + \delta^2)(P + C^T\Sigma_v^{-1}C) \\
\delta^2 P & 0 & \Sigma_v^{-1} \\
\delta A^TP & 0 & 0 & (1 + \delta^2)P
\end{bmatrix} \succeq 0
\end{aligned}
$$

with the optimal $\hat{E}$ given by $\hat{E} = P^{-1}$.

Proof. Expanding the right hand side of (19) and regrouping the terms, we get

$$
\hat{E} = A\hat{E}A^T + B\Sigma_wB^T - A\hat{E}C^TC\hat{E}A^T S^{-1} + (L - A\hat{E}C^TS^{-1})S(L - A\hat{E}C^TS^{-1})^T
$$

where $S$ is as in (26). Nullifying the term that involves $L$ will minimize $\hat{E}$, which yields (25) and (27).

To show (28) and (29), consider the following equation:

$$
\hat{E}_\Omega = (A - LC)\hat{E}_\Omega(A - LC)^T + \delta^2 LC\hat{E}_\Omega C^TL^T + (1 + \delta^2)L\Sigma_vL^T + B\Sigma_wB^T + \Omega
$$

where $\Omega = \Omega^T \geq 0$. It is clear that $\hat{E}_\Omega$ is a monotonically increasing function of $\Omega$. Hence, in view of (30) and considering the optimal $L$ in (27), the problem

$$
\min_{\hat{E}} \text{Tr}(\hat{E}), \text{ subject to } \\
\hat{E} > A\hat{E}A^T + B\Sigma_wB^T - A\hat{E}C^TC\hat{E}A^T S^{-1}
$$

gives the (unique) solution of $\hat{E}$ to (25). Now, applying the matrix inversion lemma and standard matrix manipulations, we can rewrite (32) as

$$
\hat{E} > (1 + \delta^2)^{-1}A[\delta^2 \hat{E} + (\hat{E}^{-1} + C^T\Sigma_v^{-1}C)^{-1}]A^T + B\Sigma_wB^T
$$

Denoting $P = \hat{E}^{-1}$ and applying Schur’s complement, it can be readily verified that the latter inequality is equivalent to (29). Finally, it is easy to check that $\min \text{Tr}(\hat{E})$ is the same as $\min \text{Tr}(Q)$ subject to (28).

3.4 Illustrative Example

We now give an example to demonstrate the accuracy of the estimate $\hat{E}$ and the design of estimation gain. We will call the optimal $L$ in (27) a robust estimation gain due to the fact that it is designed to mitigate quantization errors. The gain $L$ designed without considering quantization errors will be called the Kalman gain.

The example we consider is a low-pass filtered random process corrupted by a measurement noise. More specifically, the system model is given by (1) with
and $\Sigma_w = 1$. Different values of $\Sigma_v$ will be considered. The filter has a normalized bandwidth of approximately 0.25 (where 1 corresponds to the Nyquist bandwidth).

Two cases, $\Sigma_v = 1$ and $\Sigma_v = 1/16$, are tested. The range of $\delta$ for the tests is chosen to be [0, 0.3]. For a given $\Sigma_v$ and $\delta$, we have designed two estimator gains, one taken as the Kalman gain designed by ignoring the quantization error and the other being the robust gain computed using (27). Quadratic stability of (8) is verified using (16) for both gains at $\delta = 0.3$.

Figs. 4 and 5 show the simulated values of $\text{Tr}(E)$ for both estimator gains along with their estimates $\text{Tr}(\hat{E})$. Fig. 4 is for $\Sigma_v = 1$ and Fig. 5 for $\Sigma_v = 1/16$. From these figures, we see that when the measurement noise is relatively large ($\Sigma_v = 1$), the Kalman gain performs well (and is actually slightly better than the robust gain). But when the measurement noise is relatively low ($\Sigma_v = 1/16$), the robust gain performs significantly better than the Kalman gain. This is because when $\Sigma_v$ is small, Kalman estimation relies heavily on the measurement, which is thus sensitive to quantization errors. In contrast, the robust gain is designed to cope with quantization errors, so it performs better when $\Sigma_v$ is small and the quantization error dominates. Also seen in Figs. 4 and 5 is that, in all cases, the estimate $\text{Tr}(\hat{E})$ matches the actual $\text{Tr}(E)$ very well, especially for small $\delta$.

4 State Estimation with Finite-Level Quantization

In this section, we study state estimation with a finite-level quantizer. The estimator structure (2)-(4) is used.

4.1 Truncated Logarithmic Quantization

A finite-level quantizer can be designed by simply truncating a logarithmic quantizer, i.e., we saturate the signal when it is too large (in magnitude) and have a dead-zone when the signal is too small. A $2N$-level logarithmic quantizer with quantization density $\rho$ is given by

$$Q(\varepsilon) = \begin{cases} 
\rho^i\mu_0, & \text{if}\; \frac{1}{1+\delta}\rho^i\mu_0 < \varepsilon < \frac{1}{1-\delta}\rho^i\mu_0, \\
\mu_0, & \text{if}\; \varepsilon < 0,
\end{cases}$$

(33)

where $\delta$ and $\mu_0$ both are to be optimized for a given $N$.

To set up this optimization problem, we assume that Conditions C1 and C2 hold. Recall that the zero-mean property for the estimation error, prediction error and quantization error is guaranteed when $x_0$, $w(k)$ and $v(k)$ have even probability density functions. The optimization problem can be written as follows:

$$\min J(\mu_0, \delta)$$

(34)

where $J(\mu_0, \delta)$ is the asymptotic variance of the quantization error. We have that

$$J(\mu_0, \delta) = 2(I_1 + I_2 + I_3)$$

(35)

where

Fig. 5. Infinite-level logarithmic quantization for $\Sigma_v = 1/16$.

Fig. 4. Infinite-level logarithmic quantization for $\Sigma_v = 1$. 

\begin{align*}
A &= \begin{bmatrix} 2.4744 & -2.8110 & 1.7038 & -0.5444 & 0.0723 \\
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
\end{bmatrix}, \\
B^T &= \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}, \\
C &= \begin{bmatrix} 0.245 & 0.236 & 0.384 & 0.146 & 0.035 \end{bmatrix}
\end{align*}
The joint optimization of \( \mu_0 \) and \( \delta \) is a difficult problem. Therefore, we settle for a suboptimal solution. We first choose \( \mu_0 \) with the aim that the input signal to the quantizer will be within the unsaturated region as much as possible, i.e., we want to choose \( \mu_0 \) to maximize \( I_1 \) for a fixed \( \delta \). Since \( I_1 \) corresponds to the case where no saturation occurs, our earlier analysis for infinite-level quantization still applies and we get

\[
I_1 := \int \frac{(1-\delta^{-1})^{-1}}{(1-\delta^{-1})^{-1}} (\varepsilon - Q(\varepsilon))^2 p(\varepsilon) d\varepsilon
\]

and \( p(\varepsilon) \) is the probability density function of \( \varepsilon(k) \).

The results of this optimization are shown in Table 1 for different values of \( \delta \), and setting it to zero, we get the maximizing value \( \mu_0^* \) for \( \mu_0 \) as follows:

\[
\mu_0^* = \sigma_e (1 - \delta) \mu
\]

where

\[
\mu = \sqrt{\frac{6N \ln(1/\rho)}{1 - \rho^2 N}}.
\]

Table 1

<table>
<thead>
<tr>
<th>( N_b )</th>
<th>( N )</th>
<th>( \delta )</th>
<th>( \rho )</th>
<th>( \mu_0/\sigma_e )</th>
<th>( (\mu_0^*, \delta)/\sigma_e^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>0.5338</td>
<td>0.3040</td>
<td>1.7699</td>
<td>0.1457</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>0.3253</td>
<td>0.5091</td>
<td>2.7220</td>
<td>0.04892</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>0.1909</td>
<td>0.6794</td>
<td>3.4887</td>
<td>0.01568</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>0.1095</td>
<td>0.8026</td>
<td>4.0931</td>
<td>0.00494</td>
</tr>
<tr>
<td>6</td>
<td>32</td>
<td>0.0619</td>
<td>0.8834</td>
<td>4.5774</td>
<td>0.00153</td>
</tr>
<tr>
<td>7</td>
<td>64</td>
<td>0.0346</td>
<td>0.9331</td>
<td>4.9779</td>
<td>0.00047</td>
</tr>
<tr>
<td>8</td>
<td>128</td>
<td>0.0191</td>
<td>0.9625</td>
<td>5.3134</td>
<td>0.00014</td>
</tr>
</tbody>
</table>

Optimized quantization density.

is given in (19). The term \( \tilde{\delta}^2 L(C \tilde{E}C^T + \Sigma_v) L^T \) in (19) represents the asymptotic quantization error variance \( \sigma_q^2 \) and thus needs to be replaced with \( J(\mu_0, \delta) \). Using (40), we can simply replace \( \tilde{\delta}^2 \) with \( \tilde{J}(\mu, \delta) \) using the optimized \( \delta \). That is, (19) is revised to be

\[
\tilde{E} = (A - LC) \tilde{E}(A - LC)^T + BS_{\Sigma_v} B^T + L\Sigma_v L^T + \tilde{J}(\mu, \delta)L(C \tilde{E}C^T + \Sigma_v) L^T.
\]

Example

The results above are demonstrated using the same example as in the previous section. Monte-Carlo simulations are shown in Figs. 6 and 7. For our example using the robust estimation gain. For C1, we compute the correlation coefficient \( \rho_u(\mu) \) and each component of \( \tilde{e}(k+1) = A\tilde{e}(k) - L\tilde{z}(k) + Bu(k) \) and denote by \( \rho \) the vector whose entries are these correlation coefficients. For our example with \( \Sigma_v = 1/16 \), the \( \ell_2 \) norm of the sequence \( \rho \) is found to be 0.028 for \( N_b = 2 \), 0.007 for \( N_b = 3 \) and 0.0013 for \( N_b = 4 \). It is clear that Condition C1 holds well. Fig. 8 shows the normalized autocorrelation function of the asymptotic prediction error for the case of \( \Sigma_v = 1/16 \). We see that the prediction error samples are slightly correlated for \( N_b = 2 \) but practically uncorrelated for \( N_b > 2 \). Fig. 9 shows the probability density function of the asymptotic prediction error for the case of \( N_b = 2 \), computed using simu-
Fig. 6. Finite-level logarithmic quantization for $\Sigma_v = 1$.

Fig. 7. Finite-level logarithmic quantization for $\Sigma_v = 1/16$.

lated data and normalized to have a unity variance, along with a standard Gaussian probability density function. We see that the computed probability density function fits a Gaussian probability density well even for $N_b = 2$. Hence, Condition C2 holds well too.

4.2 Dynamic Scaling

The use of any finite-level quantizer can potentially create a stability problem when the system (1) is unstable and the initial state is too large or there is a burst of large process noise. To overcome this problem, we introduce a dynamic scaling method borrowed from Fu & Xie (2009). The idea is to scale $\varepsilon(k)$ so that it is within the quantization range $\pm [\rho^{N_b-1}\mu_0, \mu_0]$ as much as possible. To do so, we modify (2) and (4) to respectively

$$\hat{x}(k + 1) = A\hat{x}(k) + Lg_k^{-1}Q(g_k \varepsilon(k))$$  \hspace{1cm} (43)

$$\varepsilon_q(k) = \varepsilon(k) - g_k^{-1}Q(g_k \varepsilon(k))$$  \hspace{1cm} (44)

where $\varepsilon(k)$ is the prediction error, i.e. $\varepsilon(k) = y(k) - C\hat{x}(k)$, and $g_k$ is the scaling parameter at time $k$ defined recursively by $g_0 = 1$ and

$$g_{k+1} = \begin{cases} 
  g_k \gamma_1, & \text{if } |Q(g_k \varepsilon(k))| = \mu_0, \\
  g_k / \gamma_2, & \text{if } |Q(g_k \varepsilon(k))| = \rho^{N_b-1}\mu_0, \\
  g_k, & \text{otherwise}
\end{cases}  \hspace{1cm} (45)$$

where $\gamma_1, \gamma_2 \in (0, 1)$ are design parameters: $\gamma_1$ makes the $g_{k+1}$ smaller than $g_k$, thus plays a zoom-out role; similarly, $\gamma_2$ plays a zoom-in role. Note that the change of $g_k$ is implicitly expressed in the quantized output by checking whether it is saturated, in the dead zone or not. Thus, no explicit transmission of $g_k$ is required.

The following result is quoted from Fu & Xie (2009):
Theorem 4.1 Consider the estimation error dynamics (8) with the infinite-level logarithmic quantizer (6). Suppose L and ρ are chosen such that (8) is quadratically stable, i.e. (by Theorem 3.1 (b)), there exists a Lyapunov matrix \( P = P^T > 0 \) such that

\[
(A - L(1-\Delta)C)^T P (A - L(1-\Delta)C) < P, \quad \forall |\Delta| \leq \delta
\]

(46)

where \( \delta \) is related to \( \rho \) by (7). Define

\[
N_0 = 1 + \frac{2 \log(\gamma_2 - \sqrt{1-\eta}) - \log(L^T PLCP^{-1}C^T)}{2 \log(\rho)}
\]

(47)

where \( 0 < \eta < 1 \) is chosen such that

\[
(A - L(1-\Delta)C)^T P (A - L(1-\Delta)C) \leq (1-\eta)P
\]

(48)

for all \( |\Delta| \leq \delta \), and \( \gamma_2 \) satisfies \( \sqrt{1-\eta} < \gamma_2 < 1 \). Also, take \( 0 < \gamma_1 < 1 \) such that the matrix \( \gamma_1 A \) is Schur stable and let \( N \geq N_0 \). If the \( 2N \)-level quantizer (33) together with the estimator (43) are used instead, then the estimation error dynamics is bounded asymptotically if the noise signals are uniformly bounded, i.e., \( |w(k)| \leq \bar{w}, |v(k)| \leq \bar{v} \) for some constants \( \bar{w} \) and \( \bar{v} \).

5 Conclusion

In this paper, we have studied the use of logarithmic quantizers in the quantized state estimation problem. Both infinite-level and finite-level quantizers are treated. For an infinite-level static quantizer, a number of results are given to approximate the asymptotic variance of the state estimation error for a given quantization density, which in turn yields clear relationship between quantization density and the asymptotic estimation error variance. The aforementioned results have also been generalized to the case where a fixed-rate finite-level quantizer is used. This allows us to understand the effect of a given bit rate to the asymptotic error variance. For unstable systems, we have also introduced a dynamic scaling parameter for the quantizer to ensure the stability of the state estimation error dynamics.

References


10