

TECHNICAL RESEARCH REPORT

A NOTE ON THE MEAN-SQUARE QUANTIZATION ERROR

by Maben Rabi

TR 2005-110



The Center for Satellite and Hybrid Communication Networks is a NASA-sponsored Commercial Space Center also supported by the Department of Defense (DOD), industry, the State of Maryland, the University of Maryland and the Institute for Systems Research. This document is a technical report in the CSHCN series originating at the University of Maryland.

Web site <http://www.isr.umd.edu/CSHCN/>

A NOTE ON THE MEAN-SQUARE QUANTIZATION ERROR

Progress report for ENEE 699 (Spring 2000)

Maben Rabi

15th April, 2000

Abstract : This report summarizes the understanding I have gained in my studies for the independent studies course ENEE 699 up to 15th April, 2000. It includes the derivation of an extension of a result of Wong and Brockett on the behaviour of scalar quantizers.

1 Control under limited communication

Let,

$$\dot{x}(t) = f(x(t), u(t)) + n(t) \quad (1)$$

be a controlled dynamical system where the state $x(\cdot) \in R^n$, the control $u(\cdot) \in R^m$, and the noise process $n(\cdot) \in R^l$ are all defined on a suitable probability space (Ω, \mathcal{F}, P) and are assumed to be \mathcal{F} -measurable. We assume further that $f(\cdot, \cdot) : R^n \times R^m \rightarrow R^n$ satisfies regularity properties that ensure that (1) has smooth solutions on at least a finite time interval $[T_0, T_1]$ for a class \mathcal{U} , of admissible controls that includes the set of piecewise continuous (or more specifically, piecewise-constant) controls.

We are interested in generating feedback control laws in \mathcal{U} based on state observations $y(t) = x(t) + n'(t)$, where the observation noise process $n'(\cdot) \in R^l$ is also defined on (Ω, \mathcal{F}, P) and is \mathcal{F} -measurable. The observer is physically removed from the plant (1) and so the observation signals are sent to it over a digital communication link of finite capacity (because of finite bandwidth as well as noise). Hence, the observation signals necessarily have to be sampled, quantized and coded for transmission over the digital communication link.

Considerations of the complexity of implementation lead us to settle for a uniform sampling rate and allocation of fixed word lengths for the transmitted code-words.

The optimal coding and decoding scheme would be in general time-varying and could need to use all of the past observations ($y(t)$ for the encoder and the received binary code-words for the decoder). Again, we opt for a sub-optimal but simpler scheme of coding and decoding that uses not the entire past history of observations but a finite-dimensional statistic of it. This is the notion of ‘Finitely recursive state estimation’ proposed in [Wong-Brockett,1]. The simplified system then takes the form:

$$X(i+1) = X(i) + F(X(i), U(i)) + N(i),$$

$$\begin{aligned}\hat{X}(i) &= G(X(i), N'(i), A(i)), \\ A(i+1) &= H(A(i), X(i), N'(i)), \\ U(i) &= J(i, \hat{X}(i)),\end{aligned}$$

where $X(i)$ and $U(i)$ are time-discretized versions of $x(\cdot)$ and $u(\cdot)$. $N(i)$ and $N'(i)$ are discrete-time noise processes derived from $n(\cdot)$ and $n'(\cdot)$. $\hat{X}(i) \in R^n$ is the sequence of state estimates made by the observer and $A(i) \in R^k$ represents the finite memory of the encoder-decoder scheme. All of the discrete-time variables declared above are defined on a modified probability space $(\Omega_\Delta, \mathcal{F}_\Delta, P_\Delta)$. The functions $G : R^n \times R^{l'} \times R^k \rightarrow R^n$ and $H : R^k \times R^n \times R^{l'} \rightarrow R^k$ are maps representing the state estimation process. Hidden in G and J , are the time-delays due to the finite bandwidth of the link.

An object that we need to keep track of is $\{E[|\hat{X}(i) - X(i)|^2]\}$, the sequence of state-estimation error variances. We would desire the convergence to zero or at least the boundedness of this sequence. The next section deals with the study of when this is possible.

2 The quality of state estimation and quantization

In the setup of [Mitter-Borkar], the system dynamics is linear *i.e.* $F(x(t), u(t)) = Ax(t) + Bu(t)$. The sequence of error covariance matrices R_k is kept bounded by assuming that the singular values of the matrix A are less than 1 in magnitude (a condition stronger than the schur-stability of A). [Wong-Brockett,1] study the convergence behaviour of $\{E[|\hat{X}(i) - X(i)|^2]\}$ for some special cases. The main component of their analysis is the derivation of some explicit inequalities governing the error-variance of a single step of encoder-decoder operation. In the remainder of this section, we will study a key equation that leads to the proof of these inequalities.

For simplicity, we treat (as in [Wong-Brockett,1]) the case of scalar quantization. Let x be a real valued random variable with a probability density function $p(x)$. Assume

that $E[x] = \mu < \infty$ and $E[(x - \mu)^2] = \sigma^2 < \infty$. A finite code-book quantizer is one that partitions the real line into a finite collection of sets $S = \{S_i\}$, members of which are mutually disjoint and which together cover the entire real line (or at least the support of $p(x)$). To each S_i , the quantizer assigns a codeword that represents the quantized value $q(x)$ of x when $x \in S_i$. The distortion measure of the quantizer is the expectation value of a given distance function $d(q(x), x)$. We seek a quantizer that minimises the squared error distortion function : $E[(q(x) - x)^2]$. It is proved in [Gersho-Gray] that:

- (i) If the set of values of the quantized levels is prescribed, the optimal partition sets S_i are intervals each containing one quantization level, and
- (ii) If the partition sets are pre-specified, the optimal quantization levels are the conditional means : $E[x|x \in S_i]$.

Given partition sets, we can find an expression for the minimum variance of the quantization error. Such a result is presented in [Wong-Brockett,1] with a small mistake in notation. We reproduce it in the form of a lemma . Let $p_i = \text{prob}\{x \in S_i\}$, $\mu_i = E[x|x \in S_i]$, and $\sigma_i^2 = E[(x - \mu_i)^2|x \in S_i]$.

Lemma 1 *The minimum variance of the quantization error is*

$$E[\epsilon^2] = \sigma^2 - \sum_{i=1}^{|S|} p_i (\mu_i - \mu)^2 \quad (2)$$

Proof:

$$\begin{aligned} E[\epsilon^2] &= \sum_{i=1}^{|S|} p_i E \left[(x - E[x|x \in S_i])^2 \middle| x \in S_i \right], \\ &= E[x^2] - \sum_{i=1}^{|S|} p_i E[x|x \in S_i]^2, \\ &= \sigma^2 + \mu^2 - \sum_{i=1}^{|S|} p_i \mu_i^2 \end{aligned}$$

$$\begin{aligned}
&= \sigma^2 - \left(\sum_{i=1}^{|S|} p_i \mu_i^2 - 2\mu^2 + \mu^2 \right), \\
&= \sigma^2 - \left(\sum_{i=1}^{|S|} p_i \mu_i^2 - 2\mu \cdot \mu + \mu^2 \cdot 1 \right), \\
&= \sigma^2 - \left(\sum_{i=1}^{|S|} p_i \mu_i^2 - 2\mu \sum_{i=1}^{|S|} p_i \mu_i + \mu^2 \sum_{i=1}^{|S|} p_i \right), \\
&= \sigma^2 - \sum_{i=1}^{|S|} (p_i \mu_i^2 - 2\mu p_i \mu_i + \mu^2 p_i) \\
&= \sigma^2 - \sum_{i=1}^{|S|} p_i (\mu_i - \mu)^2.
\end{aligned}$$

□

The above lemma tells us that the variance of the quantization error is always less than or equal to that of x itself. For specific choices of the density function $p(x)$, we can calculate the minimum error variance as $\rho(S)\sigma^2$ with $0 < \rho(S) < 1$. We list some examples below:

(i) Uniform density function

Let

$$p(x) = \begin{cases} \frac{1}{b-a} & a < x \leq b, \\ 0 & \text{elsewhere,} \end{cases}$$

and

$$|S| = n, S_i = \left(a + (i-1)\frac{b-a}{n}, a + i\frac{b-a}{n} \right]$$

Then,

$$\mu = \frac{a+b}{2}, \sigma^2 = \frac{(b-a)^2}{12}, E[\epsilon^2] = \frac{(b-a)^2}{12n^2}, \rho(S) = 1/n^2.$$

(ii) Gaussian density function with two quantization levels

Let

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

and

$$|S| = 2, S_1 = (-\infty, 0], S_2 = (0, +\infty)$$

Then,

$$E[\epsilon^2] = \frac{\sigma^2(\pi-2)}{\pi}, \rho(S) = \frac{\pi-2}{\pi}.$$

(iii) Gaussian density function with n quantization levels (From[Gray])

Let

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$$

Then, there exists a n -element $S, n \gg 1$ such that:(Bennet's formula)

$$E[\epsilon^2] \approx \frac{\pi\sqrt{3}}{2n^2}\sigma^2, \quad \rho(S) \approx \frac{\pi\sqrt{3}}{2n^2}$$

For any density $p(x)$ with a finite support interval $[a, b]$, it is easy to find a partition S such that $\rho(S) < 1$. The equi-spaced partition into n sets with $n = \lceil \frac{b-a}{\sigma} \rceil$ for instance. Theorem 1 of [Wong-Brockett,1] states that if on N intervals, $p(x)$ (not necessarily of finite support) is either non-decreasing or concave and non-increasing, there exists a n -level quantizer ($n = \log_2 \lceil (N + 1) \rceil$) that satisfies

$$E[\epsilon^2] < \frac{3}{4}\sigma^2$$

Though the class of density functions that satisfy this theorem is broad, it excludes common choices such as the Gaussian density function. Also, it is not a high resolution result in that, $E[\epsilon^2]$ does not tend to 0 as $n \uparrow \infty$. In the next section, we state and prove a new result that fills both of these gaps.

3 A result on scalar quantization

The proof of the theorem of [Wong-Brockett,1] used some inequalities that go in a direction inverse to that of the Cauchy-Schwartz inequality. It is on examining this class of inequalities (as in [Diaz-Metcalf]) that I was led to think of the following result:

Theorem 1 *If $p(x)$ is a piece-wise continuous probability density function for the real-valued random variable x such that (i) $E[x] = \mu < \infty$, (ii) $E[(x - \mu)^2] = \sigma^2 < \infty$, and (iii) $p(x) \leq M \forall x \in R$ then, given any $\rho > 0$, it is possible to find a finite code-book quantizer whose quantization error variance is not greater than $\rho\sigma^2$.*

Proof: We carry out the proof in four steps.

Step 1: Bounding the contribution of the overload region

The function $f : R \rightarrow [0, \sigma^2]$ defined by $\phi(t) = \int_{-\infty}^t (x - \mu)^2 p(x) dx$ is continuous and non-decreasing. Also,

$$\lim_{t \rightarrow \infty} \phi(t) = \sigma^2.$$

Hence, given any γ such that $0 < \gamma < 1/2$, it is possible to find finite L, U such that $L < U$ and $\int_{-\infty}^L (x - \mu)^2 p(x) dx = \gamma\sigma^2$ and $\int_U^{\infty} (x - \mu)^2 p(x) dx = \sigma^2 - \int_{-\infty}^U (x - \mu)^2 p(x) dx = \gamma\sigma^2$. The set $(-\infty, L] \cup (U, \infty)$ is the overload region for our proposed quantizer. We call the granular region (viz. $(L, U]$) as $G(L, U)$.

Step 2: Calculating $E[\epsilon^2]$ on the granular region

We arrange matters so as to be able to use the Pólya-Szegö inequality [Diaz-Metcalf]

which states that if on the finite interval $a \leq x \leq b$, $f(x)$ and $g(x)$ are two Riemann integrable functions such that

$$0 < m_1 \leq f(x) \leq M_1 \quad (3)$$

$$0 < m_2 \leq g(x) \leq M_2, \quad (4)$$

then,

$$\left(\int_a^b f^2(x) dx \right) \left(\int_a^b g^2(x) dx \right) \leq \frac{(M_1 M_2 + m_1 m_2)^2}{4m_1 m_2 M_1 M_2} \left(\int_a^b f(x) g(x) dx \right)^2 \quad (5)$$

For any $\delta > 0$, define

$$G_\delta(L, U) = \{x : x \in G(L, U), p(x) \leq \delta^2\} \cup \{x : x \in G(L, U), |x - \mu| \leq \delta\}.$$

Also define $G^c(L, U) = G(L, U) - G_\delta(L, U)$. Given $\gamma' > 0$, we can choose δ such that

$$\int_{G_\delta(L, U)} (x - \mu)^2 p(x) dx < \gamma' \sigma^2 \quad (6)$$

For, if $K = \max(|L - \mu|, |U - \mu|)$, we have

$$\int_{G_\delta(L, U)} (x - \mu)^2 p(x) dx \leq K^2 \delta^2 (U - L) + \delta^2 M (U - L).$$

We can choose

$$\delta < \sqrt{\frac{\gamma' \sigma^2}{(K^2 + M)(U - L)}}$$

to make (6) true.

The set $G^c(L, U)$ is non-empty and has finite, non-zero length because

$$\int_{G^c(L, U)} (x - \mu)^2 p(x) dx \geq \sigma^2 (1 - 2\gamma - \gamma') \quad (7)$$

On $G^c(L, U)$, the following are true:

$$(i) \quad 0 < \delta^2 \leq |x - \mu| \sqrt{p(x)} \leq K \sqrt{M} \quad (8)$$

$$(ii) \quad 0 < \delta \leq \sqrt{p(x)} \leq \sqrt{M} \quad (9)$$

and from (i) and (ii) above, and (4), for any measurable subset Ω of $G^c(L, U)$,

$$\left(\int_\Omega |x - \mu|^2 p(x) dx \right) \left(\int_\Omega p(x) dx \right) \leq \frac{(KM + \delta^3)^2}{4\delta^3 KM} \left(\int_\Omega |x - \mu| p(x) dx \right)^2 \quad (10)$$

Step 3: Writing down the high resolution version of (7)

We now describe a partition of $G^c(L, U)$. For positive integers n_1 and n_2 , let

$$\alpha(n_1) = \left(\frac{K\sqrt{M}}{\delta^2} \right)^{1/n_1}, \quad \beta(n_2) = \left(\frac{\sqrt{M}}{\delta} \right)^{1/n_2}.$$

For $i = 1, \dots, n_1; j = 1, \dots, n_2$, define

$$A_{ij} = \{x : x \in G^c(L, U), \delta^2 \alpha^{i-1}(n_1) \leq |x - \mu| \sqrt{p(x)} \leq \delta^2 \alpha^i(n_1)\} \\ \cap \{x : x \in G^c(L, U), \delta \beta^{j-1}(n_2) \leq \sqrt{p(x)} \leq \delta \beta^j(n_2)\}$$

By (4),

$$\left(\int_{A_{ij}} |x - \mu|^2 p(x) dx \right) \left(\int_{A_{ij}} p(x) dx \right) \leq \frac{(\alpha(n_1)\beta(n_2) + 1)^2}{4\alpha(n_1)\beta(n_2)} \left(\int_{A_{ij}} |x - \mu| p(x) dx \right)^2 \quad (11)$$

The utility of (8) comes from the fact that if $\Delta = \alpha(n_1)\beta(n_2) - 1$,

$$\left| \frac{(\alpha(n_1)\beta(n_2) + 1)^2}{4\alpha(n_1)\beta(n_2)} - 1 \right| = \frac{(\alpha(n_1)\beta(n_2) + 1)^2}{4\alpha(n_1)\beta(n_2)} - 1 = \\ \frac{(\alpha(n_1)\beta(n_2) - 1)^2}{4\alpha(n_1)\beta(n_2)} = \frac{\Delta^2}{4(1 + \Delta)} \leq \frac{\Delta^2}{4}.$$

so that given $\gamma'' > 0$, we can choose n_1, n_2 such that $\Delta = \alpha(n_1)\beta(n_2) - 1$ is small enough for

$$\left| \frac{(\alpha(n_1)\beta(n_2) + 1)^2}{4\alpha(n_1)\beta(n_2)} - 1 \right| \leq \gamma''$$

to be true. Hence, we can choose n_1, n_2 such that,

$$\left(\int_{A_{ij}} |x - \mu| p(x) dx \right)^2 \geq \frac{1}{1 + \gamma''} \left(\int_{A_{ij}} |x - \mu|^2 p(x) dx \right) \left(\int_{A_{ij}} p(x) dx \right) \quad (12)$$

Step 4: Specifying the quantizer partition cells

Let $A_{ij}^+ = A_{ij} \cap \{x : x > \mu\}$ and $A_{ij}^- = A_{ij} \cap \{x : x \leq \mu\}$. Then on A_{ij}^+ , $|x - \mu| = x - \mu$ and on A_{ij}^- , $|x - \mu| = -(x - \mu)$. Now, (9) is brought to the form :

$$\left(\int_{A_{ij}^+} (x - \mu) p(x) dx \right)^2 \geq \frac{1}{1 + \gamma''} \left(\int_{A_{ij}^+} (x - \mu)^2 p(x) dx \right) \left(\int_{A_{ij}^+} p(x) dx \right) \quad (13)$$

$$\left(\int_{A_{ij}^-} (x - \mu)p(x)dx \right)^2 \geq \frac{1}{1 + \gamma''} \left(\int_{A_{ij}^-} (x - \mu)^2 p(x)dx \right) \left(\int_{A_{ij}^-} p(x)dx \right) \quad (14)$$

We now specify the quantizer completely:

$$q(x) = \begin{cases} E[x|x \in A_{ij}^+] & \forall x \in A_{ij}^+ \ i = 1, \dots, n_1; j = 1, \dots, n_2 \\ E[x|x \in A_{ij}^-] & \forall x \in A_{ij}^- \ i = 1, \dots, n_1; j = 1, \dots, n_2 \\ \frac{U+L}{2} & \forall x \in G_\delta(L, U) \\ L & \forall x \in (-\infty, L] \\ U & \forall x \in (U, \infty) \end{cases} \quad (15)$$

Note that the partition cells are not necessarily connected sets. From (3),(6),(10) and (11) :

$$E[\epsilon^2] \leq \sigma^2 - \sum_{\substack{i=1, \dots, n_1 \\ j=1, \dots, n_2}} \frac{\left(\int_{A_{ij}^+} (x - \mu)p(x)dx \right)^2}{\left(\int_{A_{ij}^+} p(x)dx \right)} - \sum_{\substack{i=1, \dots, n_1 \\ j=1, \dots, n_2}} \frac{\left(\int_{A_{ij}^-} (x - \mu)p(x)dx \right)^2}{\left(\int_{A_{ij}^-} p(x)dx \right)}, \quad (16)$$

$$= \sigma^2 - \frac{1}{1 + \gamma''} \sigma^2 (1 - 2\gamma - \gamma'), \quad (17)$$

$$= \sigma^2 \left[1 - \frac{(1 - 2\gamma - \gamma')}{(1 + \gamma'')} \right], \quad (18)$$

which can be made as small as possible (hence less than $\rho\sigma^2$) by choosing appropriate γ, δ, n_1, n_2 . Note that the error variance is further reduced when we split those cells A_{ij}^+, A_{ij}^- that are not connected sets, into their connected components. \square

4 Conclusions and future work

The chief merit of the theorem of the previous section is that it generalizes Bennet's approximate formula (example (iii) of section 2) and theorem 1 of [Wong-Brockett,1] to any piece-wise continuous density function that is bounded and has finite mean and variance. Also, it is a high resolution result in that, it says that using a large enough number of code-words, we can make the quantization error variance as small as desired. While it gives a constructive procedure to design a finite code-book quantizer that meets a prescribed distortion bound, the resulting quantizer is not necessarily the one that uses the minimum size code-book for that bound. Using this result, one can try to see whether the strong assumption on the stability of the uncontrolled dynamics of the system in [Mitter-Borkar] can be replaced by one on the

number of quantization levels and the sampling frequency. One can also try to work out the details that show that the notion of ‘Containability’ of [Wong-Brockett,2] reduces to that of ‘Stability’ (in the usual sense) in the case of infinite capacity of the communication link. I would like to read the paper of Williamson and the monograph of Moroney to get an idea of how the implementation side of these issues are dealt with. Another direction is to address a problem mentioned in [Zhang-Berger] : To find the rate-constrained analogue of the Cramer-Rao inequality.

References

- [Conway-Sloane] J.H CONWAY, N.J.A. SLOANE, ‘Voronoi regions of lattices, second moments of polytopes and quantization’, IEEE transactions on Information theory, vol:28, no:2, pp:211-226 (March 1982).
- [Curry] R.E. CURRY, Estimation and control with quantized measurements, MIT Press (1970).
- [Diaz-Metcalf] J.B. DIAZ, F.T. METCALF, ‘Complementary inequalities II: inequalities complementary to the Buniakowski-Schwartz inequality for integrals’, Journal of Mathematical analysis and applications, vol:9, pp:278-293 (1964).
- [Gray] R.M. GRAY, ‘Quantization’, IEEE transactions on Information theory, vol:44, no:6, pp:2325-2383 (October 1998).
- [Gersho-Gray] A. GERSHO, R.M. GRAY, Vector quantization and signal compression, Kluwer academic publishers (1992).
- [Mitrinović] D.S. MITRINOVIĆ, Analytic inequalities, Springer-Verlag (1970).
- [Mitter-Borkar] S.K. MITTER, V.S. BORKAR, ‘LQG control with communication constraints’, in Communications, computation, control and signal processing, edited by : A. Paulraj, V. Roychowdhury, C.D. Schaper. pp:365-373 (1997).
- [Moroney] P. MORONEY, Issues in the implementation of digital feedback compensators, MIT Press (1983).
- [Wong-Brockett,1] W.S. WONG, R.W. BROCKETT, ‘Systems with finite bandwidth constraints I: State estimation problems’, IEEE transactions on Automatic control, vol:42, no:9, pp:1294-1298 (September 1997).
- [Wong-Brockett,2] W.S. WONG, R.W. BROCKETT, ‘Systems with finite bandwidth constraints II: Stabilization with limited information feedback’, IEEE transactions on Automatic control, vol:44, no:5, pp:1049-1053 (May 1999).

[Williamson] D. WILLIAMSON, 'Finite wordlength design of digital Kalman filters for state estimation', IEEE transactions in Automatic control, vol:30, no:10, pp:930-939 (October 1985).

[Zhang-Berger] Z. ZHANG, T. BERGER, 'Estimation via compressed information', IEEE transactions on Information theory, vol:34, no:2, pp:198-211 (March 1988).