Optimal Quantization of Signals for System Identification

Koji Tsumura

May 12, 2009

Abstract: In this paper, we examine the optimal quantization of signals for system identification. We deal with memoryless quantization for the output signals and derive the optimal quantization schemes. The objective functions are the errors of least squares parameter estimation subject to a constraint on the number of subsections of the quantized signals or the expectation of the optimal code length for either high or low resolution. In the high-resolution case, the optimal quantizer is found by solving Euler–Lagrange’s equations and the solutions are simple functions of the probability densities of the regressor vector.

In order to clarify the minute structure of the quantization, the optimal quantizer in the low resolution case is found by solving recursively a minimization of a one-dimensional rational function. The solution has the property that it is coarse near the origin of its input and becomes dense away from the origin in the usual situation. Finally the required quantity of data to decrease the total parameter estimation error, caused by quantization and noise, is discussed.

Keywords: system identification, quantization, networked control, least squares method, FIR model, entropy

1 Introduction

The recent rapid improvement in the transmission capacity of computer networks has made long-distance automatic control more realistic and the necessity of understanding the effects of transmission limitations on the information in control systems has become more widely accepted. In particular, quantization of the signals to reduce the information content of the transmitted signals in control systems has been discussed actively by several control research groups during the last few years and interesting results have been achieved.

The problem of signal quantization has a long history going back to the 1940s, and is one of main themes in the area of information theory (e.g., see [13]). The problem is to attain low distortion between the original and the quantized signals subject to constraints on the amount of information. Naturally, the situations and objectives for data transmission and those for control systems are essentially different and the need for research on the latter case has been recognized. However, although elementary discussion in the control community dates from the 1970s (e.g., see [5]), rigorous analysis did not begin until the late 1980s. The main difficulty of quantization in control systems lies in their dynamics; the result by [6, 7] is recognized as a breakthrough, in which the behavior of control systems and their stability or state estimation are analyzed in detail. In the last few years, stabilization problems of quantized systems have been actively investigated in several different situations, e.g., [26, 27, 3, 16, 8, 17, 23, 18]. Of these, a logarithmic quantizer was shown to be coarsest, in some sense, to achieve a kind of asymptotic stability [8] and reveal the variations in the importance of signals, depending on their magnitudes and the directions in the signal space, from the viewpoint of system control.
With this background, our interests naturally shifted to the system identification problem; that is, what quantization scheme is optimal for system identification? We expect that the answer to this question will clarify the amount of information in the signals necessary for parameter estimation. Unfortunately however, compared to the research activity in the stabilization or estimation problem, the optimal quantization problem for system identification [10] has not been adequately considered. The main subject of this paper is to answer this fundamental question.

In this paper, we consider the optimal memoryless quantization problem of output signals that are used for parameter estimation. The identified system is a simple single input single output (SISO) finite impulse response (FIR) model, in order to reveal the essential properties of the optimal quantization in system identification and help intuitively understanding it. By optimality in this paper we mean the minimization of the variance of the parameter estimation error given by the least squares method with a constraint on the number of quantization steps or the expectation of the code length of the optimally coded quantized signals. We consider this problem for two cases: (1) high quantization resolution with weak assumptions on input, (2) low quantization resolution, however with some specific assumptions on input. The difficulty with the problem is in the complex correlation between the input signals and the quantization errors, and solving this is the key for the optimization problem.

In the high resolution case (Section 3), we introduce a key concept, the density of the number of quantized subsections, and by using calculus of variations, analytic solutions are derived subject to the constraint on the number of quantization steps or the optimal code length. The solutions are functions of the probability density of the input signals and we can rigorously calculate the profile of the density of the number of the optimally quantized subsections. Moreover, these results suggest several insights into system identification with finite information. We illustrate these facts for some cases and describe the complexity of the problem of system identification.

The results in Section 3 show that the quantization resolution around the origin of the signals relatively becomes coarse in usual cases. In order to clarify the minute structure of the quantization and complement the results in Section 3, we consider the low resolution case in Section 4. We give the optimal quantizer with a condition of uniform distribution of input signals. The optimal quantizer is given by minimizing a one-dimensional rational function recursively. In a special case, we show that the optimal quantization is not uniform and it is coarse near the origin of the quantized signals and becomes dense away from the origin. This fundamental property is opposite to the case of stabilization in [8] and reveals duality between system identification and stabilization.

Finally, in Section 5, we compare the effects of the resolution of quantization and the I/O data length. The results show that the former is more effective for decreasing quantization error in the estimated system parameters, on the other hand, the latter is more effective in reducing noise error. From this, there exists a trade-off between these two error terms subject to a constant amount of data and we can find an appropriate quantizer resolution to balance them by using the results in Section 5.

Note that the main purpose of this paper is to reveal the essential properties of the optimal quantization for system identification; therefore, the focus of this paper is on the analysis of this problem and not on practical system identification methods.

In this paper, most of the proofs of theorems, lemmas, or propositions are collected in the appendix for ease of understanding the main theme and the outline of this paper. Refer to these in Appendix A if necessary.
Notation:

\[ d_j: \text{ eq. (4) and (5)} \]
\[ E[x]: \text{ expectation of } x, \ E_\bullet[x]: \text{ eq. (48)} \]
\[ e(t) = y'(t) - y(t): \text{ quantization error at } t \]
\[ e(\tilde{\phi}_1(t)) = e(t): \text{ quantization error specified by } \tilde{\phi}_1 \]
\[ f(x): \text{ probability density of } x \]
\[ g(\bullet): \text{ eq. (25)} \]
\[ H(\bullet): \text{ entropy of } \bullet, \ H(\bullet, \bullet), H_\bullet(\bullet): \text{ eq. (38)} \]
\[ j: \text{ index of quantized subsections} \]
\[ M: \text{ number of quantization subsections} \]
\[ M': \text{ associate number of quantization subsections (53)} \]
\[ N: \text{ data length} \]
\[ n: \text{ order of FIR model} \]
\[ O(\bullet), o(\bullet): \text{ orders of } \bullet \text{ (Landau’s symbols)} \]
\[ P(\bullet): \text{ eq. (90)} \]

\[ r_j, r_j^\circ: \text{ ratio or optimal ratio of } d_j \text{ and } d_{j+1} \text{ (54)} \]
\[ S_j^\bullet: j\text{-th subsection on the space of } \bullet \]
\[ T: \text{ variable transformation matrix} \]
\[ V[x]: \text{ expectation of } \|x\|_2^2, \ V_\bullet[x]: \text{ eq. (82)} \]
\[ y(t) = \phi(t)\theta: \text{ output of FIR model at } t \]
\[ y_\circ(t): \text{ observed output (1)} \]
\[ \theta \in \mathbb{R}^n: \text{ parameter vector of FIR model} \]
\[ \phi(t): \text{ regressor vector eq. (1)} \]
\[ \hat{\phi}_1: 1\text{-st element of } \hat{\phi} \]
\[ \sigma(\hat{\phi}_1): \text{ eq. (27)} \]
\[ \bullet_i, (\bullet)_i: i\text{-th element of vector } \bullet \]
\[ \bullet': \text{ quantized number of } \bullet \]
\[ \bullet'_j: j\text{-th quantized number for } S_j^\bullet \]
\[ \hat{\bullet}: \text{ transformed vector or matrix of } \bullet \text{ by } T \]

2 Problem Formulation

The objective of this paper is to show the effect of I/O signal quantizers for parameter estimation error intuitively understandable form as possible. In general, the quantization error has a strong correlation with the original signal, therefore, analysis of the quantization problem in system identification in general model is difficult because several types of correlation are used for parameter estimation. In order to derive analytic and intuitively understandable results for the quantization problem in system identification, we should formulate the problem in feasible forms appropriately.

From the above observations, in this paper, we deal with a system identification problem by least square criterion for a simple discrete time SISO FIR model. The plant is:

\[ y_\circ(t) = q(y(t)) + w(t), \ y(t) = \phi(t)\theta, \]
\[ \phi(t) := \left[ u(t) \ u(t-1) \ \cdots \ u(t-n+1) \right], \ \theta := \left[ \theta_1 \ \theta_2 \ \cdots \ \theta_n \right]^T, \]
\[ y_\circ, \ y, \ w, \ u \in \mathbb{R}, \ \phi \in \mathbb{R}^{1 \times n}, \ \theta \in \mathbb{R}^{n \times 1}, \]

where \( w \) is random noise, \( q \) is the quantized original analogue output \( y \), \( y_\circ \) is the observed output, \( \phi \) is the regressor vector, \( \theta \) is a system parameter, \( n \) is the dimension of the FIR model, \( u \) is the input, and \( t \) is the time index.

We assume that \( u \) and \( w \) are independent. The input \( u \) and the associated regressor vector \( \phi \) are a realization of a stochastic process with a joint density function \( f(\phi_1, \phi_2, \ldots, \phi_n) \) of \( \phi_1, \phi_2, \ldots, \phi_n \), where \( \phi_i \) denotes the \( i \)-th element of \( \phi \). The class of \( f(\phi_1, \phi_2, \ldots, \phi_n) \) considered in this paper is described below.

**Note 2.1** We also consider noise to be

\[ y_\circ(t) = q(y(t) + w(t)) \]

in [24] (the long version of this paper). The result suggests that the noise when (2) increases the effect of quantization on the
magnitude of the parameter estimation error by approximately twice that of (1). From that result, it is enough to analyze the form of (1) in order to know the essential property of the optimal quantization. To avoid complicated notation and focus on the quantization effect for system identification, we treat the plant (1) in this paper.

The quantizer \( q \) is a memoryless symmetric type defined by:

\[
q(y) := y_{(j)} \quad \text{when} \quad y \in S^q_j
\]

\[
S^q_0 := \{ y = 0 \}, \quad S^q_j := \{ y : d_{j-1} < y \leq d_j \}, \quad j > 0, \quad S^q_{-j} := \{ y : d_j \leq y < d_{j+1} \}, \quad j < 0
\]

\[
d_0 = 0 < d_1 < d_2 \cdots, \quad d_{-1} = -d_1, \quad d_{-2} = -d_2, \quad \ldots
\]

where \( y_{(j)} \) is the assigned quantized value to the subsection \( S^q_j \). The quantizer \( q \) is symmetrical with respect to the origin, and hereinafter we may omit references on the negative subsections \( S^q_{-1}, S^q_{-2}, \ldots \) if they are obvious from the context. Note that a form \( S^q_0 = \{ y : -d_1 \leq y \leq d_1 \} \) is also possible for \( S^q_0 \), however it is clarified not to be optimal in Section 4 and without loss of generality, we consider the form of (4) hereafter.

Following the standard least squares method, we propose the estimated parameter \( \hat{\theta} \) with a sufficient length of I/O data, \( \{ u(t) \} \) and \( \{ y_o(t) \} \), as:

\[
\hat{\theta} = (U^T U)^{-1} U^T Y_o = (U^T U)^{-1} U^T (Y' + W) = (U^T U)^{-1} U^T (Y + E + W),
\]

where

\[
U := [ \phi(1) \phi(2) \cdots \phi(N) ]^T, \quad W := [ w(1) \ w(2) \ \ldots \ w(N) ]^T,
\]

\[
Y_o := [ y_o(1) \ y_o(2) \ \cdots \ y_o(N) ]^T, \quad Y := [ y(1) \ y(2) \ \cdots \ y(N) ]^T,
\]

\[
Y' := [ y'(1) \ y'(2) \ \cdots \ y'(N) ]^T, \quad y'(t) := q(y(t)),
\]

\[
e(t) := y'(t) - y(t).
\]

and \( N \) is the I/O data length. We call \( e \) as the quantization error between \( y' \) and \( y \). The estimated parameter \( \hat{\theta} \) can be also written as:

\[
\hat{\theta} = (U^T U)^{-1} U^T (U \theta + E + W) = \theta + \Delta E + \Delta W,
\]

\[
E := [ e(1) \ e(2) \ \cdots \ e(N) ]^T, \quad \Delta E := (U^T U)^{-1} U^T E, \quad \Delta W := (U^T U)^{-1} U^T W.
\]

This shows that the estimation error \( \hat{\theta} - \theta \) can be evaluated from the magnitudes of the quantization error term \( \Delta E \) and the noise error term \( \Delta W \).

In the quantization-free case, i.e. \( e = 0 \), (6) is the standard least squares estimation. When \( e \neq 0 \), (6) is still a realistically reasonable estimation subject to the minimization of

\[
E[\|\Delta E\|_2^2]
\]

because

\[
E[\|\hat{\theta} - \theta\|_2^2] = E[\|\Delta E + \Delta W\|_2^2] = E[\|\Delta E\|_2^2] + E[\|\Delta W\|_2^2].
\]
The reduction of the noise error term $\Delta W$ is the main theme of normal system identification and has been well investigated. On the other hand, although the quantization error term $\Delta E$ can be reduced, in general, when the resolution of quantizer becomes high, there exists a limitation in the reduction because of the constraint of the resolution of the quantizer and good quantizers for reducing $\Delta E$ are expected. Here we show an original quantization problem in this paper which is resolved into feasible ones in Section 3 and 4.

**Problem 2.1** Find an optimal quantizer $q(y)$:

\[
\min_{q} E[\|\Delta E\|^2] \\
\text{s.t. } E[\Delta E] = 0
\]  

under constraint on the quantization resolution.

Note that the latter condition is for bias-free of the estimated parameters.

**Note 2.2** In the field of information theory, the quantization problem is also one of the research themes and its objective is reducing the distortion between the original signal and the quantized signal subject to constraints on the information in the transmitted signals [1, 15, 11, 2, 9]. The constraint on the information in signals can be given by the number of the quantization steps or the mean code length of the associated code. The former is called “fixed-rate quantization” and the latter “variable-rate quantization”. In contrast, the purpose in system identification should be the reduction of the estimation error and this is the definitive difference.

In an ordinary probabilistic framework, a conventional, and reasonable, method to evaluate the noise error term $\Delta W$ is to show the convergence rate of:

\[
N(U^TU)^{-1} \xrightarrow{N \to \infty} \frac{1}{\sigma_u^2} I, \quad \frac{1}{N}U^TW \xrightarrow{N \to \infty} O,
\]

where $\sigma_u^2$ is the covariance of $u$, by using Slutsky’s theorem (see Appendix A), subject to an assumption of the mutual independence of the input signal $u$ and the noise $w$. This methodology is also basically applicable to the evaluation of $\Delta E$ in the probabilistic framework. However, different from the case of the noise error term, $u$ and $e$ are not independent in general and the evaluation of $U^TE$ is much more complicated. This means the problem seems to be a vector quantization on $U^TE$ with a complex multidimensional distribution. In general, multidimensional optimal quantization is known to be a difficult problem for analytical solution except in special cases.

Our idea to resolve the above difficulty is in showing that the original problem, i.e., minimizing the cost function on the magnitude of $\Delta E$, can be reduced to a feasible problem; “minimization of a functional of a weighted one-dimensional quantizer,” by following two steps: 1. finding an equivalent orthogonal quantization on the space of the regressor vector to the original quantization of the output signals, 2. reduction of the cost functions to a suitable form by using one of the base axes in the regressor vector space. Step 1 is described in this section and Step 2 is described in Section 3 and 4.

We define subsets $S_j^\phi$ of the regressor vector $\phi$ associated with the subsection $S_j^y$ by:

\[
S_j^\phi := \{ \phi : y = \phi \theta \in S_j^y \}.
\]
We also consider the following variable transformation:

\[ y = \phi \theta = \phi T^{-1} \theta = \tilde{\phi} \tilde{\theta}, \quad \tilde{\theta} := T^{-1} \theta = [\tilde{\theta}_1 \ O \ \tilde{\theta}_2 \ \cdots \ \tilde{\theta}_n] \]  

(12)

where \( T \) is an orthogonal matrix. Note that such \( T \) always exists for any \( \theta \). Then, \( S^\phi_j \) is represented by:

\[
S^\phi_j := \begin{cases} 
\{ \phi : \tilde{\phi}_1 \tilde{\theta}_1 \in (d_{j-1}, d_j) \}, & j > 0, \\
\{ \tilde{\phi}_1 = 0 \}, & j = 0, \\
\{ \phi : \tilde{\phi}_1 \tilde{\theta}_1 \in [d_j, d_{j+1}) \}, & j < 0.
\end{cases}
\]

We also define subsections on the space \( \tilde{\phi}_1 \):

\[
S^{\tilde{\phi}_1}_j := \begin{cases} 
\{ \tilde{\phi}_1 : \tilde{\phi}_1 \tilde{\theta}_1 \in (d_{j-1}, d_j) \}, & j > 0, \\
\{ \tilde{\phi}_1 = 0 \}, & j = 0, \\
\{ \tilde{\phi}_1 : \tilde{\phi}_1 \tilde{\theta}_1 \in [d_j, d_{j+1}) \}, & j < 0.
\end{cases}
\]

Then, subsections \( S^y_j, S^\phi_j, \) and \( S^{\tilde{\phi}_1}_j \) correspond to each other, and the probability distribution of \( y \) depends only on that of \( \tilde{\phi}_1 \). Therefore, the variable \( \tilde{\phi}_1 \) and its subsection \( S^{\tilde{\phi}_1}_j \) are convenient for analyzing the probability distribution of \( y \) and the error \( e \).

Fig. 1 and Fig. 2 are representations of the relationship between \( S^y_j, S^\phi_j, \) and \( S^{\tilde{\phi}_1}_j \), or \( y, \phi, \) and \( \tilde{\phi}_1 \).

Associated with \( T \), the quantization error term \( \Delta E \) and \( U \) are also transformed to:

\[ \Delta \tilde{E} := T^{-1} \Delta E, \quad \tilde{U} := UT \]  

(13)
and $\Delta \tilde{E}$ can be represented as:

$$
\Delta \tilde{E} = T^{-1} (U^T U)^{-1} U^T E = (\tilde{U}^T \tilde{U})^{-1} \tilde{U}^T E
$$

Note that $\|\Delta \tilde{E}\|_2^2 = \|\Delta E\|_2^2$ because $T$ is an orthogonal matrix. From the above, it is known that the quantizer can be considered to be an orthogonal and symmetric type along each axis $\hat{\phi}_i$ in the sense that each axis $\hat{\phi}_i$ is partitioned in the same rule (see Fig. 3).

In Sections 3 and 4, we first derive key lemmas, respectively, to show that the quantity $\|\Delta E\|_2^2 = \|\Delta \tilde{E}\|_2^2$ can be represented as a functional of the one-dimensional marginal density function $f(\hat{\phi}_1)$ and the quantizer on $\hat{\phi}_1$, subject to appropriate assumptions.

## 3 High Resolution Quantization

In this section, we derive optimal quantizers under considerably weak conditions on the probability densities $f(\phi)$ where the quantizers are assumed to be high resolution. At first, we show the following assumption:

**Assumption 3.1** The input $u$ and the density function $f(\phi)$ satisfy the following conditions:

1: $u(t), t = \ldots, 1, 2, \ldots$ are mutually independent.

2: $f(\phi)$ is a continuous function s.t. $f(\hat{\phi})$ satisfies:

$$
f(\hat{\phi}) = \delta_0 + \sum_i \delta_i(\hat{\phi}_i - \hat{\phi}_i^*) + \sum_{i,j} \delta_{ij}(\hat{\phi}_i - \hat{\phi}_i^*)(\hat{\phi}_j - \hat{\phi}_j^*) + O((\hat{\phi}_i - \hat{\phi}_i^*)(\hat{\phi}_j - \hat{\phi}_j^*)(\hat{\phi}_k - \hat{\phi}_k^*)), \quad |\delta_*| < \infty
$$

(15)

in the neighborhood of an arbitrary $\hat{\phi}^* = [\hat{\phi}_1^* \hat{\phi}_2^* \cdots \hat{\phi}_n^*] \in \{\hat{\phi}\}$.

These conditions are not strong in usual setting of system identification. In particular, the essence of (15) is for guaranteeing the continuity of $f(\phi)$ and it is usually satisfied; e.g., (15) is satisfied when $f(\phi)$ is a multidimensional normal distribution.

This technical condition is used in the proof of Lemma 3.1.

The first Assumption 3.1.1 gives the convergence of $\frac{1}{N} U^T U$ or $\frac{1}{N} \tilde{U}^T \tilde{U}$ to $\sigma_u^2 I$, where $\sigma_u^2$ is a covariance of $u$, at $N \to \infty$, and therefore,

$$
N\|\Delta E\|_2^2 = \lim_{N \to \infty} \text{tr} \left[ \frac{1}{N^2} U^T U U^T U \right] = \frac{1}{\sigma_u^2} \text{plim}_{N \to \infty} \left[ \frac{1}{N} E^T U U^T E \right]
$$

(16)

by Slutsky’s theorem (see Appendix A). Moreover, we get:

$$
\text{plim}_{N \to \infty} \left[ \frac{1}{N} E^T U U^T E \right] = \frac{1}{N} \text{Var} \left[ U^T E \right] = \frac{1}{N} \text{Var} \left[ \tilde{U}^T E \right]
$$

(17)
therefore,
\[ \| \Delta E \|^2 \sim \frac{1}{\sigma_u^4 N^2} \mathbb{V}[U^T E] \] (18)
at enough large \( N \). Then, it is reasonable to find an optimal quantizer that:

1) minimizes \( \mathbb{V}[U^T E] \left( = \mathbb{V}[\tilde{U}^T E] \right) \)

2) subject to constraints on the resolution of the quantizer, free of bias from the quantization error term, such as: \( \mathbb{E}[U^T E] = 0 \) (equivalently \( \mathbb{E}[\tilde{U}^T E] = 0 \)).

The minimization of \( \mathbb{V}[U^T E] \) in arbitrary resolution cases of the quantizer is too complex to expect meaningful results, however, it is possible to derive the analytic solution in *high resolution* as shown in the following of this section.

**Note 3.1** The multidimensional optimal quantization problem has been investigated (e.g., see [13, 12, 19, 9]) and the research focus is on the derivation of analytic solutions. In the general resolution case, it is known to be a difficult problem and limited cases have been solved. One of these is the case of one-dimensional quantization and another is the asymptotic case when the resolution of quantizers is sufficiently high. Note that cost functions are \( \mathbb{E}[\| X - q(X) \|^r] \) in these studies. However, we consider the cost function \( \mathbb{E}[\| U^T E \|^2] \) in this paper, which originates in system identification parameter estimation. The evaluation of the latter is much more complicated because it contains many correlations of variables and resolving this difficulty is one of main themes of this paper (Note that the latter is not simple weighted square-error distortion because of the correlation between \( \hat{\phi}_1 \) and \( e = \hat{\phi}_1 \hat{\theta}_1 - q(\hat{\phi}_1 \hat{\theta}_1) \)). The key lemmas (Lemma 3.1 and 4.1) show that this quantity can be represented as a functional of one-dimensional functions with one-dimensional quantization rules under appropriate assumptions and, by using them, we can find the optimal quantizers.

On the above minimization problem, the bias-free condition \( \mathbb{E}[U^T E] = 0 \) is equivalent to \( \mathbb{E}[\tilde{U}^T E] = 0 \) from the relation \( \tilde{U}^T E = T^T U^T E \), where \( T \) is nonsingular and orthogonal. From (14), this condition is equivalent to
\[
\mathbb{E} \left[ \sum_{i=1}^{N} \tilde{\phi}_k(t)e(t) \right] = N \cdot \mathbb{E} \left[ \tilde{\phi}_k \cdot e(\tilde{\phi}_1) \right] = N \int \tilde{\phi}_k e(\tilde{\phi}_1) f(\tilde{\phi}_1, \tilde{\phi}_k) d\tilde{\phi}_1 d\tilde{\phi}_k = 0
\]
for \( k = 2, 3, \ldots, n \) and
\[
\mathbb{E} \left[ \sum_{i=1}^{N} \tilde{\phi}_1(t)e(t) \right] = N \cdot \mathbb{E} \left[ \tilde{\phi}_1 \cdot e(\tilde{\phi}_1) \right] = N \int \tilde{\phi}_1 e(\tilde{\phi}_1) f(\tilde{\phi}_1) d\tilde{\phi}_1 = 0
\]
for \( k = 1 \). Note that we use the notation \( e(\tilde{\phi}_1(t)) \) when we intend to specify that \( e(t) \) is a function of \( \tilde{\phi}_1(t) \), which can be seen from (14). The notation \( f(\tilde{\phi}_1) \) represents a marginal density function:
\[
f(\tilde{\phi}_1) := \int f(\tilde{\phi}_1, \tilde{\phi}_2, \ldots, \tilde{\phi}_n) d\tilde{\phi}_2 \ldots d\tilde{\phi}_n.
\]
The notations \( f(\tilde{\phi}_1, \tilde{\phi}_2), f(\tilde{\phi}_1, \tilde{\phi}_2, \tilde{\phi}_3), \ldots \) are similarly defined.

With the continuity condition of \( f(\phi) \) in Assumptions 3.1.2, (19) and (20), i.e., the bias-free condition \( \mathbb{E}[U^T E] = 0 \left( \mathbb{E}[\tilde{U}^T E] = 0 \right) \), are asymptotically satisfied as the widths of the quantization steps tend to 0 with the setting of \( y(\delta) \) at
the center of the quantization subsections. On the other hand, for the cost function $V[U^T E] = V[\hat{U}^T E]$, which can be represented by 

$$V[U^T E] = \sum_{k=1}^{N} E \left[ \left( \sum_{t=1}^{N} \phi_k(t)e(t) \right)^2 \right] = \sum_{k=1}^{N} E \left[ \left( \sum_{t=1}^{N} \tilde{\phi}_k(t)e(\tilde{\phi}_1(t)) \right)^2 \right],$$

we derive the following key lemma.

**Lemma 3.1** Assume that $f(\hat{\phi})$ satisfies (15) in Assumption 3.1.2. Then, 

$$E \left[ \left( \sum_{t=1}^{N} \tilde{\phi}_k(t)e(\tilde{\phi}_1(t)) \right)^2 \right] \xrightarrow{\Delta y_{\text{max}} \to 0} N \Delta y_{\text{max}} e^2(\hat{\phi}_1),$$

where $\Delta y_{\text{max}}$ is the maximum width of the subsections $S^y$ of the quantizer defined by $\Delta y := \max |d_{j+1} - d_j|$. The proof of this lemma is given in Appendix A.

From this lemma, the cost function $V[U^T E] = V[\hat{U}^T E]$ can be approximated by:

$$V \left[ U^T E \right] \xrightarrow{\Delta y_{\text{max}} \to 0} N \sum_{k=1}^{N} E[\hat{\phi}_k^2 e^2(\hat{\phi}_1)] = N \sum_{k=1}^{N} \int \hat{\phi}_k^2 e^2(\hat{\phi}_1) f(\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_n) d\hat{\phi}_1 d\hat{\phi}_2 \cdots d\hat{\phi}_n$$

$$= N \int \left( \int \sum_{k=1}^{N} \hat{\phi}_k^2 f(\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_n) d\hat{\phi}_2 \cdots d\hat{\phi}_n \right) e^2(\hat{\phi}_1) d\hat{\phi}_1. \tag{24}$$

in the high resolution case. Therefore, the focus of the problem is on the calculation of the r.h.s. of (24) for general $f(\phi)$ and its minimization. A key concept in solving this problem is the introduction of the following quantity in the distribution of quantization subsections, which is a reasonable concept in the high resolution case.

**Definition 3.1** The quantity $g(\hat{\phi}_1)$, which satisfies 

$$g(\hat{\phi}_1)d\hat{\phi}_1 = \text{number of quantized subsections in } d\hat{\phi}_1,$$ 

is called the density of the number of quantized subsections.

This quantity is the same as that introduced in [1, 15] and from this definition, $g(\hat{\phi}_1)^{-1}$ represents the width of the quantization step at $\hat{\phi}_1$.

We also assume a form of smoothness of $f(\phi)$ and $g(\hat{\phi}_1)$ in the following.

**Assumption 3.2** The density function $f(\phi)$ and $g(\hat{\phi}_1)$ satisfy the following conditions:

1: $f(\phi)$ is a continuous function s.t.

$$\frac{d(\sigma^2(\hat{\phi}_1)f(\hat{\phi}_1))}{d\hat{\phi}_1} < \infty, \tag{26}$$

$$\sigma(\hat{\phi}_1) := \left( f(\hat{\phi}_1)^{-1} \int \left( \sum_{k=1}^{N} \hat{\phi}_k^2 \right) f(\hat{\phi}_1, \ldots, \hat{\phi}_n) d\hat{\phi}_2 \cdots d\hat{\phi}_n \right)^{\frac{1}{2}}, \tag{27}$$

where $f(\hat{\phi}_1)$ is the marginal density function on the space of $\hat{\phi}_1$ defined by (21).
the resolution of quantizer is sufficiently high and the density \( g(\tilde{\phi}_1) \) satisfies:

\[
\frac{dg(\tilde{\phi}_1)}{d\tilde{\phi}_1} < \infty.
\]

**Note 3.2** The essence of Assumption 3.2 is the smoothness of \( f(\tilde{\phi}_1) \) and \( g(\tilde{\phi}_1) \) such as they guarantee the approximation of (24) in the following. Assumption 3.2.1 describes a form of the continuity of \( f(\tilde{\phi}) \) or \( f(\hat{\phi}_1) \) and it is not a strong assumption in the usual situation of system identification; e.g., \( f(\tilde{\phi}) \) or \( f(\hat{\phi}) \) in \( C^1 \) is enough and it is satisfied when they are multidimensional normal distributions. Assumption 3.2.2 also describes a form of the continuity of the quantizer and \( g(\tilde{\phi}_1) \) or \( g(y) \in C^2 \) is enough. Such technical conditions come from our intention to make the necessary conditions for deriving (28) weak as possible.

With Assumption 3.2.2, we can select a value \( g_j^{-1} \sim g(\hat{\phi}_1)^{-1} \) for the subsection \( S_{\tilde{\phi}_j} \) that satisfies \( g_j^{-1} = |S_{\tilde{\phi}_j}| \). Moreover, with \( \sigma(\tilde{\phi}_1) \) of \( f(\hat{\phi}) \) at \( \hat{\phi}_1 \) defined in (27), Assumption 3.2.1–2, and \( \Delta \tilde{\phi} := \max_j \hat{\theta}_1^{-1}|d_{j+1} - d_j| \), for the objective function (24), we calculate the following directly:

\[
\frac{(24)}{N} = \int (\tilde{\theta}^2(\hat{\phi}_1) - 2 \sigma^2(\hat{\phi}_1)) f(\hat{\phi}_1) d\hat{\phi}_1 + O(\Delta \tilde{\phi}).
\]

See Appendix A for the derivation of (28). From this,

\[
\hat{\theta}_1^2 \int \frac{1}{12} g(\hat{\phi}_1)^{-2} \sigma^2(\hat{\phi}_1) f(\hat{\phi}_1) d\hat{\phi}_1
\]

is considered to be a reasonable cost function when Assumption 3.1 and 3.2 are satisfied.

In the following we assume Assumption 3.1 and Assumption 3.2 and give the optimal quantizers, which minimize (29), subject to a constraint on the number of quantization steps (Section 3.1) or on the expectation of the code length, where the quantized data is optimally encoded (Section 3.2). The former case is referred to as “fixed-rate quantization” because it is identical to a “fixed-code length” case; the latter case is referred to as “variable-rate quantization” and the code length is not fixed.

### 3.1 Fixed-rate quantization

From the previous derivation, the original optimization problem of (24) can be replaced by the minimization of (29) in \( N \to \infty \) and the high resolution case:

**Problem 3.1** Find

\[
g_k(\tilde{\phi}_1) := \arg \min_g \int \mathcal{F}(g(\tilde{\phi}_1)) d\tilde{\phi}_1
\]

s.t.

\[
\int g(\tilde{\phi}_1) d\tilde{\phi}_1 = M,
\]

where

\[
\mathcal{F}(g(\tilde{\phi}_1)) := \frac{1}{12} \hat{\theta}_1^2 g(\hat{\phi}_1)^{-2} \sigma^2(\hat{\phi}_1) f(\hat{\phi}_1).
\]
Theorem 3.1 The solution of (30) is:

\[ g_t(\tilde{\phi}_1) = K \sigma^2(\tilde{\phi}_1) f^{-\frac{1}{2}}(\tilde{\phi}_1) \quad (33) \]
\[ K = D^{-1} M \quad (34) \]
\[ D = \int \sigma^2(\tilde{\phi}_1) f^{-\frac{1}{2}}(\tilde{\phi}_1) d\tilde{\phi}_1. \quad (35) \]

Moreover, the optimized value is given by:

\[ \int \mathcal{F}(g_t(\tilde{\phi}_1)) d\tilde{\phi}_1 = \frac{1}{12} \tilde{\sigma}_0^2 D^3 M^{-2}. \quad (36) \]

The minimization problem can be rigorously solved by applying the calculus of variations. See Appendix A for the proof.

From Theorem 3.1, the asymptotic optimal quantization at high resolution is readily calculated analytically, or numerically, if the marginal density functions \( f(\tilde{\phi}_1) \) are known.

Note 3.3 The optimal quantization scheme on \( y \) (call it as \( g_t(y) \)) is also given by using the above results. With the relation \( y = \tilde{\phi}_1 \tilde{\theta}_1 \) and the fact that the optimal \( g_t(\tilde{\phi}_1) \) is given only by \( f(\tilde{\phi}_1) \), \( g_t(y) \) on \( y \) is a simple scaling of \( g_t(\tilde{\phi}_1) \). Therefore, \( g_t(y) \) on \( y \) is given by: (i) using the knowledge of \( \tilde{\theta}_1 \) and \( g_t(\tilde{\phi}_1) \), or (ii) \( f(y) \) on \( y \) such as \( g_t(y) = K' \sigma^{\frac{1}{2}}(y) f^{\frac{1}{2}}(y) \), where \( f(y) \) is obtained by the observation of the output data \( \{y(t)\} \). The situation (i) is a standard problem setting of control systems under limitation of channel capacity, where the quantizer (encoder) is supposed that it can fully utilize information on systems in order to optimally compress the data. The situation (ii) is also a natural problem setting.

Example 3.1 When \( f(\tilde{\phi}) \) is a multidimensional normal distribution:

\[ f(\tilde{\phi}_1, \tilde{\phi}_2, \ldots, \tilde{\phi}_n) = \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{\det \Gamma}} \exp \left( -\frac{1}{2} \tilde{\phi}^T \Gamma^{-1} \tilde{\phi} \right), \quad \Gamma = \text{diag}(\sigma_o, \sigma_o, \ldots, \sigma_o), \]

where \( \Gamma \) is a covariance matrix of \( \tilde{\phi} \), then

\[ \sigma^2(\tilde{\phi}_1) = \tilde{\sigma}_1^2 + (n-1)\sigma_o^2. \]

For simplicity, in the case that the order \( n \) of the FIR model is sufficiently large,

\[ \sigma^2(\tilde{\phi}_1) f(\tilde{\phi}_1) \sim n \sigma_o^2 f(\tilde{\phi}_1). \]

Therefore:

\[ D \sim n^2 \tilde{\sigma}_0^2 \int f^{\frac{1}{2}}(\tilde{\phi}_1) d\tilde{\phi}_1, \quad g_t(\tilde{\phi}_1) \sim M \left( \int f^{\frac{1}{2}}(\tilde{\phi}_1) d\tilde{\phi}_1 \right)^{-1} f^{\frac{1}{2}}(\tilde{\phi}_1), \]

\[ \int \mathcal{F}(g_t(\tilde{\phi}_1)) d\tilde{\phi}_1 \sim \frac{1}{12} \tilde{\sigma}_0^2 \left( \int f^{\frac{1}{2}}(\tilde{\phi}_1) d\tilde{\phi}_1 \right)^3 n\sigma_o^2 M^{-2} = \frac{1}{12} \tilde{\sigma}_0^2 n \sqrt{3} \pi n\sigma_o^2 M^{-2} \sim 0.8658 \tilde{\sigma}_0^2 n \sigma_o^2 M^{-2}. \quad (37) \]

\[ \diamond \]

Example 3.2 Here we consider another simple case \( n = 1 \), where the cost function becomes

\[ \mathcal{V} [U^T E] = N \int \tilde{\sigma}_1^2 e^2 (\tilde{\phi}_1) f(\tilde{\phi}_1) d\tilde{\phi}_1. \]
Then, the optimal quantization $g_t(\hat{\phi}_1)$ for this is given by

$$g_t(\hat{\phi}_1) = K_{\hat{\phi}_1} f^*(\hat{\phi}_1), \quad K = D^{-1} M, \quad D = \int_{\hat{\phi}_1} f^*(\hat{\phi}_1) d\hat{\phi}_1.$$ 

We illustrate $g_t(\hat{\phi}_1)$ for the cases where $\sigma^2(\hat{\phi}_1) = \tilde{\phi}_1^2 + \sigma_o^2$ and $f(\hat{\phi}_1)$ is the uniform distribution, normal distribution, or power law as follows.

Fig. 4 is the case that $f(\hat{\phi}_1)$ is the uniform distribution. From the figure, we observe that the optimal quantization is coarse near the origin of $\hat{\phi}_1$ and dense near the boundary of the domain of $\hat{\phi}_1$. Theorem 3.1 shows that the increasing rate of resolution with enough large $\tilde{\phi}_1$ is about $\tilde{\phi}_1^2$.

When $f(\hat{\phi}_1)$ is the normal distribution, the profile of the density $f(\hat{\phi}_1)$ near the origin is flat; therefore, the optimal quantizer must have a similar profile to that where $\hat{\phi}_1$ is the uniform distribution near the origin. We can see such a profile of $g_t(\hat{\phi}_1)$ in Fig. 5. This property is, in some sense, the dual result to that of the quantization problem for stabilization by [8]; that is, the coarsest quantization scheme for stabilization is dense near the origin and becomes coarser as distance from the origin increases. These observations suggest that there appears to exist a trade-off between parameter estimation and stabilization in the quantization scheme for a type of adaptive control system. On the other hand, in the area of the tail of $f(\hat{\phi}_1)$, $g_t(\hat{\phi}_1)$ decreases. However, contrary to our intuition, the resolution remains high, e.g., $g_t(3) \sim 0.208 \sim 45\%$ of $\max g_t(\hat{\phi}_1)$ or $g_t(4) \sim 0.0774 \sim 17\%$ of $\max g_t(\hat{\phi}_1)$, where $f(\hat{\phi}_1)$ is sufficiently small.

Finally $f(\hat{\phi}_1) \sim \tilde{\phi}_1^{-2}$ at the tail of the distribution is an example of a power law. In this case, $g_t$ is constant in the tail and is marginal for the solution’s existence (see Fig. 6). This result shows the difficulty of system identification at sufficient accuracy by using finite information from the system when the tail of the probability density function $f(\hat{\phi}_1)$ is heavier than $O(\tilde{\phi}_1^{-2})$. That is, this explains the complexity of the power law from the viewpoint of parameter estimation in system identification.
Note 3.4 As known from Fig. 4 and Fig. 5, when $f(\tilde{\phi})$ is the normal distribution, uniform distribution or other probable distributions in usual situation of system identification, the marginal density $f(\tilde{\phi}_1)$ is approximately flat near the origin and the quantization becomes coarse in such subsection. Therefore, in order to clarify the minute structure of the optimal quantizer around the origin, we should consider the problem in the coarse resolution with a flat marginal density $f(\tilde{\phi}_1)$. Such case is rigorously analyzed in Section 4.

3.2 Variable-rate quantization

The previous subsection presents the optimal quantizer to minimize the identification error (24) (i.e. (29)) subject to a constraint on the number of quantization steps, i.e., fixed-rate quantization, with high resolution. On the other hand, to reduce the information in the observed data, it is reasonable to apply variable-rate coding for the quantized signals and evaluate the mean code length from the information theoretic viewpoint. From this observation, we consider the minimization problem of (24) (i.e., (29)) subject to a constraint of the expectation of the optimal code length in this subsection, that is, variable-rate quantization, with high resolution.

Let $C(\cdot)$ be an encoder that is a mapping from source alphabets to code alphabets and $l(\cdot)$ be the code length. We regard the quantized output $q(\tilde{\phi}_1)$ as the corresponding source alphabets, then, $l(C(q(\tilde{\phi}_1)))$ represents the code length of $q(\tilde{\phi}_1)$. The expectation of the optimal variable-rate code length for a quantized signal is related to the entropy of the source alphabets by the following well-known source coding theorem.

**Proposition 3.1** [20, 4] Let $x$ be source alphabets, then:

$$\mathbb{E}[l(C(x))] \geq H(x),$$

where $H(x)$ represents the entropy of $x$.

With this proposition, the optimization problem of the quantizer for the code length is reduced to the minimization problem of (24) (i.e., (29)) subject to a constraint on the entropy of the quantized signals.
The basic concept for representing the quantizer with high resolution is the same as that of the previous subsection. That is, subject to Assumption 3.2.1 and 3.2.2, we obtain the asymptotic approximation of the entropy of the quantized signal:

\[ \sum_j -p_j \log p_j \sim \sum_j - \int S_{\tilde{\phi}_1} f(\tilde{\phi}_1) d\tilde{\phi}_1 \log f_j g_j^{-1} \sim \int -f(\tilde{\phi}_1) \log \left( f(\tilde{\phi}_1) g(\tilde{\phi}_1)^{-1} \right) d\tilde{\phi}_1 \]

\[ = H_d(f) + \int -f(\tilde{\phi}_1) \log g(\tilde{\phi}_1) d\tilde{\phi}_1 =: H(f,g), \quad (38) \]

where \( H_d(f) := \int -f(\tilde{\phi}_1) \log f(\tilde{\phi}_1) d\tilde{\phi}_1 \).

By using this asymptotic approximation of the entropy (38), we consider the following problem.

**Problem 3.2** Find

\[ g_v(\tilde{\phi}_1) := \arg \min_g \int \mathcal{F}(g(\tilde{\phi}_1)) d\tilde{\phi}_1 \quad (39) \]

s.t. \( H(f,g) = \log M, \quad (40) \)

where \( \mathcal{F}(\cdot) \) is defined in (32).

Note that \( M \) is the expected number of quantization steps in the sense of (40). We can derive the following theorem:

**Theorem 3.2** The solution of (39) is:

\[ g_v(\tilde{\phi}_1) = KM \sigma(\tilde{\phi}_1), \quad (41) \]

\[ K = \exp L \quad (42) \]

\[ L := -H_d(f) - \int \log \sigma(\tilde{\phi}_1) d\tilde{\phi}_1 = \int f(\tilde{\phi}_1) \log \frac{f(\tilde{\phi}_1)}{\sigma(\tilde{\phi}_1)} d\tilde{\phi}_1. \]

Moreover, the optimized value is:

\[ \int \mathcal{F}(g_v(\tilde{\phi}_1)) d\tilde{\phi}_1 = \frac{1}{12} \tilde{\theta}_1^2 K^{2} M^{-2}. \]

(44)

The proof is in Appendix A.

**Note 3.5** It is interesting that the optimal \( g_v \) is a simple linear function of \( \sigma(\tilde{\phi}_1) \). The constant coefficient is also linear with respect to the number of expected quantization steps \( M \). On the other hand, the convergence rate of the minimized cost function is \( M^{-2} \); this is in common with the fixed-rate quantization. \( \diamond \)

**Example 3.3** When \( f_{\tilde{\phi}} \) is the density function in a multidimensional normal distribution and \( n \) is sufficiently large, as described in Example 3.1,

\[ g_v(\tilde{\phi}_1) = KM \sigma(\tilde{\phi}_1) \sim M \cdot \exp(-H_d(f)) \]

\[ \int \mathcal{F}(g_v(\tilde{\phi}_1)) d\tilde{\phi}_1 \sim \frac{1}{12} \tilde{\theta}_1^2 \exp(2H_d(f)) n \sigma_v^2 M^{-2} = \frac{1}{12} \tilde{\theta}_1^2 2e \pi n \sigma_v^4 M^{-2} \sim 0.4533 \pi \tilde{\theta}_1^2 n \sigma_v^4 M^{-2}. \]

(45)

By comparison with (37) and (45), it can be seen that variable-rate optimal coding achieves approximately half the magnitude of the square of the quantization error compared with \( g_l \) for fixed-rate quantization. \( \diamond \)
4 Quantization in Coarse Resolution

In the previous section, we give the optimal quantization in high resolution for general probability densities of input signals. The results are enough for understanding the profile of the optimal quantization, however, as explained in Note 3.4, its minute structure around the origin is not clear in the case of coarse quantization. In this section, we do not necessarily suppose high resolution of quantization and derive the optimal quantization, however, under limited assumption as follows.

Assumption 4.1 $f(\phi)$ is a probability density function such that $f(\tilde{\phi})$ is uniform distribution in $\tilde{\phi}_i \in [-\kappa, \kappa]$ with a given $\kappa (\in \mathbb{R}) > 0$.

The optimization problem under this assumption has clear significance for the following cases: (1) to clarify the minute quantization scheme around the origin of $y$ because the profile of the multidimensional probability densities of usual input signals in system identification, e.g., normal distribution, is flat around the origin. In such subsection, the quantization is comparatively coarse and the probability density can be approximated as a uniform distribution. The important fact is that such property of the flatness of the probability density around the origin does not depend on the choice of the base in the space of $\phi$. This means the condition of Assumption 4.1 is always satisfied around the origin in usual situation of system identification. (2) to consider the first order systems where input signals obey a uniform distribution. In this case, the analytic optimal solution in coarse quantization can be given and it is enough for the main subject of this paper to clarify the essential properties of the optimal quantizers for parameter estimation.

When Assumption 4.1 is satisfied, as similar to the case of Section 3, $\frac{1}{N}U^T\hat{U}$ and $\frac{1}{N}\hat{U}^T\hat{U}$ also converge to $\sigma_u^2I$ when $N \to \infty$, then the optimal quantization problem is also reduced to minimize $\mathbb{E} [U^TE]$ (equivalently $\mathbb{V} [\hat{U}^TE]$) of (22) subject to a bias free condition: $\mathbb{E} [U^TE] = 0 \left(\text{equivalently } \mathbb{E} [\hat{U}^TE] = 0\right)$, i.e. (19) and (20).

Under Assumption 4.1, it is obvious that
\[
\int \phi_k f(\tilde{\phi}_1, \tilde{\phi}_k) d\tilde{\phi}_k = 0 \quad \text{(46)}
\]
for $k \neq 1$, then, (19) is automatically satisfied. Therefore, the bias-free condition is reduced to (20). Moreover, (20) means
\[
\int \phi_1 e(\tilde{\phi}_1) f(\tilde{\phi}_1, \tilde{\phi}_2, \ldots, \tilde{\phi}_n) d\tilde{\phi}_1 = 0 \quad \text{(47)}
\]
under Assumption 4.1. A sufficient condition for (20) is
\[
\mathbb{E}_{\tilde{\phi}_1} \left[ \phi_1 e(\tilde{\phi}_1) \right] := \int_{\tilde{\phi}_1 \in S_{\phi_1}} \phi_1 e(\tilde{\phi}_1) f(\tilde{\phi}_1) d\tilde{\phi}_1 = \int_{\tilde{\phi}_1 \in S_{\phi_1}} \phi_1 (y_j' - \tilde{\theta}_1 \tilde{\phi}_1) f(\tilde{\phi}_1) d\tilde{\phi}_1 = 0, \quad \forall j. \quad \text{(48)}
\]
This condition is sufficiently reasonable for the representative number $y'_j$ of the subsection $S_j'$ (or the corresponding $S_{\phi_1}$ on $\tilde{\phi}_1$).

On the other hand, we can derive the following key lemma for the cost function $\mathbb{V}[U^TE]$ (equivalently $\mathbb{V} [\hat{U}^TE]$) of (22):

Lemma 4.1 Subject to the conditions:
\[
\int \phi_h f(\tilde{\phi}_1, \ldots, \tilde{\phi}_h, \ldots, \tilde{\phi}_n) d\tilde{\phi}_h = 0, \quad \forall h = 1, 2, \ldots, n \quad \text{(49)}
\]
and
\[
\int \phi_1 e(\tilde{\phi}_1) f(\tilde{\phi}_1) d\tilde{\phi}_1 = 0, \quad \text{(50)}
\]
\[ E \left[ \left( \sum_{k=1}^{N} \tilde{\phi}_k(t)e(\tilde{\phi}_1(t)) \right)^2 \right] = \begin{cases} N \int \tilde{\phi}_1 e^2(\tilde{\phi}_1)f(\tilde{\phi}_1)d\tilde{\phi}_1 & \text{for } k = 1 \\ N \int \tilde{\phi}_k^2 e^2(\tilde{\phi}_1, \tilde{\phi}_k)d\tilde{\phi}_1d\tilde{\phi}_k & \text{for } k \neq 1 \end{cases} \] (51)

is satisfied.

The proof of this lemma is in Appendix A.

Assumption 4.1 automatically guarantees the condition (46), i.e. (49), and therefore with the bias-free condition (50), (51) follows from Lemma 4.1. With these preliminaries, we formulate the problem considered in this section:

**Problem 4.1** Let \( M \) be the number of quantized subsections \( S^q_j \) of \([-\kappa_y, \kappa_y] := [-\kappa_\tilde{t}_1, \kappa_\tilde{t}_1] \) on \( y \) (i.e., \( S^\tilde{\phi}_1 \) of \([-\kappa, \kappa] \) on \( \tilde{\phi}_1 \)) where \( M \geq 2 \). For the system (1) with Assumption 4.1 and a fixed \( M \), find a quantizer \( q \) that minimizes

\[ V[ U^T E] = V[ \tilde{U}^T E] = \sum_{k=1}^{n} E \left[ \left( \sum_{i=1}^{N} \tilde{\phi}_k(t)e(\tilde{\phi}_1(t)) \right)^2 \right] = N \int \sigma^2(\tilde{\phi}_1)^2 e^2(\tilde{\phi}_1)f(\tilde{\phi}_1)d\tilde{\phi}_1 \] (52)

such that \( E_{S^\tilde{\phi}_1} [\tilde{\phi}_1 e(\tilde{\phi}_1)] = 0 \) for all \( j \).

The reason for the constraint \( M \geq 2 \) is described in Note 4.1.

As described in Section 2, the quantization scheme of \([-\kappa_\tilde{t}_1, \kappa_\tilde{t}_1] \) on \( y \) is essentially equal to that of \([-\kappa, \kappa] \) on \( \tilde{\phi}_1 \) and it is completely defined by the setting of the subsections \( S^\tilde{\phi}_1, S^\tilde{\phi}_2, \ldots, S^\tilde{\phi}_{M} \), where

\[ M' := \begin{cases} \frac{1}{2} M & \text{for even } M \ (\geq 2) \\ \frac{1}{2} (M - 1) & \text{for odd } M \ (\geq 3) \end{cases} \] (53)

and the assigned quantized values

\[ q(y)|_{y \in S^q_j} = q(\tilde{\phi}_1 \tilde{t}_1)|_{\tilde{\phi}_1 \in S^\tilde{\phi}_1} = y_{(j)} \]

for each subsection \( S^q_j \) or \( S^\tilde{\phi}_1 \) (see Fig. 7). Therefore, optimization of the quantization is reduced to a minimization problem of \( V[ U^T E] \) of approximately \( 2M \)-variables \( (d_{-(M'-1)}, \ldots, d_{M'-1} \) and \( y'_{(-M')}, \ldots, y'_{(M')} \), note that \( d_{M'} = \kappa_\tilde{t}_1 \) and \( d_{-M'} = -\kappa_\tilde{t}_1 \).

16
In this section, we consider the case of even $M$. The case of odd $M$, that is, $S^y_0 \neq \{0\}$ ($\tilde{S}^\phi_1 \neq \{0\}$), is reduced to the even case and the reason is explained in Note 4.1. We also refer to the positive domain $S^y_1, S^y_2, \ldots$ because of quantization symmetry.

It is known that when a subsection $S^y_j$ is fixed (i.e. $d_{j-1}$ and $d_j$ are fixed), $y'_j$ is given by the bias-free condition $E_{\tilde{S}^\phi_1} \left[ \tilde{\phi}_1 e(\tilde{\phi}_1) \right] = 0$. Therefore, the optimization problem is reduced to finding optimal $d_{-(M'-1)}, \ldots, d_{M'}$. Corresponding to $d_j$, we introduce key variables, ratios $r_j$ ($j = 1, \ldots, M - 1$) between $d_j$ and $d_{j+1}$ defined by:

$$d_j = r_j d_{j+1}, \quad r_j \in [0, 1].$$  \hspace{1cm} (54)

Note that determining optimal $d_{-(M'-1)}, \ldots, d_{M'}$ is equal to determining optimal $r_{-(M'-1)}, \ldots, r_{M'}$ and we derive the following result.

**Proposition 4.1** The optimal ratios $r^o_j$ for Problem 4.1 are given by solving the following recursive optimization problem iteratively.

$$r^o_j = \arg \min_{r \in [0, 1]} \left( d_{j+1}^5 \psi(r; \psi^\text{min}_{j-1}) + 20 \kappa_y^2 (n-1) d_{j+1}^3 \xi(r; \xi^\text{min}_{j-1}) \right)$$  \hspace{1cm} (55)

$$\psi(r; \alpha) := \alpha r^5 - 18(1-r)^5 + 45(1+r)^2(1-r)^3 + 5(1-r)^7(1+r)^{-2}$$  \hspace{1cm} (56)

$$\psi^\text{min}_j := \psi(r^o_j; \psi^\text{min}_{j-1})$$

$$\psi^\text{min}_0 := 32$$

$$\xi(r; \alpha) := \alpha r^3 + 3(1-r)^3 + \frac{(1-r)^5}{(1+r)^2}$$  \hspace{1cm} (57)

$$\xi^\text{min}_j := \xi(r^o_j; \xi^\text{min}_{j-1})$$

$$\xi^\text{min}_0 := 4.$$
The optimal value of (52) is given by

\[
\min_q V[U^T E] \left( = \min_q V[\tilde{U}^T E]\right) = \min_q \sum_{j=-M'}^{M'} V[\tilde{S}_j \tilde{E}] = N \frac{\kappa_1^4}{2160} \left( \psi_{M'-1}^{\min} + 20(n-1)\xi_{M'-1}^{\min} \right) = N \frac{\tilde{\theta}_1^4 \kappa_1^4}{2160} \left( \psi_{M'-1}^{\min} + 20(n-1)\xi_{M'-1}^{\min} \right).
\] (58)

See Appendix A for the proof.

Note 4.1 For odd \( M \), there must not exist a subsection \( S_0^{\phi_1} \) (i.e. \( S_0^{\tilde{\phi}_1} \)) of nonzero width that contains the origin of \( y \) (i.e., origin of \( \tilde{\phi}_1 \)) because for any such subsection and setting \( y_{(0)}, E[S_0^{\tilde{\phi}_1} \tilde{\phi}_1 e(\tilde{\phi}_1)] \neq 0 \). This means that \( S_0^{\mu} \) (i.e. \( S_0^{\tilde{\phi}_1} \)) should be \( \{0\} \) and consequently the problem is equal to the case of even \( M \) with the setting \( M' = \frac{1}{2}(M-1) \).

\[\diamondsuit\]

Example 4.1 Consider the following second-order FIR model as an example of (1):

\[y(t) = \theta_1 u(t) + \theta_2 u(t-1),\]

where \( \theta_1 = \sqrt{2} \) and \( \theta_2 = \frac{1}{2} \) and the system is noise free. We generate 50 sets of I/O data sequences with a length \( N = 10,000 \) for the system (59) that obey Assumption 4.1 and \( \kappa = 4 \) (i.e., \( \kappa_y = 4 \)). Fig. 8 is one of the histogram of 10,000 samples of \( \tilde{\phi}_1 \) from 50 sets.

Next, quantize the output data \( y \) with the optimal quantizers given by Proposition 4.1 and with uniform quantizers, for comparison, subject to the constraints \( M' = 5 \) (\( M = 10 \)). Fig 9 shows the step function \( q \) for \( y \) of the optimal quantizer for \( M' = 5 \). Fig 9 indicates a basic property of the optimal quantizer, that is, it is coarse near the origin and becomes denser away from the origin.

The bias term \( \frac{1}{N} \sum_{t=1}^{N} \tilde{\phi}_1(t)e(t) \) and the quantization error term \( \Delta E \) were calculated; Table 1 shows a summary of the results. From Table 1, the optimal quantizer, which minimize \( V[U^T E] \) attains a lower \( \|\Delta E\|_2^2 \) than that of the uniform quantizer.
Table 1 The ratios of the biases and the squares of errors for $M' = 5$ (averages of 50 sets)

| $|\sum_{t=1}^{10000} \hat{\phi}_1(t)c(t)|$ by opt. quant. | $|\sum_{t=1}^{10000} \hat{\phi}_1(t)c(t)|$ by unif. quant. | $\|\Delta E\|_2^2$ by opt. quant. | $\|\Delta E\|_2^2$ by unif. quant. |
|---------------------------------------------------------|----------------------------------------------------------|---------------------------|---------------------------|
|                                                                      | 0.1107                                                   | 0.0132                    |

Proposition 4.1 shows that the problem is in a category of the typical dynamic programming and we can solve it by numerical calculation. In general, the computation complexity of this problem is high; however, the optimization problem (55) can be solved by very few calculation steps in special cases $n = 1$ or $n \gg 1$, respectively, as shown in the following theorem:

**Theorem 4.1** When $n = 1$, the optimal ratios $r_j^o$ for Problem 4.1 are given by solving the following optimization problem iteratively.

$$r_j^o = \arg \min_{r \in [0, 1]} \psi(r; \psi_{j-1}^\text{min})$$

$$\psi_{j}^\text{min} := \psi(r_j^o; \psi_{j-1}^\text{min})$$

$$\psi_0^\text{min} := 32.$$  \hspace{2cm} (60)

The optimal value of (52) is given by

$$\min_q V [U^T E] \left( = \min_q V [\tilde{U}^T E] \right) = \frac{N}{2160} \beta_1^4 \kappa^4 \psi_{M'}^\text{min} - 1.$$ \hspace{2cm} (62)

Similarly, when $n \gg 1$, the optimal ratios $r_j^o$ for Problem 4.1 converge to the solution of the following optimization problem.

$$r_j^o = \arg \min_{r \in [0, 1]} \xi(r; \xi_{j-1}^\text{min})$$

$$\xi_{j}^\text{min} := \xi(r_j^o; \xi_{j-1}^\text{min})$$

$$\xi_0^\text{min} := 4.$$ \hspace{2cm} (63)

The optimal value of (52) converges to

$$\frac{N}{108} \beta_1^4 \kappa^4 (n - 1) \xi_{M'}^\text{min}.$$ \hspace{2cm} (65)

**Note 4.2** The definitive difference of the optimization problems (55) and (60) or (63) is that in the former case, $r_j^o$ depends on $d_{j+1}$ and this requires a complex calculation such as dynamic programming, on the other hand, in the latter cases, $r_j^o$ does not depend on $d_{j+1}$ and $\{r_j^o\}$ can be given by solving (60) or (63) from $j = 1$ to $j = M' - 1$ in turn only once. This means that the original minimization problem of approximately $2M$-variable function $V [U^T E]$ can be reduced to a recursive minimization problem of a single one-variable rational function when $n = 1$ or $n \gg 1$. Moreover, when $n = 1$, from Lemma A.1 in Appendix A, the local minimum of $\phi(r; \alpha)$, $\alpha > 0$, in $r \in (0, 1)$ is unique. Therefore, finding the minimizer does not require a highly complex calculation.

In the following of this section, we focus on the case $n = 1$ because it is a basic problem and reveals typical property of the optimal quantization. We call the optimal quantization scheme as $Q_{\text{opt}}$ hereafter.
Every optimal ratio $r_j^o$ can be explicitly determined by solving (60) – (61) iteratively; however, the properties of the sequence $r_1^o, r_2^o, \ldots$ are not clear from (60) – (61). For the asymptotic characteristics of the optimal ratios $r_j^o (j = 1, 2, \ldots)$ and related quantities, we derive the following series of Lemma 4.2 – 4.5.

**Lemma 4.2** The optimal ratios $r_j^o$ satisfies:

$$r_j^o < r_{j+1}^o, \forall j > 0,$$

$$r_j^o \to 1, j \to \infty.$$

**Lemma 4.3** The width of the subsections $S_j^y$ or $\hat{S}_j^{\hat{y}_1}$ of $Q_{\text{opt}}$ satisfy:

$$|S_j^y| > |S_{j+1}^y|, \quad |\hat{S}_j^{\hat{y}_1}| > |\hat{S}_{j+1}^{\hat{y}_1}|, \forall j > 0,$$

where $| \cdot |$ denotes the width of the subsection.

The proofs of these lemmas are in Appendix A.

Lemma 4.3 shows that the optimal quantization scheme $Q_{\text{opt}}$ has the property that it is coarse near the origin of $y$ and becomes denser as $y$ tends to the boundaries of $[-\kappa y, \kappa y]$. This property coincides with the results in Section 3 and it is also the dual result to that of the quantization problem for stabilization by [8] as mentioned in Section 3.

Next, consider the unboundedness of $\prod_{j=1}^\infty \frac{1}{r_j^o}$. If it is bounded and $\prod_{j=1}^\infty \frac{1}{r_j^o} = \gamma < \infty$, then this causes a contradiction as to the optimality of $Q_{\text{opt}}$, that is, when a region $[-\gamma, \gamma]$ of $\hat{y}_1$ is quantized, the width of $\hat{S}_1^{\hat{y}_1}$, for example, is never smaller than 1 even if the number of quantization levels increases to infinity. Of course, this is not true and $\prod_{j=1}^\infty \frac{1}{r_j^o}$ is therefore unbounded. The next lemma strictly describes this fact. Refer to [24] for the proof.

**Lemma 4.4** The optimal ratios $r_j^o$ satisfies:

$$\prod_{j=1}^\infty \frac{1}{r_j^o} = \infty$$

From Lemma 4.2 to Lemma 4.4, we know the outline of the quantization of the region $[-\kappa y, \kappa y]$.

Next, to clarify the profile of $\nabla \left[ U^T E \right]$ with respect to $M'$, the following lemma confirms the asymptotic characteristics of $\psi_{\text{min}}^{M'}$.

**Lemma 4.5** The minimized quantity $\psi_{\text{min}}^{M'}$ of (56) at $j = M'$ converges as

$$\psi_{\text{min}}^{M'} \to \Psi^b_a(M'), \quad M' \to \infty,$$

where $a = -5 \cdot 3^{-\frac{4}{3}}$ and $b = \frac{3}{2}$, and $\Psi^b_a(m)$ is a function of integer $m$ defined as the solution of the following recurrence formula with an appropriate initial number $\psi(0) = \psi_o$:

$$\hat{\psi}(m) - \hat{\psi}(m-1) = a \hat{\psi}^b(m-1). \quad (66)$$

The proof is in Appendix A.
Note that the recurrence formula (66) is from (90) in Appendix A, and it can be approximated by \( \tilde{\psi} \), which is a solution of a differential equation:

\[
\frac{d\tilde{\psi}(m)}{dm} = (a + \nu)\tilde{\psi}(m) + a\tilde{\psi}^2(m) = P(\tilde{\psi}(m)), \quad m \in \mathbb{R},
\]

where \( P(\cdot) \) is defined in (90) and \( \nu > 0 \) is an appropriate constant number satisfying \( a + \nu < 0 \) and the above inequality (such \( \nu \) always exists). We can show \( \tilde{\psi}(m) \gtrsim \tilde{\psi}(m) \) at sufficiently large integer \( m \) when \( \tilde{\psi}(0) \gtrsim \tilde{\psi}(0) \) in Lemma A.2. Then, we obtain the solution

\[
\tilde{\psi}(m) = \{(b + 1)(a + \nu)m + B\}^{1/2}
\]

for an appropriate constant \( B \). From (62) and (67), we obtain

\[
\min_q V[U^T E] \leq \frac{N}{2160} \kappa^4 \left((-3/2 + 1)((-5 \cdot 3^{-\frac{3}{2}} + \nu)(M' - 1) + B)\right)^{-\frac{3}{2}}.
\]

\[
A := \frac{N}{540} \left(5 \cdot 3^{-\frac{3}{2}} - \nu\right)^{-2}, \quad B := (5 \cdot 3^{-\frac{3}{2}} - \nu)^{-1} B.
\]

This (68) approximately shows the relationship between the optimized quantization error \( \min_q V[U^T E] \) and the number of quantization levels.

**Example 4.2** Consider the following first-order FIR model for verifying the above results:

\[
y(t) = \theta u(t),
\]

where \( \theta = 2 \) and the system is noise free. We also generate 50 sets of I/O data sequences with a length \( N = 10,000 \) for the system (69) that obey Assumption 4.1 and \( \kappa = 4 \) (i.e., \( \kappa_y = 8 \)).

Next, quantize the output data \( y \) with the optimal quantizers given by Theorem 4.1 and with uniform quantizers, for comparison, subject to the constraints \( M' = 5 \) \( (M = 10) \). Fig 10 shows the step function \( q \) for \( y \) of the optimal quantizer for \( M' = 5 \). From the comparison with Fig 9, Fig 10 more clearly shows the property of the optimal quantizer, that is, it is coarse near the origin and becomes denser away from the origin.

Table 2 shows comparison of the bias term \( \frac{1}{N} \sum_{t=1}^N \tilde{\phi}_1(t)e(t) \) and the quantization error term \( \Delta E \). From Table 2, the optimal quantizer, which minimize \( V[U^T E] \) attains a lower \( \| \Delta E \|^2 \) than those of the uniform quantizer.
Table 2. The ratios of the biases and the squares of errors for \( M' = 5 \) (averages of 50 sets)

| \( \sum_{t=1}^{10000} \tilde{\phi}_1(t)e(t) \) by \( Q_{\text{opt}} \) / \( \sum_{t=1}^{10000} \tilde{\phi}_1(t)e(t) \) by unif. quant. | 0.0135 |
| \( \|\Delta E\|_2^2 \) by \( Q_{\text{opt}} \) / \( \|\Delta E\|_2^2 \) by unif. quant. | 0.0116 |

5 Resolution of the Quantizer and I/O Data Length

In the system identification of (1), it is important to clarify the relationship between the estimation error and the amount of signal data used for the estimation. The amount of signal data is the resolution of the quantization multiplied by the length of signal sequence. Using the results in the previous sections, we evaluate the magnitudes of the error term \( \Delta \hat{E} \) and \( \Delta \hat{W} \) based on the approach in [25] and compare the effects of the resolution of quantizers and the length of signal sequence.

First, the evaluation of the magnitude of \((\hat{U}^T\hat{U})^{-1}\).

**Lemma 5.1** [25] Assume that \( \tilde{\phi} \) satisfies Assumption 3.1 and 3.2 with \( \mathcal{V}[\tilde{\phi}_1(t)] = \sigma_{\phi_1}^2\), \( \mathcal{V}[\tilde{\phi}_2(t)] = \eta \). Then, for any reliability index \( \beta_1 > 0 \), where \( 1 - \beta_1 > 0 \), and \( \sigma_{\phi_1}^2 \sqrt{N} - n \sqrt{\frac{\sigma_{\phi_1}^2}{\sigma_{\phi_2}^2}} (\sqrt{n} + (n-1)\sigma_{\phi_2}^2) > 0 \), the following inequality is satisfied.

\[
\mathbb{P} \left( \| (\hat{U}^T\hat{U})^{-1} \|_1 \geq \epsilon_1 \right) \leq \beta_1 \\
\epsilon_1 := \frac{1}{\sigma_{\phi_1}^2 N - n \sqrt{\frac{\sigma_{\phi_1}^2}{\sigma_{\phi_2}^2}} (\sqrt{n} + (n-1)\sigma_{\phi_2}^2)}
\]

Using Lemma 5.1, we evaluate \( \|\Delta \hat{E}\|_\infty \) in the following theorem.

**Theorem 5.1** For the system (1) with the optimal quantizer \( q(y) \) defined by (3) – (5), (33), assume Assumption 3.1 and 3.2. Then, for the reliability indices \( \beta_1, \beta_2 > 0 \), a length of data \( N \) and the number of quantization levels \( M \), where \( 1 - \beta_1 - \beta_2 > 0 \), and \( \sigma_{\phi_1}^2 \sqrt{N} - n \sqrt{\frac{\sigma_{\phi_1}^2}{\sigma_{\phi_2}^2}} (\sqrt{n} + (n-1)\sigma_{\phi_2}^2) > 0 \), the following inequality asymptotically holds at \( \Delta y \to 0 \):

\[
\mathbb{P} \left( \|\Delta \hat{E}\|_\infty \leq \epsilon_1 \epsilon_2 \right) \geq 1 - \beta_1 - \beta_2 \\
\epsilon_1 := \frac{1}{\sigma_{\phi_1}^2 N - n \sqrt{\frac{\sigma_{\phi_1}^2}{\sigma_{\phi_2}^2}} (\sqrt{n} + (n-1)\sigma_{\phi_2}^2)} , \quad \epsilon_2 := \frac{1}{M} \sqrt{\frac{1}{12} \beta_1^2 D^3} \sqrt{nN} / \beta_2.
\]

The proof is in Appendix A.

From this theorem, we know that the convergence rate of the error term \( \|\Delta \hat{E}\|_\infty \) has an order of \( M^{-1} \) for sufficiently large \( M \) and of \( N^{-\frac{1}{12}} \). Approximately, the total amount of information in the quantized output transmitted from identified systems to the observers is approximately \( N \log_2 M =: K \) using binary coding. Therefore, subject to a constraint of such a total amount of information, it is known that a large \( M \) is preferable to a large \( N \) to reduce the estimation error by observing:

\[
M^{-1} N^{-\frac{1}{12}} = M^{-1} \left( \frac{K}{\log_2 M} \right)^{-\frac{1}{12}} = K^{-\frac{1}{12}} M^{-1} (\log_2 M)^{\frac{1}{12}} \xrightarrow{M \to \infty} 0.
\]

Of course, this is valid only for the error term \( \|\Delta \hat{E}\|_\infty \) and the situation is different for the noise error term \( \Delta \hat{W} \). We introduce the result for \( \Delta \hat{W} \) in the following proposition.
Proposition 5.1 [25] Assume that \( \hat{\phi} \) satisfies Assumption 3.1 and 3.2 and \( w(t) \) is i.i.d. random variable with \( \text{V}[\hat{\phi}_1(t)] = \sigma_{\hat{\phi}_1}^2 \), and \( \text{V}[w(t)] = \sigma_w^2 \), respectively. Then, for reliability indices \( \beta_1, \beta_2 > 0 \), and a length of data \( N \), where \( 1 - \beta_1 - \beta_2 > 0 \), and \( \sigma_{\hat{\phi}_1}^2 N - n \sqrt{\frac{N}{\beta_1}} \left( \sqrt{\eta} + (n - 1)\sigma_{\hat{\phi}_1}^2 \right) > 0 \), the following inequality is satisfied.

\[
\text{Prob}\left( \|\Delta \hat{W}\|_\infty \leq \epsilon_1 \epsilon_2 \right) \geq 1 - \beta_1 - \beta_2
\]

(73)

\[
\epsilon_1 := \frac{1}{\sigma_{\hat{\phi}_1}^2 N - n \sqrt{\frac{N}{\beta_1}} \left( \sqrt{\eta} + (n - 1)\sigma_{\hat{\phi}_1}^2 \right)}, \quad \epsilon_2 := \frac{\sigma_{\hat{\phi}_1} \sigma_w \sqrt{nN}}{\beta_2}
\]

(74)

This result shows that a large \( N \) is preferable for reducing \( \Delta \hat{W} \). By combining Theorem 5.1 and Proposition 5.1, it can be seen that there exists a trade-off between \( \Delta \hat{E} \) and \( \Delta \hat{W} \) (also \( \Delta E \) and \( \Delta W \)) for reducing the total identification error subject to the constraint on the amount of information transmitted from the identified systems to the estimators.

6 Conclusion

In this paper, we show that the optimal quantizers for system identification can be derived analytically and their essential properties investigated with a simple FIR model. The results of this paper are summarized as follows:

(1) General cases of the distribution of regressor vectors can be treated for high resolution quantizers by introducing the concept of the density of quantization subsections (Section 3).

(2) The optimization problems in (1) are reduced to minimizations of functionals and the solutions can be found by solving Euler–Lagrange differential equations (Section 3).

(3) When the regressor vector has a form of uniform distribution, the optimal quantization problem is reduced to a recursive minimization, which can be solved by a dynamic programming (Section 4).

(4) In usual situation, the optimal quantizer is coarse near the origin of the output signals and tends to be dense away from the origin (Section 3 and Section 4).

(5) Subject to a limitation on the total quantity of information in the quantized I/O data, there exists a trade-off between the magnitudes of the quantization error and noise error (Section 5).

In this paper, we restrict the model to a SISO FIR model. For more realistic situations, we must extend the results to: a) ARX models, or MIMO systems, b) quantized input signal, and c) online system identification and adaptive control. These remain for future study.

Acknowledgement

The author expresses deep gratitude for discussion and help to Professor Jan Maciejowski, University of Cambridge.

References

[1] W. R. Bennett, “Spectra of quantized signals,” The Bell System Technical Journal, vol. 27, pp. 446–472, 1948.
[2] T. Berger, “Optimum quantizers and permutation codes,” *IEEE Transactions on Information Theory*, vol. 18, no. 6, pp. 759–765, 1972.

[3] R. W. Brockett and D. Liberzon, “Quantized feedback stabilization of linear systems,” *IEEE Trans. Automat. Control*, vol. 45, no. 7, pp. 1279–1289, 2000.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley series in telecommunications. John Wiley & Sons, Inc., New York, 1991.

[5] R. E. Currie, *Estimation and control with quantized measurements*, M.I.T. Press, Cambridge, MA, 1970.

[6] D. F. Delchamps, “Extracting state information from a quantized output record,” *Systems & Control Letters*, vol. 13, pp. 365–372, 1989.

[7] D. F. Delchamps, “Stabilizing a linear system with quantized state feedback,” *IEEE Trans. Automat. Control*, vol. 35, no. 8, pp. 916–924, 1990.

[8] N. Elia and S. K. Mitter, “Stabilization of linear systems with limited information,” *IEEE Trans. on Automatic Control*, vol. 46, no. 9, pp. 1384–1400, 2001.

[9] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression Control*, Kluwer international series in engineering and computer science. Kluwer Academic Publishers, 1992.

[10] M. Gevers and G. Li, *Parametrization in Control, Estimation and Filtering Problems: Accuracy aspects*, Communications and control engineering series. Springer-Verlag, Berlin, 1993.

[11] H. Gish and J. N. Pierce, “Asymptotically efficient quantizing,” *IEEE Transactions on Information Theory*, vol. 14, no. 5, pp. 676–683, 1968.

[12] S. Graf and H. Luschgy, *Foundations of Quantization for Probability Distributions*, Communications and control engineering series. Springer-Verlag, Berlin Heidelberg, 2000.

[13] R. M. Gray and D. L. Neuhoff, “Quantization,” *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 1–63, 1998.

[14] G. Grimmett and D. Stirzaker, *Probability and Random Processes*, Oxford, 3rd ed., 2001.

[15] S. P. Lloyd, “Least squares quantization in PCM,” *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[16] G. N. Nair and R. J. Evans, “Stabilization with data-rate-limited feedback: tightest attainable bounds,” *Systems and Control Letters*, vol. 41, pp. 49–56, 2000.

[17] G. N. Nair and R. J. Evans, “Mean square stabilisability of stochastic linear systems with data rate constraints,” *Proceedings of the 41st IEEE Conference on Decision and Control*, pp. 1632–1637, 2002.

[18] G. N. Nair and R. J. Evans, “Stabilizability of stochastic linear systems with finite feedback data rates,” *SIAM J. Control Optim.*, vol. 43, no. 2, pp. 413–436, 2004.

[19] L. L. Scharf, *Statistical Signal Processing*, Addison-Wesley, 1991.

[20] C. E. Shannon, “A mathematical theory of communication,” *Bell Sys. Tech. Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[21] K. Tsumura, “Asymptotic property of optimal quantization for system identification,” *Technical Report of The Univ. Tokyo*, METR 2004–10, February 2004.

[22] K. Tsumura and J. Maciejowski, “Optimal quantization of signals for system identification,” *Technical Report of The Univ. Cambridge*, CUED/F-INFENG/TR445, 2002 (also in Proceedings of the Euromean Control Conference 2003, Cambridge, UK, 2003).

[23] K. Tsumura and J. Maciejowski, “Stabilizability of SISO control systems under constraints of channel capacities,” *Proceedings of the 42th Conference on Decision and Control*, pp. 193–198, Maui, USA, 2003.

[24] K. Tsumura and J. Maciejowski, “Analysis on optimal quantization of signals for system identification and the effect of noise,” *Technical Report of The Univ. Tokyo*, METR 2005–04, January 2005 (http://www.keisu.t.u-tokyo.ac.jp/Research/METR/2005/METR05-04.pdf).
A Appendix

Slutsky’s Theorem (e.g. [14])

For sequences of stochastic variables $X(i), Y(i)$, assume that $\lim_{i \to \infty} [X(i)]$ and $\lim_{i \to \infty} [Y(i)]$ converge to constants. Then,

$$\lim_{i \to \infty} [X(i)]^{-1}Y(i) = \left( \lim_{i \to \infty} [X(i)] \right)^{-1} \lim_{i \to \infty} [Y(i)]$$

holds.

Proof of Lemma 3.1

The outline of the proof is similar to that of Lemma 4.1 and we evaluate the value of: $E \left[ \hat{o}_a e(\hat{o}_b) \hat{o}_c e(\hat{o}_d) \right]$ for possible cases in (23): $a \neq b \neq c \neq d$, $a = b \neq c \neq d$, $a = b = c = d$, and $a = c \neq b = d$ (the other possible cases in (23) are essentially identical to these cases).

Let $S_{\hat{o}_a}, S_{\hat{o}_b}, S_{\hat{o}_c}, S_{\hat{o}_d}$ be a quantized subsection of the axis of $\hat{o}_a, \hat{o}_b, \hat{o}_c, \hat{o}_d$, respectively, and consider a subset $S_{\hat{o}_a} \times S_{\hat{o}_b} \times S_{\hat{o}_c} \times S_{\hat{o}_d}$ in the space of $\hat{o}$. Moreover, let $\hat{\phi}_a', \hat{\phi}_b', \hat{\phi}_c', \hat{\phi}_d'$ be the quantized values, which are midpoints of $S_{\hat{o}_a}, S_{\hat{o}_b}, S_{\hat{o}_c}, S_{\hat{o}_d}$, respectively. The partial integral of $E \left[ \hat{o}_a e(\hat{o}_b) \hat{o}_c e(\hat{o}_d) \right]$ restricted to this subset is

$$\int_{S_{\hat{o}_a} \times S_{\hat{o}_b} \times S_{\hat{o}_c} \times S_{\hat{o}_d}} \hat{o}_a e(\hat{o}_b) \hat{o}_c e(\hat{o}_d) f(\hat{o}_a, \hat{o}_b, \hat{o}_c, \hat{o}_d) d\hat{o}_a d\hat{o}_b d\hat{o}_c d\hat{o}_d.$$

Let $2\Delta \hat{o}$ be the width of the largest side of the possible hyperrectangular parallelepipeds in $\hat{\phi}$ given by quantization, then, when $a \neq b \neq c \neq d$:

$$\int_{S_{\hat{o}_a} \times S_{\hat{o}_b} \times S_{\hat{o}_c} \times S_{\hat{o}_d}} \hat{o}_a e(\hat{o}_b) \hat{o}_c e(\hat{o}_d) f(\hat{o}_a, \hat{o}_b, \hat{o}_c, \hat{o}_d) d\hat{o}_a d\hat{o}_b d\hat{o}_c d\hat{o}_d$$

$$= \int_{S_{\hat{o}_a} \times S_{\hat{o}_b} \times S_{\hat{o}_c} \times S_{\hat{o}_d}} \hat{o}_a e(\hat{o}_b) \hat{o}_c e(\hat{o}_d)$$

$$\times \left( \delta_0 + \sum_{i} \delta_i (\hat{o}_i - \hat{o}_i') + \sum_{i,j} \delta_{ij} (\hat{o}_i - \hat{o}_i') (\hat{o}_j - \hat{o}_j') + O((\hat{o}_i - \hat{o}_i') (\hat{o}_j - \hat{o}_j') (\hat{o}_k - \hat{o}_k')) \right) d\hat{o}_a d\hat{o}_b d\hat{o}_c d\hat{o}_d$$

$$= \hat{o}_a' \hat{o}_c' \delta_a \delta_c \frac{2^4}{3^2} \Delta \hat{o}^8 + O(\Delta ^9), \quad (75)$$

and similarly, when $a = b \neq c \neq d$:

$$\int_{S_{\hat{o}_a} \times S_{\hat{o}_b} \times S_{\hat{o}_c} \times S_{\hat{o}_d}} \hat{o}_a e(\hat{o}_b) \hat{o}_c e(\hat{o}_d) f(\hat{o}_a, \hat{o}_b, \hat{o}_c, \hat{o}_d) d\hat{o}_a d\hat{o}_b d\hat{o}_c d\hat{o}_d = (\hat{o}_a' \hat{o}_c' \delta_a \delta_c + \hat{o}_b' \delta_b \delta_d) \frac{2^4}{3^2} \Delta \hat{o}^8 + O(\Delta ^9), \quad (76)$$

and when $a = b \neq c = d$:

$$\int_{S_{\hat{o}_a} \times S_{\hat{o}_b} \times S_{\hat{o}_c} \times S_{\hat{o}_d}} \hat{o}_a e(\hat{o}_b) \hat{o}_c e(\hat{o}_d) f(\hat{o}_a, \hat{o}_b, \hat{o}_c, \hat{o}_d) d\hat{o}_a d\hat{o}_b d\hat{o}_c d\hat{o}_d$$

$$= \hat{o}_a' \hat{o}_c' \delta_a \delta_c + \hat{o}_b' \delta_b \delta_d \frac{2^4}{3^2} \Delta \hat{o}^8 + O(\Delta ^9). \quad (77)$$

25
Alternatively, when \(a = b = c = d\):

\[
\int_{S_{\tilde{\phi}} \times S_{\tilde{\phi}} \times S_{\tilde{\phi}} \times S_{\tilde{\phi}}} \tilde{\phi}_a e(\tilde{\phi}_a) \tilde{\phi}_b e(\tilde{\phi}_b) f(\tilde{\phi}_a, \tilde{\phi}_b, \tilde{\phi}_c, \tilde{\phi}_d) d\tilde{\phi}_a d\tilde{\phi}_b d\tilde{\phi}_c d\tilde{\phi}_d = \tilde{\phi}_a^2 \tilde{\phi}_b^2 \frac{2^4}{3} \Delta \tilde{\phi}^6 + O(\Delta \tilde{\phi}^7) \quad (78)
\]

and similarly, when \(a = c \neq b = d\):

\[
\int_{S_{\tilde{\phi}} \times S_{\tilde{\phi}} \times S_{\tilde{\phi}} \times S_{\tilde{\phi}}} \tilde{\phi}_a e(\tilde{\phi}_a) \tilde{\phi}_b e(\tilde{\phi}_b) f(\tilde{\phi}_a, \tilde{\phi}_b, \tilde{\phi}_c, \tilde{\phi}_d) d\tilde{\phi}_a d\tilde{\phi}_b d\tilde{\phi}_c d\tilde{\phi}_d = \tilde{\phi}_a^2 \tilde{\phi}_b^2 \frac{2^4}{3} \Delta \tilde{\phi}^6 + O(\Delta \tilde{\phi}^7). \quad (79)
\]

The above show that, when \(\Delta \tilde{\phi} \to 0\), the rate of convergence of (75) – (77) to 0 is faster than that of (78) and (79). Therefore, we have the following:

\[
E \left[ \left( \sum_{t=1}^{N} \tilde{\phi}_k(t) e(\tilde{\phi}_1(t)) \right)^2 \right] \to_{\Delta y_{\text{max}} \to 0} N \ E \left[ \tilde{\phi}_k^2 e^2(\tilde{\phi}_1) \right].
\]

\[
\square
\]

Derivation of eq. (28)

\[
(24)/N \overset{(i)}{=} \int \sigma^2(\tilde{\phi}_1) e^2(\tilde{\phi}_1) f(\tilde{\phi}_1) d\tilde{\phi}_1 = \sum_j \int_{S_{\tilde{\phi}}^j} \sigma^2(\tilde{\phi}_1) g_j^{-1}(\tilde{\theta}_1 \tilde{\phi}_1)^2 \cdot \sigma^2(\tilde{\phi}_1) f(\tilde{\phi}_1) d\tilde{\phi}_1
\]

\[
= \sum_j \int_{S_{\tilde{\phi}}^j} \tilde{\theta}_1(\tilde{\phi}_1)^2 \cdot \sigma^2(\tilde{\phi}_1) f(\tilde{\phi}_1) d\tilde{\phi}_1
\]

\[
\overset{(i)}{=} \sum_j \int_{S_{\tilde{\phi}}^j} \tilde{\theta}_1(\tilde{\phi}_1)^2 \cdot \sigma^2(\tilde{\phi}_1) f(\tilde{\phi}_1) d\tilde{\phi}_1 + O(\Delta \tilde{\phi}^3)
\]

\[
= \tilde{\theta}_1^2 \sum_j \int_{S_{\tilde{\phi}}^j} \tilde{\theta}_1(\tilde{\phi}_1)^2 \cdot \sigma^2(\tilde{\phi}_1) f(\tilde{\phi}_1) d\tilde{\phi}_1 + O(\Delta \tilde{\phi}^3)
\]

\[
\overset{(ii)}{=} \tilde{\theta}_1^2 \sum_j \int_{S_{\tilde{\phi}}^j} \tilde{\theta}_1(\tilde{\phi}_1)^2 \cdot \sigma^2(\tilde{\phi}_1) f(\tilde{\phi}_1) d\tilde{\phi}_1 + O(\Delta \tilde{\phi}^3)
\]

\[
\overset{(iii)}{=} \tilde{\theta}_1^2 \sum_j \int_{S_{\tilde{\phi}}^j} \tilde{\theta}_1(\tilde{\phi}_1)^2 \cdot \sigma^2(\tilde{\phi}_1) f(\tilde{\phi}_1) d\tilde{\phi}_1 + O(\Delta \tilde{\phi}^3)
\]

where \((\tilde{\phi}_1)_{(j)}^j\) is the midpoint of \(S_{\tilde{\phi}}^j\). (i) is by Assumption 3.2.1, (ii) is by Assumption 3.2.2, and (iii) is by Assumption 3.2.1.

\[
\square
\]

Proof of Theorem 3.1

The optimal solution can be given by using a similar technique to that in [1, 15]. With the calculus of variations, the following Euler–Lagrange equation:

\[
\frac{d}{d\tilde{\phi}_1} \left( \frac{\partial F}{\partial \tilde{\phi}_1} \right) - \frac{\partial F}{\partial G} = 0,
\]

where

\[
G(\tilde{\phi}_1) := \int_{-\infty}^{\tilde{\phi}_1} g(\tilde{\phi}_1) d\tilde{\phi}_1,
\]

gives a differential equation:

\[
\frac{d}{d\tilde{\phi}_1} \left( -2g(\tilde{\phi}_1)^{-3} \sigma^2(\tilde{\phi}_1) f(\tilde{\phi}_1) \right) = 0,
\]

and the solution is:

\[
g(\tilde{\phi}_1) = K \sigma^3(\tilde{\phi}_1) f(\tilde{\phi}_1), \quad K : \text{constant.}
\]
The constant number $K$ is directly calculated by the condition (31), and the value of the objective function is derived as follows.

$$
\int \mathcal{F}(g(\hat{\phi}_1))d\hat{\phi}_1 = \int \frac{1}{12} \hat{\partial}^2_1 (K g^\frac{1}{2} (\hat{\phi}_1) f^\frac{1}{2} (\hat{\phi}_1))^{-2} \sigma^2(\hat{\phi}_1) f(\hat{\phi}_1) d\hat{\phi}_1
$$

$$
= \int \frac{1}{12} \hat{\partial}^2_1 K^{-2} \sigma^2(\hat{\phi}_1) f^\frac{1}{2} (\hat{\phi}_1) d\hat{\phi}_1 = \frac{1}{12} \hat{\partial}^2_1 K^{-2} D = \frac{1}{12} \hat{\partial}^2_1 D^3 M^{-2}
$$

Proof of Theorem 3.2

We use a similar technique to that in [11, 2]. Let $\lambda$ be a Lagrange multiplier and consider the minimization of the following quantity.

$$
\int \mathcal{F}(g(\hat{\phi}_1))d\hat{\phi}_1 + \lambda H_d(f, g) = \int \frac{1}{12} \hat{\partial}^2_1 \left( \frac{1}{g(\hat{\phi}_1)} \right)^2 \sigma^2(\hat{\phi}_1) f(\hat{\phi}_1) - \lambda f(\hat{\phi}_1) \log \left( g^{-1}(\hat{\phi}_1) \right) d\hat{\phi}_1 + \lambda H_d(f)
$$

By applying the calculus of variations, we obtain:

$$
\frac{\partial}{\partial g} \left( g^{-2} \sigma^2(\hat{\phi}_1) + \lambda \log g \right) = -2g^{-3} \sigma^2(\hat{\phi}_1) + \lambda g^{-1} = \text{constant}.
$$

Fix the constant to be zero, then,

$$
g = \left( \frac{2}{\lambda} \right)^{\frac{1}{2}} \sigma(\hat{\phi}_1),
$$

and by substituting this for $H(f, g)$, we obtain:

$$
H(f, g) = \int -f \log g^{-1} d\hat{\phi}_1 = \log \left( \frac{2}{\lambda} \right)^{\frac{1}{2}} + \int -f \log \frac{f}{\sigma(\hat{\phi}_1)} d\hat{\phi}_1 = \log M.
$$

Therefore,

$$
\left( \frac{2}{\lambda} \right)^{\frac{1}{2}} = \exp \left( \int f \log \frac{f}{\sigma(\hat{\phi}_1)} d\hat{\phi}_1 + \log M \right),
$$

and (41) is derived. By substituting $g_\sigma$ for the objective integral, the following is derived.

$$
\int \frac{1}{12} \hat{\partial}^2_1 g^{-2}(\hat{\phi}_1) \sigma^2(\hat{\phi}_1) f(\hat{\phi}_1) d\hat{\phi}_1 = \frac{1}{12} \hat{\partial}^2_1 \left( \frac{\lambda}{2} \right) = \frac{1}{12} \hat{\partial}^2_1 K^{-2} M^{-2}
$$

Proof of Lemma 4.1

The left hand side of (51) is extended:

$$
E \left[ \left( \sum_{t=1}^{N} \hat{\phi}_k(t) e(\hat{\phi}_1(t)) \right)^2 \right] = E \left[ \sum_{t=1}^{N} \hat{\phi}_k(t) e^2(\hat{\phi}_1(t)) \right] + 2E \left[ \sum_{t=1}^{N-1} \hat{\phi}_k(t) e(\hat{\phi}_1(t)) \hat{\phi}_k(t+1) e(\hat{\phi}_1(t+1)) \right] + \cdots
$$

$$
= N E \left[ \hat{\phi}_k^2(\hat{\phi}_1) \right] + 2(N-1)E \left[ \hat{\phi}_k(\hat{\phi}_1) \hat{\phi}_{k+1}(\hat{\phi}_1) e(\hat{\phi}_2) \right] + \cdots. \quad (80)
$$

In (80), terms of the form $E \left[ \hat{\phi}_a e(\hat{\phi}_b) \hat{\phi}_c e(\hat{\phi}_d) \right]$ appear and in general, when (49) and (50) are satisfied, $E \left[ \hat{\phi}_a e(\hat{\phi}_b) \hat{\phi}_c e(\hat{\phi}_d) \right]$ can be calculated according to the combinations of $a$, $b$, $c$ and $d$ as follows.
When \( a \neq b \neq c \neq d \),
\[
E \left[ \tilde{\phi}_a e(\tilde{\phi}_b) \tilde{\phi}_c e(\tilde{\phi}_d) \right] = \int \tilde{\phi}_a e(\tilde{\phi}_b) \tilde{\phi}_c e(\tilde{\phi}_d) f(\tilde{\phi}_a, \tilde{\phi}_b, \tilde{\phi}_c, \tilde{\phi}_d) d\tilde{\phi}_a d\tilde{\phi}_b d\tilde{\phi}_c d\tilde{\phi}_d \\
= \int e(\tilde{\phi}_b) \tilde{\phi}_c e(\tilde{\phi}_d) \left( \int \tilde{\phi}_a f(\tilde{\phi}_a, \tilde{\phi}_b, \tilde{\phi}_c, \tilde{\phi}_d) d\tilde{\phi}_a \right) d\tilde{\phi}_b d\tilde{\phi}_c d\tilde{\phi}_d \quad (49) = \int e(\tilde{\phi}_b) \tilde{\phi}_c e(\tilde{\phi}_d) \times 0 \times d\tilde{\phi}_b d\tilde{\phi}_c d\tilde{\phi}_d = 0,
\]
and similarly, when \( a = b \neq c \neq d \),
\[
E \left[ \tilde{\phi}_a e(\tilde{\phi}_b) \tilde{\phi}_c e(\tilde{\phi}_d) \right] = \int \tilde{\phi}_a e(\tilde{\phi}_b) \tilde{\phi}_c e(\tilde{\phi}_d) f(\tilde{\phi}_a, \tilde{\phi}_b, \tilde{\phi}_c, \tilde{\phi}_d) d\tilde{\phi}_a d\tilde{\phi}_b d\tilde{\phi}_c d\tilde{\phi}_d \\
= \int \tilde{\phi}_a e(\tilde{\phi}_a) e(\tilde{\phi}_d) \left( \int \tilde{\phi}_c f(\tilde{\phi}_a, \tilde{\phi}_b, \tilde{\phi}_c, \tilde{\phi}_d) d\tilde{\phi}_c \right) d\tilde{\phi}_a d\tilde{\phi}_b d\tilde{\phi}_d \quad (49) = \int \tilde{\phi}_a e(\tilde{\phi}_a) e(\tilde{\phi}_d) \times 0 \times d\tilde{\phi}_a d\tilde{\phi}_d = 0,
\]
and when \( a = b \neq c = d \),
\[
E \left[ \tilde{\phi}_a e(\tilde{\phi}_b) \tilde{\phi}_c e(\tilde{\phi}_d) \right] = \int \tilde{\phi}_a e(\tilde{\phi}_b) \tilde{\phi}_c e(\tilde{\phi}_d) f(\tilde{\phi}_a, \tilde{\phi}_b, \tilde{\phi}_c) d\tilde{\phi}_a c_{\tilde{\phi}_c} d\tilde{\phi}_d \\
= \int \tilde{\phi}_a e(\tilde{\phi}_a) e(\tilde{\phi}_d) \left( \int \tilde{\phi}_c f(\tilde{\phi}_a, \tilde{\phi}_b, \tilde{\phi}_c) d\tilde{\phi}_c \right) d\tilde{\phi}_a c_{\tilde{\phi}_d} (50) = \int \tilde{\phi}_a e(\tilde{\phi}_a) e(\tilde{\phi}_d) \times 0 \times d\tilde{\phi}_a = 0.
\]
On the other hand, there is no term when \( a = c \neq b \neq d \) or \( b = d \neq a \neq c \) in (80). Finally, when \( a = c, b = d \),
\[
E \left[ \tilde{\phi}_a e(\tilde{\phi}_b) \tilde{\phi}_c e(\tilde{\phi}_d) \right] = E \left[ \tilde{\phi}_a^2 e^2(\tilde{\phi}_b) \right].
\]
The other cases are essentially equivalent to one of the above cases (for example, \( a = d \neq b \neq c \) is equivalent to \( a = b \neq c \neq d \)).

From the above, it follows that:
\[
E \left[ \left( \sum_{i=1}^{N} \tilde{\phi}_k(t) e(\tilde{\phi}_1(t)) \right)^2 \right] = NE \left[ \tilde{\phi}_1^2 e^2(\tilde{\phi}_1) \right].
\]

\( \square \)

**Proof of Proposition 4.1**

Consider \( \mathcal{S}^y = (0, d_1] \) (equivalently \( \mathcal{S}^{\tilde{\phi}_1} \) on \( \tilde{\phi}_1 \)) and \( \mathcal{S}^y_2 = (d_1, d_2] \) (equivalently \( \mathcal{S}^{\tilde{\phi}_2} \) on \( \tilde{\phi}_2 \)) where their boundaries \( d_1, d_2 \) have the relationship:
\[
d_1 = r_1 d_2, \quad r_1 \in [0, 1] \quad (81)
\]
with an appropriate ratio \( r_1 \). The quantized values \( y_{y}(1) \) and \( y_{y}(2) \) for the subsections \( \mathcal{S}^y \) on \( y \) (or \( \mathcal{S}^{\tilde{\phi}_1} \) on \( \tilde{\phi}_1 \)) and \( \mathcal{S}^y_2 \) (or \( \mathcal{S}^{\tilde{\phi}_2} \)) satisfying the bias-free condition:
\[
E_{\mathcal{S}^{\tilde{\phi}_j}} \left[ \tilde{\phi}_1 \cdot e(\tilde{\phi}_1) \right] = 0, \quad j = 1, 2
\]
are given as follows. Let \( y_{y}(1) = \frac{d_1}{2} + h_1 \), where \( h_1 \) is an offset from the center of \( \mathcal{S}^{y}_{1} \), then,
\[
E_{\mathcal{S}^{\tilde{\phi}_1}} \left[ \tilde{\phi}_1 \cdot e(\tilde{\phi}_1) \right] = \int_{-k_1}^{k_1} \left( \frac{r_1 d_2}{2} + z \right) (z - h_1) \frac{1}{2\kappa_y} dz = -\frac{1}{2\kappa_y} \left( \frac{2}{3} k_1^3 - r_1 d_2 h_1 k_1 \right), \quad k_1 := \frac{d_1}{2},
\]
and therefore,
\[
h_1 = \frac{2}{3} \frac{k_1^2}{r_1 d_2} = \frac{1}{6} r_1 d_2.
\]
Similarly, let \( y_{y}(2) := \frac{(1 + r_1) d_2}{2} + h_2 \), where \( h_2 \) is the offset, then,
\[
E_{\mathcal{S}^{\tilde{\phi}_2}} \left[ \tilde{\phi}_1 \cdot e(\tilde{\phi}_1) \right] = \int_{-k_2}^{k_2} \left( \frac{d_2 + r_1 d_2}{2} + z \right) (z - h_2) \frac{1}{2\kappa_y} dz = -\frac{1}{2\kappa_y} \left( \frac{2}{3} k_2^3 - (d_2 + r_1 d_2) h_2 k_2 \right), \quad k_2 := \frac{d_2 (1 - r_1)}{2},
\]
and therefore,
\[ h_2 = \frac{2}{3} k_2^2 \frac{1}{d_2 (1 + r_1)} = \frac{1}{6} \frac{(1 - r_1)^2}{(1 + r_1)} d_2. \]

By using these \( y'_{(1)} \) and \( y'_{(2)} \), the variances of \( \tilde{\phi}_1 e(\tilde{\phi}_1) \) in each subsection can be calculated as follows. Let \( V_{S^y_1} \left[ \sigma(\tilde{\phi}_1)e(\tilde{\phi}_1) \right] \)
denote the quantity:

\[
V_{S^y_1} \left[ \sigma(\tilde{\phi}_1)e(\tilde{\phi}_1) \right] := \int_{S^y_1} \sigma^2(\tilde{\phi}_1)e^2(\tilde{\phi}_1)f(\tilde{\phi}_1)d\tilde{\phi}_1,
\]
where
\[
\sigma^2(\tilde{\phi}_1) = \phi_1^2 + \frac{1}{3} \kappa_y^2(n - 1),
\]
then, for even \( M \),

\[
V_{S^y_1} \left[ \sigma(\tilde{\phi}_1)e(\tilde{\phi}_1) \right] = \int_{-k_1}^{k_1} \left( \frac{r_1 d_2}{2} + z \right)^2 + \frac{1}{3} \kappa_y^2(n - 1) \right) (z - h_1)^2 \frac{1}{2\kappa_y} dz
\]
\[= \frac{1}{2160} \frac{1}{2\kappa_y} d_2^2 \left( 32r_1^5 \right) + \frac{1}{27} \frac{1}{2\kappa_y} \kappa_y^2(n - 1)d_2^3 r_1^3 \]

and similarly

\[
V_{S^y_2} \left[ \sigma(\tilde{\phi}_1)e(\tilde{\phi}_1) \right] = \int_{-k_2}^{k_2} \left( \frac{d_2(1 + r_1)}{2} + z \right)^2 + \frac{1}{3} \kappa_y^2(n - 1) \right) (z - h_2)^2 \frac{1}{2\kappa_y} dz
\]
\[= \frac{1}{2160} \frac{1}{2\kappa_y} d_2^2 \left( 32r_1^5 - 18(1 - r_1)^5 + 45(1 + r_1)^2(1 - r_1)^3 + 5(1 - r_1)^7(1 + r_1)^{-2} \right)
\]
\[+ \frac{1}{108} \frac{1}{2\kappa_y} \kappa_y^2(n - 1)d_2^3 \left( 3(1 - r_1)^3 + \frac{(1 - r_1)^5}{(1 + r_1)^2} \right) \]

Therefore, the sum of \( V_{S^y_1} \left[ \sigma(\tilde{\phi}_1)e(\tilde{\phi}_1) \right] \) and \( V_{S^y_2} \left[ \sigma(\tilde{\phi}_1)e(\tilde{\phi}_1) \right] \) is:

\[
V_{S^y_1} \left[ \sigma(\tilde{\phi}_1)e(\tilde{\phi}_1) \right] + V_{S^y_2} \left[ \sigma(\tilde{\phi}_1)e(\tilde{\phi}_1) \right] = \frac{1}{2160} \frac{1}{2\kappa_y} \left( d_2^2 \psi(r_1; 32) + 20\kappa_y^2(n - 1)d_2^3 \xi(r_1; 4) \right),
\]
\[
\psi(r_1; 32) := 32r_1^5 - 18(1 - r_1)^5 + 45(1 + r_1)^2(1 - r_1)^3 + 5(1 - r_1)^7(1 + r_1)^{-2},
\]
\[
\xi(r_1; 4) := 4r_1^3 + 3(1 - r_1)^3 + \frac{(1 - r_1)^5}{(1 + r_1)^2}. \]

The minimizer \( r_1^* \) of this sum is given by:

\[
r_1^* = \arg \min_{r_1 \in [0, 1]} \left( d_2^2 \psi(r_1; 32) + 20\kappa_y^2(n - 1)d_2^3 \xi(r_1; 4) \right),
\]
\[
\psi_1^{\min} := \psi(r_1^*; 32),
\]
\[
\xi_1^{\min} := \xi(r_1^*; 4),
\]
and

\[
\left( V_{S^y_1} \left[ \sigma(\tilde{\phi}_1)e(\tilde{\phi}_1) \right] + V_{S^y_2} \left[ \sigma(\tilde{\phi}_1)e(\tilde{\phi}_1) \right] \right) \bigg|_{r_1 = r_1^*} = \frac{1}{2160} \frac{1}{2\kappa_y} \left( d_2^2 \psi_1^{\min} + 20\kappa_y^2(n - 1)d_2^3 \xi_1^{\min} \right). \]

Note that the optimal \( r_1^* \) is independent of the value of \( d_2 \), which is the upper boundary of \( S^y \).

Next, we successively consider another subsection \( S^y_3 \) on \( y \) (or \( S^y_3^{\hat{1}} \) on \( \tilde{\phi}_1 \)) together with \( S^y_1 \) (or \( S^y_1^{\hat{1}} \)) and \( S^y_2 \) (or \( S^y_2^{\hat{1}} \)). Assume the relation between \( d_2 \) and \( d_3 \) is:

\[ d_2 = r_2 d_3, \]

29
where $r_2$ is an appropriate number in $[0, 1]$. Similar to the case of $S^y_1$ and $S^y_2$, the offset $h_3$ of $y^{(3)}_i$ for the subsection $S^y_3$ on $y$ (or $S^\phi_3$ on $\tilde{\phi}_1$) satisfying $E_{S^\phi_3} \left[ \tilde{\phi}_1 e(\tilde{\phi}_1) \right] = 0$ is

$$h_3 = \frac{2}{3} k_3^2 \frac{1}{d_3(1 + r_2)} = \frac{1}{6} \frac{(1 - r_2)^2}{(1 + r_2)} d_3, \quad k_3 := \frac{d_3(1 - r_2)}{2}$$

and $V_{S^\phi_3} \left[ \sigma(\tilde{\phi}_1)e(\tilde{\phi}_1) \right]$ can be given as

$$V_{S^\phi_3} \left[ \sigma(\tilde{\phi}_1)e(\tilde{\phi}_1) \right] = \int_{-k_3}^{k_3} \left( \frac{d_3(1 + r_2)}{2} + z \right)^2 \frac{1}{3} \kappa_y^2(n - 1) \left( z - h_3 \right)^2 \frac{1}{2 \kappa_y} dz$$

$$= \frac{1}{2160} \frac{1}{2 \kappa_y} d_3^3 \left\{ -18(1 - r_2)^5 + 45(1 + r_2)^2(1 - r_2)^3 + 5(1 - r_2)^7(1 + r_2)^2 \right\}$$

$$+ \frac{1}{108} \frac{1}{2 \kappa_y} \kappa_y^2(n - 1) d_3^3 \left\{ 3(1 - r_2)^3 + \frac{(1 - r_2)^5}{(1 + r_2)^2} \right\}$$

Therefore, the optimal $r_2^o$ that minimizes $V_{S^\phi_1} \left[ \sigma(\tilde{\phi}_1)e(\tilde{\phi}_1) \right] + V_{S^\phi_2} \left[ \sigma(\tilde{\phi}_1)e(\tilde{\phi}_1) \right] + V_{S^\phi_3} \left[ \sigma(\tilde{\phi}_1)e(\tilde{\phi}_1) \right]$ is found by solving the following minimization problem:

$$r_2^o := \arg \min_{r_2} \left( V_{S^\phi_1} \left[ \sigma(\tilde{\phi}_1)e(\tilde{\phi}_1) \right] + V_{S^\phi_2} \left[ \sigma(\tilde{\phi}_1)e(\tilde{\phi}_1) \right] + V_{S^\phi_3} \left[ \sigma(\tilde{\phi}_1)e(\tilde{\phi}_1) \right] \right)$$

$$\psi(r_2; \psi_1^{\min}) := \psi_1^{\min} r_2^o - 18(1 - r_2)^5 + 45(1 + r_2)^2(1 - r_2)^3 + 5(1 - r_2)^7(1 + r_2)^2,$$

$$\xi(r_2; \xi_1^{\min}) := \xi_1^{\min} r_2^o + 3(1 - r_2)^3 + \frac{(1 - r_2)^5}{(1 + r_2)^2}.$$  \hspace{1cm} (84)

By repeating the above process, we obtain the result. \hfill \square

**Lemma A.1** \hspace{0.5cm} A rational function

$$\psi(r) := \alpha r^5 - 18(1 - r)^5 + 45(1 + r)^2(1 - r)^3 + 5(1 - r)^7(1 + r)^2$$

has only one local minimum in $r \in (0, 1)$ when $\alpha > 0$.

Refer to [24] for the proof.

**Slutsky’s theorem**

$$\lim_{i \to \infty} [X(i)^{-1} Y(i)] = (\lim_{i \to \infty} [X(i)])^{-1} \lim_{i \to \infty} [Y(i)]$$

subject to that $\lim_{i \to \infty} [X(i)]$ and $\lim_{i \to \infty} [Y(i)]$ exist.

**Proof of Lemma 4.2**

From Lemma A.1, it is known that $\psi(r, \psi_0^{\min} = 32)$ has only one local minimum in $r \in (0, 1)$. Moreover, from

$$\psi(0; \alpha) = 32, \quad \forall \alpha > 0, \quad \psi(1; \psi_1^{\min}) = \psi_1^{\min}, \quad \psi_0^{\min} = 32,$$

the minimum value $\psi_1^{\min}$ satisfies

$$\psi_1^{\min} < 32.$$  \hspace{1cm}

Next, $\psi(r; \psi_1^{\min})$ satisfies

$$\psi(0; \psi_1^{\min}) = 32, \quad \psi(1; \psi_1^{\min}) = \psi_1^{\min} < 32,$$
and also \( \psi(r; \psi_1^{\min}) \) has only one local minimum in \( r \in (0, 1) \). This means

\[
\psi_1^{\min} > \psi_2^{\min}.
\]

The difference between \( \psi(r; \psi_0^{\min}) \) and \( \psi(r; \psi_1^{\min}) \) is only the coefficient of the term \( r^\phi \) and \( r^\delta \) is a strictly increasing function in \( (0, 1] \). Therefore, with \( \psi_0^{\min} > \psi_1^{\min} \),

\[
r_1^\phi < r_2^\phi < 1.
\]

By repeating the same process, we finally obtain:

\[
r_1^\phi < r_2^\phi < r_3^\phi < \cdots < 1.
\]

Next to show \( \lim_{j \to \infty} r_j^\phi = 1 \). Let \( \lim_{j \to \infty} r_j^\phi = r_\infty \). Then, \( r_\infty \) satisfies:

\[
r_\infty := \arg\min_{r \in [0, 1]} \psi(r; \psi_\infty^{\min})
\]

\[
\psi_\infty^{\min} = \psi(r_\infty; \psi_\infty^{\min}).
\]

Note that if \( \psi_\infty^{\min} > 0 \), \( \psi(r; \psi_\infty^{\min}) \) also has only one local minimum in \( r \in (0, 1) \). On the other hand, when \( \psi_\infty^{\min} = 0 \), it is also known that \( \psi(r; \psi_\infty^{\min}) \) is a decreasing function in \( r \in [0, 1] \) from the proof of Lemma A.1 and \( \min_r \psi(r; \psi_\infty^{\min}) = \psi(1; \psi_\infty^{\min}) \). From (56), \( \psi(1; \psi_\infty^{\min}) = \psi_\infty^{\min} \), and the minimum is at \( r = 1 \). This means \( r_\infty = 1 \) (and \( \psi_\infty^{\min} = 0 \)). \[ \Box \]

**Proof of Lemma 4.3**

On the subsections \( S_j^{\phi_1} (S_j^\delta) \) and \( S_{j+1}^{\phi_1} (S_{j+1}^\delta) \), i.e., the general case for (81) – (84), from:

\[
\int_{-k_j}^{k_j} \left( \frac{d_j + d_{j+1}}{2} + z \right) (z - h_j) \, dz = \frac{2}{3} k_j^3 - (d_j + d_{j+1})h_j k_j,
\]

the offsets \( h_j \) and \( h_{j+1} \) such that \( E_{S_j^{\phi_1}} [\tilde{\phi}_1 e(\tilde{\phi}_1)] = 0 \) and \( E_{S_{j+1}^{\phi_1}} [\tilde{\phi}_1 e(\tilde{\phi}_1)] = 0 \) are given by:

\[
h_j = \frac{2}{3} d_j + d_{j+1} k_j^2, \quad k_j := \frac{d_{j+1} - d_j}{2}, \quad h_{j+1} = \frac{2}{3} d_{j+1} + d_{j+2} k_{j+1}^2, \quad k_{j+1} := \frac{d_{j+2} - d_{j+1}}{2}.
\]

On the other hand, \( V_{S_j^{\phi_1}} [\tilde{\phi}_1 e(\tilde{\phi}_1)] \) is calculated by:

\[
V_{S_j^{\phi_1}} [\tilde{\phi}_1 e(\tilde{\phi}_1)] = \int_{-k_j}^{k_j} \left( \frac{d_j + d_{j+1}}{2} + z \right)^2 (z - h_j)^2 \, dz = A (d_{j+1} - d_j)^5 + B (d_j + d_{j+1})^2 (d_{j+1} - d_j)^3,
\]

where

\[
A := \frac{1}{5} \cdot \frac{1}{2^4} - \frac{1}{3^2} \cdot \frac{1}{2^4} < 0, \quad B := \frac{1}{3} \cdot \frac{1}{2^4} > 0.
\]

Therefore:

\[
V_{S_j^{\phi_1}} [\tilde{\phi}_1 e(\tilde{\phi}_1)] + V_{S_{j+1}^{\phi_1}} [\tilde{\phi}_1 e(\tilde{\phi}_1)] = A (d_{j+1} - d_j)^5 + B (d_{j+1} + d_j)^2 (d_{j+1} - d_j)^3 + A (d_{j+2} - d_{j+1})^5 + B (d_{j+2} + d_{j+1})^2 (d_{j+2} - d_{j+1})^3 =: Z(d_{j+1}).
\]

(85)

For given \( d_j \) and \( d_{j+2} \), consider which side the minimum point of \( Z(d_{j+1}) \) is on from the center of \( d_j \) and \( d_{j+2} \). From \( A < 0 \) and \( B > 0 \) and the symmetric structure of \( Z(d_{j+1}) \), except for the terms \( (d_{j+1} + d_j)^2 \) and \( (d_{j+2} + d_{j+1})^2 \) where
that and with (88), we obtain:

This contradicts the assumption employing the Taylor series expansion, then by Chebyshev’s inequality, we obtain:

Proof of Theorem 5.1

From Lemma 4.2 and its proof, it is known that when \( j \to \infty \), \( r^o_j \) and \( \psi_j^{\min} \) converge to 1 and 0, respectively. Therefore, by employing the Taylor series expansion, \( \psi(r; \psi_j^{\min}) \) can be represented by:

\[
\psi(r; \psi_j^{\min}) = \psi_{j-1}^{\min}(1 - 5(1 - r) + 10(1 - r)^2 - 10(1 - r)^3) + 45 \cdot 2^2(1 - r)^3 + O((1 - r)^4)
\]

near \( r = 1 \) at sufficiently large \( j \). By applying a variable transformation \( 1 - r = \epsilon \), we obtain

\[
\psi(\epsilon; \psi_j^{\min}) = \psi_{j-1}^{\min}(1 - 5\epsilon + 10\epsilon^2 - 10\epsilon^3) + 180\epsilon^3 + O(\epsilon^4)
\]

at \( \epsilon \to 0 \). Denote the local minimum of \( \psi(\epsilon; \psi_j^{\min}) \) as \( \epsilon_j \), then \( \epsilon_j \) must satisfy:

\[
\psi_{j-1}^{\min}(-5 + 20\epsilon_j - 30\epsilon_j^2) = 540\epsilon_j^3 + O(\epsilon_j^3) = 0.
\]

From (87), it is simple to verify that:

\[
\epsilon_j = \left( \frac{1}{108} \psi_{j-1}^{\min} \right)^{1/2} + o\left( \left( \psi_{j-1}^{\min} \right)^{1/2} \right)
\]

at \( \psi_{j-1}^{\min} \to 0 \). On the other hand, from (86), \( \psi_j^{\min} \) is represented by:

\[
\psi_j^{\min} = \psi_{j-1}^{\min}(1 - 5\epsilon_j + 10\epsilon_j^2 - 10\epsilon_j^3) + 180\epsilon_j^3 + O(\epsilon_j^3),
\]

and with (88), we obtain:

\[
\psi_j^{\min} - \psi_{j-1}^{\min} = -5\left( \frac{1}{108} \right)^{1/2} \psi_{j-1}^{3/2} + 180 \left( \frac{1}{108} \right)^{3/2} \psi_{j-1}^{3/2} + O(\psi_{j-1}^{3/2})
\]

With the convergence \( \psi_j^{\min} \to 0 \), we derive the statement of the lemma.

Lemma A.2 \( \tilde{\psi}(m) \geq \hat{\psi}(m) \) at \( m = 0, 1, \ldots \), when \( \hat{\psi}(0) \geq \hat{\psi}(0) \).

Proof First define \( \hat{\psi}(m) \) for \( m \in \mathcal{R} \), which is a simple linear interpolation of \( \hat{\psi}(m) \) at \( m = 0, 1, \ldots \), and the gradient between \( \hat{\psi}'(m - 1) \) and \( \hat{\psi}'(m) \) (\( m = 1, 2, \ldots \)) is a constant \( \mathcal{P}(\hat{\psi}'(m - 1)) = a\hat{\psi}'(m - 1) + o(\hat{\psi}'(m - 1)) \) \((< 0) \). Assume that \( \hat{\psi}(m) \) crosses \( \hat{\psi}'(m) \) downward at \( m = m' \) between \( m - 1 \) and \( m \). Note that \( \hat{\psi}(m') < \hat{\psi}'(m - 1) = \hat{\psi}(m - 1) \), therefore,

\[
\frac{d\hat{\psi}(m)}{dm} \bigg|_{m=m'} = (a + \nu)\hat{\psi}'(m') \geq \mathcal{P}(\hat{\psi}(m')) > \mathcal{P}(\hat{\psi}'(m - 1)) = a\hat{\psi}'(m - 1) + o(\hat{\psi}'(m - 1)).
\]

This contradicts the assumption \( \hat{\psi}(m') \) crosses \( \hat{\psi}'(m') \) downward.

Proof of Theorem 5.1

First evaluate the magnitude of \( \tilde{U}^T E \). From (28), (29), and (36),

\[
E \left[ \tilde{U}^T E \right] = 0, \quad \mathbb{V} \left[ \tilde{U}^T E \right] = \frac{1}{12} \beta_1 \beta_2 D^3 M^{-2} N.
\]

Then by Chebyshev’s inequality, we obtain:

\[
\text{Prob} \left( \| \tilde{U}^T E \|_{\infty} \geq \sqrt{\frac{n}{\beta_2} \frac{1}{12} \beta_1 \beta_2 D^3 M^{-2} N} \right) \leq \beta_2,
\]

32
for a reliability index $\beta_2$. Combine $(\tilde{U}^T\tilde{U})^{-1}$ and $\tilde{U}^T E$ using the norm inequality:

$$\| (\tilde{U}^T\tilde{U})^{-1}\tilde{U}^T E\|_\infty \leq \|(\tilde{U}^T\tilde{U})^{-1}\|_1 \|\tilde{U}^T E\|_\infty,$$

and this gives:

$$\text{Prob}\left(\|(\tilde{U}^T\tilde{U})^{-1}\tilde{U}^T E\|_\infty \leq \epsilon_1 \epsilon_2\right) \geq \text{Prob}\left(\|(\tilde{U}^T\tilde{U})^{-1}\|_1 \leq \epsilon_1 \text{ and } \|\tilde{U}^T E\|_\infty \leq \epsilon_2\right).$$

Therefore we have proved the statements. \qed