MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

Analysis on Optimal Quantization of Signals for System Identification and the Effect of Noise

Koji TSUMURA

(Communicated by Kazuo Murota)

METR 2005–04 January 2005

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF TOKYO
BUNKYO-KU, TOKYO 113-8656, JAPAN

WWW page: http://www.i.u-tokyo.ac.jp/mi/mi-e.htm

Analysis on Optimal Quantization of Signals for System Identification and the Effect of Noise *

Koji Tsumura† and Jan Maciejowski‡

METR 2005–04 January 31, 2005

Abstract: In this paper, we analyse the property of the optimal quantization of signals used for system identification. We deal with memoryless quantization for output signals and consider to derive optimal quantization schemes for minimizing the errors of parameter estimation given by least squares method under a constraint on the number of subsections of the quantized signals or the expectation of the optimal code length in general resolution case or high resolution case. In the case of general resulution of quantizer and a kind of uniform distribution of input signals, the optimal quantizer can be given by solving a minimization of a special 1-dimensional rational function recursively. This quantizer has the property that it is coarse around the origin of its input and goes to be dense apart from the origin. On the other hand, the optimal quantizer of high resolution can be given by solving Eular–Lagrange's equations and the solutions are represented as a simple function of the distribution density of the regressor vector. We show examples of solutions for several cases of the distribution density of the regressor vectors and discuss their meanings with respect to the feasibility of parameter estimations. Moreover, in the case of the constraint of code length, the necessary information to attain the optimal identification errors is given as a function of the entropy of the regressor vector.

Keywords: system identification, quantization, least squares method, MA model, entropy

1 Introduction

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

The problem of quantization of signals itself has a long history from the 1940s and one of main themes in the area of information theory (e.g. see [11]). The purpose of the problem is to attain low distortion between the original signals and the quantized ones under constraints on the amount of information. Of course, the situations and the objective for data transmission and for control systems are essentially different and the necessity of the research for the latter case has been recognized for a long time. However, although we can see elementary discussion in the control community from the 70s (e.g. see [5]), the strict analysis began at the late 80s. The main difficulty of quantization problem in control systems should be in their dynamics and the result by [6, 7] is recognized as a break through, in which papers the behaviour of control systems, and their stability or state estimation, are analysed in detail. Then, in the last few years, stabilization problems of quantized systems have been actively investigated for several different situations, e.g., [21, 22, 3, 14, 8, 15, 19]. Among them, a logarithmic quantizer was shown to be coarsest in some sense to attain a kind of

*This paper is the revised version of the technical reports/conference paper [18, 17] including new results.

†Koji Tsumura is with Department of Information Physics and Computing, The University of Tokyo, Hongo 7–3–1, Bunkyo-ku, Tokyo 113–8656, Japan, tel: +81–3–5841–6891, fax: +81–3–5841–6886, e-mail: tsumura@i.u-tokyo.ac.jp

‡Jan Maciejowski is with Department of Engineering, The University of Cambridge, Trumpington Street, Cambridge CB2 1PZ tel: +44–1223–332732, e-mail: jmm@eng.cam.ac.uk

asymptotic stability [8] and it reveals the difference of importance on signals depending on its magnitudes and the directions in the signal space from the view point of controlling systems.

With this background, our interests naturally grow into system identification problem; that is, a question: what quantizer is *optimal* for system identification? We expect the answer to this question will clarify the relationship between system identification and stabilization from the point of view of the information of signals. However, compared to the research activity in the stabilization or estimation problem, unfortunately, the quantization problem for system identification [9] has not been adequately considered. From such point of view, we dealt with this problem.

In this paper, we consider optimal quantization problem of output signals which are used for parameter estimation. The identified system is a simple SISO MA model in order to reveal the essential property of the optimal quantization in system identification and assist intuitive understanding of it. The *optimality* we mean in this paper is to minimize the variance of the error of the parameter estimation given by least squares method under a constraint of the number of quantization steps or the expectation of the code length when the quantized signals are coded by an optimal coder. We consider this problem for the cases of general resolution and high resulution of quantization. The difficulty of the problem is in the complex correlation between the input signals and the quantization errors and managing it is a key for solving the optimization problem.

In the general resolution case (Section 3), we give the optimal quantizer under a problem settings of a kind of uniform distribution of input signals. The optimal quantizer is given by solving a minimization of some special 1-dimensional rational function recursively. The optimal quantization is not uniform and it is coarse around the origin of the quantized signals and goes to be dense apart from it. This result shows an opposite property against stabilization given in [8] and reveals a kind of duality of system identification and stabilization.

In the high resolution case (Section 4), we consider the generalization of the previous results under considerably weak conditions. The straight forward extension of the approach in the previous result is hard to deal with because of the complexity of the calculation for quantization error. In order to solve this difficulty, we introduce a key notion; density of the number of the optimally quantized subsections, and by using calculus of variations, analytic solutions are derived under the constraint on the number of quantization steps or the optimal code length. The solutions are functions of the distribution density of input signals and we can strictly figure out the profile of the density of the number of quantized subsections. Moreover, these results suggest several insights on system identification under the condition of finite information. We illustrate such facts for some cases and discuss on the complexity of the problem of system identification.

In Section 5, we analyse the effect of noise which is added at the input of quantizers. We show that such noise equivalently twices the magnitude of quantization error compared with the case of noise which is added at the output of quantizers. Finally, in Section 6, we compare the effect of the resolution of quantizations and that of the I/O data length. The former is more effective for decreasing quantization error in estimated system parameters, however, the latter is effective for noise error. This fact shows that there exists a trade-off between these two errors.

The main purpose of this paper is to reveal the essential properties of the optimal quantization for system identification, therefore, the argument of this paper is much analytic. Read the followings with this in mind.

In the following of this paper, except for some cases, all the proofs of theorems, lemmas, or propositions are collected in the appendix for easy understanding of the main theme and the outline of this paper. Refer them in Appendix A if necessary.

Notations: $\mathrm{E}[\cdot]$: expectation, $\mathrm{V}[\cdot]$: variance, $f(x)$: probability density of x

2 Preliminaries

The objective of this paper is to show the effect of quantizers of I/O signals for system identification on its performance in analytic and intuitive form as possible. In general, the quantization error behaves as a random signal when the quantizer has enough high resolution, and such condition has been often assumed in the

area of signal processing. However, of course, the quantization error has strong correlation with the original signal and it should be analyzed strictly, because in system identification, several kinds of correlation are used for calculating the estimation parameters. Therefore, such assumptions are not appropriate in order to understand the essence of the quantization probelm and the strict analysis is desirable. On the other hand, we should also note that it is not easy to derive analytic and intuitive understanding results for general models.

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

$$\begin{aligned} & y_o(i) = q(y(i)) + w(i),\ y(i) = \phi(i)\theta \\ & \phi(i) := [\, u(i) \quad u(i-1) \quad \cdots \quad u(i-n+1)\,], \\ & \theta := [\, \theta_1 \quad \theta_2 \quad \cdots \quad \theta_n\,]^{\mathrm{T}}, \\ & y_o,\, y,\, w,\, u \in \mathcal{R},\, \phi \in \mathcal{R}^{1\times n}, \theta \in \mathcal{R}^{n\times 1}, \end{aligned} \tag{1}$$

where w: noise, q: quantizer of the original analogue output y, y_o: observed output, ϕ: regressor vector, θ: system parameter, n: dimension of MA model, u: input, i: index of time. The input u, that is, the associate regressor vector ϕ is a realization of a stochastic process with a joint density function $f(\phi_1, \phi_2, \ldots, \phi_n)$ of ϕ_1, ϕ_2, $\ldots$, ϕ_n, where ϕ_i denotes the i-th elements of ϕ. The class of $f(\phi_1, \phi_2, \ldots, \phi_n)$ considered in this paper is explained in later.

We will also discuss a case of noise as

$$y_o(i) = q(y(i) + w(i)) \tag{2}$$

in Section 5. We refer this case (2) as *pre-quantizer noise* and the case (1) as *post-quantizer noise*. However, in order to avoid complicated notations and focus on the effect for system identification by quantization, we mainly deal with the plant (1) throughout this paper until Section 5.

The quantizer q is a memoryless symmetric type defined by

$$q(y) := \mathrm{sgn}(y)\bar{y}_j,\ y \in S_j,\ \bar{y}_j \geq 0 \tag{3}$$

$$\begin{aligned} & S_0 := \{y = 0\},\ S_j := \{y : d_{j-1} < y \leq d_j\},\ j > 0, \\ & S_j := \{y : d_{j-1} \leq y < d_j\},\ j < 0 \end{aligned} \tag{4}$$

$$\begin{aligned} & d_0 = 0 < d_1 < d_2 \cdots, \\ & d_{-1} = -d_1,\ d_{-2} = -d_2,\ \ldots, \end{aligned} \tag{5}$$

where $\mathrm{sgn}(y)\bar{y}_j$ is the assigned quantized value to the subsection S_j. The quantizer q is symmetrical with respect to the origin, and hereafter we may omit references on the negative subsections S_{-1}, S_{-2}, ... if they are obvious from the context.

The estimated parameter $\hat{\theta}$ using the least squares method with an enough length of I/O data set $\{u(i)\}$ and $\{y_o(i)\}$ is given by

$$\hat{\theta} = (U^{\mathrm{T}}U)^{-1}U^{\mathrm{T}}\left(\overline{Y} + W\right), \tag{6}$$

where

$$\begin{aligned} & U := [\, \phi(1)^{\mathrm{T}} \quad \phi(2)^{\mathrm{T}} \quad \cdots \quad \phi(N)^{\mathrm{T}}\,]^{\mathrm{T}}, \\ & W := [\, w(1) \quad w(2) \quad \cdots \quad w(N)\,]^{\mathrm{T}}, \\ & \overline{Y} := [\, \bar{y}(1) \quad \bar{y}(2) \quad \cdots \quad \bar{y}(N)\,]^{\mathrm{T}}, \\ & \bar{y}(i) := q(y(i)), \end{aligned} \tag{7}$$

and N is the I/O data length. Define the quantization error between $\bar{y}$ and y by

$$e(i) := \bar{y}(i) - y(i), \tag{8}$$

then, the estimated parameter $\hat{\theta}$ can be written as

$$\begin{aligned} \hat{\theta} &= (U^{\mathrm{T}}U)^{-1}U^{\mathrm{T}}(U\theta + E + W) \\ &= \theta + \Delta E + \Delta W \end{aligned} \tag{9}$$

$$
\begin{aligned}
E &:= [\, e(1) \quad e(2) \quad \cdots \quad e(N) \,]^{\mathrm{T}}, \\
\Delta E &:= (U^{\mathrm{T}}U)^{-1}U^{\mathrm{T}}E, \\
\Delta W &:= (U^{\mathrm{T}}U)^{-1}U^{\mathrm{T}}W.
\end{aligned} \tag{10}
$$

This shows that the estimation error $\hat{\theta} - \theta$ can be evaluated from the magnitudes of the *quantization error term* ΔE and the *noise error term* ΔW.

The reduction of the noise error term ΔW is the main theme of the ordinary system identification and its characteristics in probabilistic/deterministic sense have been well investigated. On the other hand, although the quantization error term ΔE can be reduced in general when the resolution of quantizer goes to high, there exists a limitation of the reduction under a constraint of the resolution of quantizer, and we should design *good* quantizers to reduce ΔE.

In general, the objective of designing quantizers in the field of information theory is reducing the distortion between the original signals and the quantized signals under constraints on the information of the transmitted signals [1, 13, 10, 2]. The constraint on the information of 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" respectively. On the other hand, the purpose in system identification should be the reduction of the estimation error and this point is the definitive difference.

A conventional, and reasonable, method to evaluate the noise error term ΔW in probabilistic approach is to show the convergence rate of

$$
N(U^{\mathrm{T}}U)^{-1} \overset{N\to\infty}{\longrightarrow} \frac{1}{\sigma_u^2} I, \ \frac{1}{N}U^{\mathrm{T}}W \overset{N\to\infty}{\longrightarrow} O, \tag{11}
$$

where σ_u^2 is the covariance of u, under 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 ΔE in probabilistic approach. However, different from the case of the noise error term, we should note that u and e are not independent in general, and the evaluation of $U^{\mathrm{T}}E$ is much more complicated. Solving this difficulty and evaluating the magnitude of $U^{\mathrm{T}}E$ are the key technique of this paper.

Useful notions for dealing with the relationship of u (or ϕ) and e are subsections and variable transformation of ϕ explained as follows. We define subsets Φ_j of the regressor vector ϕ associated with the subsection S_j by

$$
\Phi_j := \{\phi : y = \phi\theta \in S_j\}. \tag{12}
$$

We also consider the following variable transformation:

$$
y \;=\; \phi\theta = \phi T \cdot T^{-1}\theta = \tilde{\phi}\tilde{\theta}, \ \tilde{\theta} := T^{-1}\theta = \begin{bmatrix} \tilde{\theta}_1 \\ O \end{bmatrix}, \ \tilde{\phi} := \phi T =: [\, \tilde{\phi}_1 \quad \tilde{\phi}_2 \quad \cdots \quad \tilde{\phi}_n \,] \tag{13}
$$

where T is an orthogonal matrix. Of course such T always exits for any θ. Then, Φ_j is represented as

$$
\Phi_j := \begin{cases} \left\{\phi : \tilde{\phi}_1\tilde{\theta}_1 \in (d_{j-1}, \ d_j]\right\}, & j > 0, \\ \{\phi = 0\}, & j = 0, \\ \left\{\phi : \tilde{\phi}_1\tilde{\theta}_1 \in [d_{-j+1}, \ d_{-j})\right\}, & j < 0. \end{cases} \tag{14}
$$

We also define subsections on the space of $\tilde{\phi}_1$:

$$
I_j := \begin{cases} \left\{\tilde{\phi}_1 : \tilde{\phi}_1\tilde{\theta}_1 \in (d_{j-1}, \ d_j]\right\}, & j > 0, \\ \left\{\tilde{\phi}_1 = 0\right\}, & j = 0, \\ \left\{\tilde{\phi}_1 : \tilde{\phi}_1\tilde{\theta}_1 \in [d_{-j+1}, \ d_{-j})\right\}, & j < 0, \end{cases} \tag{15}
$$

then, the subsections S_j, Φ_j, and I_j correspond to each other, and the probability distribution of y depends only on that of $\tilde{\phi}_1$. Therefore, in order to analyse the probability distribution of y and the error e, the variable $\tilde{\phi}_1$ and its subsection I_j are convenient to deal with.

The quantization error term ΔE and U are also transformed to

$$\Delta\tilde{E} := T^{-1}\Delta E,\ \tilde{U} := UT = \begin{bmatrix} \phi(1)T \\ \phi(2)T \\ \vdots \\ \phi(N)T \end{bmatrix} \tag{16}$$

by T and it can be represented as

$$\begin{aligned} \Delta\tilde{E} &= T^{-1}(U^{\mathrm{T}}U)^{-1}U^{\mathrm{T}}E = (\tilde{U}^{\mathrm{T}}\tilde{U})^{-1}\tilde{U}^{\mathrm{T}}E \\ &= (\tilde{U}^{\mathrm{T}}\tilde{U})^{-1}\begin{bmatrix} \sum_{i=1}^{N}\tilde{\phi}_1(i)e(i) \\ \sum_{i=1}^{N}\tilde{\phi}_2(i)e(i) \\ \vdots \\ \sum_{i=1}^{N}\tilde{\phi}_k(i)e(i) \end{bmatrix} = (\tilde{U}^{\mathrm{T}}\tilde{U})^{-1}\begin{bmatrix} \sum_{i=1}^{N}\tilde{\phi}_1(i)(q(\tilde{\phi}_1(i)\tilde{\theta}_1)-\tilde{\phi}_1(i)\tilde{\theta}_1) \\ \sum_{i=1}^{N}\tilde{\phi}_2(i)(q(\tilde{\phi}_1(i)\tilde{\theta}_1)-\tilde{\phi}_1(i)\tilde{\theta}_1) \\ \vdots \\ \sum_{i=1}^{N}\tilde{\phi}_k(i)(q(\tilde{\phi}_1(i)\tilde{\theta}_1)-\tilde{\phi}_1(i)\tilde{\theta}_1) \end{bmatrix}. \end{aligned} \tag{17}$$

Note that $\|\Delta\tilde{E}\|_2 = \|\Delta E\|_2$ since T is an orthogonal matrix.

In Section 3 and Section 4, which are main results of this paper, we assume the followings on $f(\phi)$ or $f(\tilde{\phi})$.

Assumptions in Section 3:

3-1) $f(\tilde{\phi})$ is a uniform distribution

Assumptions in Section 4:

4-1) $u(i) = \phi_1(i)$, $i = \ldots, 1, 2, \ldots$ are mutually independent

4-2) the resolution of quantizer is enough high

4-3) $f(\tilde{\phi})$ is symmetric about each $\tilde{\phi}_i$-axis

4-4) $f(\tilde{\phi})$ satisfies:

$$f(\tilde{\phi}) = \prod_{i=1}^{n}(H_i + K_i(\tilde{\phi}_i - \tilde{\phi}_{io}) + O((\tilde{\phi}_i - \tilde{\phi}_{io})^2)) \tag{18}$$
$$|H_i|,\ |K_i| < \infty$$

in the neighborhood of an arbitrary $\tilde{\phi}_o = [\tilde{\phi}_{1o}\ \tilde{\phi}_{2o}\ \cdots\ \tilde{\phi}_{no}] \in \{\tilde{\phi}\}$.

4-5)

$$\frac{d(\sigma^2(\tilde{\phi}_1)f(\tilde{\phi}_1))}{d\tilde{\phi}_1} < \infty \tag{19}$$

When assumption 3-1) or 4-1) is satisfied, $U^{\mathrm{T}}U$ or $\tilde{U}^{\mathrm{T}}\tilde{U}$ converges to NI, then it is reasonable to find an optimal quantizer which minimizes $\mathrm{V}\left[U^{\mathrm{T}}E\right]$ or $\mathrm{V}\left[\tilde{U}^{\mathrm{T}}E\right]$ under constraints on the resolution of quantizer, bias-free of the quantization error term such as $\mathrm{E}\left[U^{\mathrm{T}}E\right] = 0$ or $\mathrm{E}\left[\tilde{U}^{\mathrm{T}}E\right] = 0$ and so on. In order to evaluate these quantities, we prepare further notations. The marginal distribution density $f(\tilde{\phi}_1)$ on the space of $\tilde{\phi}_1$ is defined by

$$f(\tilde{\phi}_1) := \int f([\,\tilde{\phi}_1\ \ \tilde{\phi}_2\ \ \cdots\ \ \tilde{\phi}_n\,])d\tilde{\phi}_2\cdots d\tilde{\phi}_n.$$

The notations $f(\tilde{\phi}_i, \tilde{\phi}_j)$, $f(\tilde{\phi}_i, \tilde{\phi}_j, \tilde{\phi}_k)$, ... are also defined similarly. Then, a bias-free condition $\mathrm{E}\left[U^{\mathrm{T}}E\right] = 0$, that is

$$\mathrm{E}\left[\sum_{i=1}^{N}\phi_k(i)e(i)\right] = 0$$

for each k, can be written as

$$\mathrm{E}\left[\sum_{i=1}^{N}\tilde{\phi}_k(i)e(i)\right] = N\mathrm{E}\left[\tilde{\phi}_k \cdot e(\tilde{\phi}_1)\right] = N\int \tilde{\phi}_k e(\tilde{\phi}_1) f(\tilde{\phi}_1, \tilde{\phi}_k) d\tilde{\phi}_1 d\tilde{\phi}_k = 0 \tag{20}$$

for each $k \neq 1$ and

$$\mathrm{E}\left[\sum_{i=1}^{N}\tilde{\phi}_1(i)e(i)\right] = N\mathrm{E}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] = N\int \tilde{\phi}_1 e(\tilde{\phi}_1) f(\tilde{\phi}_1) d\tilde{\phi}_1 = 0 \tag{21}$$

for $k = 1$ because of $\tilde{U}^{\mathrm{T}}E = T^{\mathrm{T}}U^{\mathrm{T}}E$, where T is orthogonal, that is, nonsingular. Of course, another bias-free condition $\mathrm{E}\left[\tilde{U}^{\mathrm{T}}E\right] = 0$ is directly reduced to (20) and (21). If assumption 3-1) or 4-3) is satisfied,

$$\int \tilde{\phi}_k f(\tilde{\phi}_1, \tilde{\phi}_k) d\tilde{\phi}_k = 0 \tag{22}$$

is held for $k \neq 1$, then, (20) is automatically satisfied. Therefore, the bias-free condition is reduced to (21) under such assumption. A sufficient condition of (21) is

$$\mathrm{E}_{I_j}\left[\tilde{\phi}_1 e(\tilde{\phi}_1)\right] := \int_{\tilde{\phi}_1 \in I_j} \tilde{\phi}_1 e(\tilde{\phi}_1) f(\tilde{\phi}_1) d\tilde{\phi}_1 = 0, \ \ \forall j. \tag{23}$$

This condition shows a properness of $\overline{y}_j$ which represents the subsection I_j and it is enough reasonable to be satisfied.

On the other hand, the objective variance $\mathrm{V}[U^{\mathrm{T}}E]$ $(= \mathrm{V}[\tilde{U}^{\mathrm{T}}E])$ is written as

$$\mathrm{V}[U^{\mathrm{T}}E](= \mathrm{V}[\tilde{U}^{\mathrm{T}}E]) = \sum_{k=1}^{n}\mathrm{E}\left[\left(\sum_{i=1}^{N}\tilde{\phi}_k(i)e(i)\right)^2\right] = \sum_{k=1}^{n}\mathrm{E}\left[\left(\sum_{i=1}^{N}\tilde{\phi}_k(i)e(\tilde{\phi}_1(i))\right)^2\right]. \tag{24}$$

With respect to this formula, we can derive the following two key lemmas.

Lemma 2.1 *Under a condition:*

$$\int \tilde{\phi}_h f(\tilde{\phi}_1, \ldots, \tilde{\phi}_h, \ldots, \tilde{\phi}_n) d\tilde{\phi}_h = 0, \tag{25}$$

$$\mathrm{E}\left[\left(\sum_{i=1}^{N}\tilde{\phi}_k(i)e(\tilde{\phi}_1(i))\right)^2\right] = \begin{cases} N\int \tilde{\phi}_1^2 e^2(\tilde{\phi}_1) f(\tilde{\phi}_1) d\tilde{\phi}_1 & \textit{for } \ k = 1 \\ N\int \tilde{\phi}_k^2 e^2(\tilde{\phi}_1) f(\tilde{\phi}_1, \tilde{\phi}_k) d\tilde{\phi}_1 d\tilde{\phi}_k & \textit{for } \ k \neq 1 \end{cases} \tag{26}$$

is satisfied.

The proof of this lemma is given in Appendix A as mentioned in Section 1. Note that the condition in the lemma is satisfied under the assumption 3-1).

Lemma 2.2 *Assume that $f(\tilde{\phi})$ satisfies (18), then,*

$$\mathrm{E}\left[\left(\sum_{i=1}^{N}\tilde{\phi}_k(i)e(i)\right)^2\right] \underset{\Delta y_{\max}\to 0}{\longrightarrow} N\mathrm{E}\left[\tilde{\phi}_k^2(i)e^2(i)\right], \tag{27}$$

where $\Delta y_{\max}$ is the maximum width of the subsections S_j of quantizer defined by $\Delta y_{\max} := \max_j |d_{j+1} - d_j|$.

This lemma is for high resolution case discussed in Section 4. See Appendix A for the proof.

When Lemma 2.1 is applicable, (24) is represented by

$$\begin{aligned} \mathrm{V}\left[U^{\mathrm{T}}E\right] &= N\int\left(\sum_{k=1}^{n}\tilde{\phi}_k^2\right)e^2(\tilde{\phi}_1)f(\tilde{\phi}_1,\ldots,\tilde{\phi}_n)d\tilde{\phi}_1\cdots d\tilde{\phi}_n \\ &= N\int\sigma^2(\tilde{\phi}_1)e^2(\tilde{\phi}_1)f(\tilde{\phi}_1)d\tilde{\phi}_1, \end{aligned} \tag{28}$$

where $\sigma(\tilde{\phi}_1)$ is a kind of variance of $f(\tilde{\phi})$ at $\tilde{\phi}_1$ defined by

$$\sigma(\tilde{\phi}_1) := \left(f(\tilde{\phi}_1)^{-1}\int\left(\sum_{k=1}^{n}\tilde{\phi}_k^2\right)f(\tilde{\phi}_1,\ldots,\tilde{\phi}_n)d\tilde{\phi}_2\cdots d\tilde{\phi}_n\right)^{\frac{1}{2}}. \tag{29}$$

And also in the case that Lemma 2.2 is applicable, we get

$$\mathrm{V}\left[U^{\mathrm{T}}E\right]\left(=\mathrm{V}\left[\tilde{U}^{\mathrm{T}}E\right]\right)\underset{\Delta y_{\max}\to 0}{\longrightarrow} N\int\sigma^2(\tilde{\phi}_1)e(\tilde{\phi}_1)f(\tilde{\phi}_1)d\tilde{\phi}_1. \tag{30}$$

Another possible objective function is $\mathrm{V}\left[(\tilde{U}^{\mathrm{T}}E)_1\right]$, that is the variance of the first element of $\tilde{U}^{\mathrm{T}}E$, which focuses on the quantization error in the unique nonzero element $\tilde{\theta}_1$ of $\tilde{\theta}$. Under the condition (21), this formula is also represented by

$$\mathrm{V}\left[(\tilde{U}^{\mathrm{T}}E)_1\right] = \mathrm{V}\left[\tilde{\phi}_1 e(\tilde{\phi}_1)\right] = N\int\tilde{\phi}_1^2 e^2(\tilde{\phi}_1)f(\tilde{\phi}_1)d\tilde{\phi}_1. \tag{31}$$

Based on these observations, the original optimization problem is reduced to the feasible forms and the main result in the following of this paper is summarized as deriving the optimal quantizer for minimizing:

1) (31) when $\tilde{\phi}$ satisfies assumption 3-1) (Section 3, **low resolution case**),

2) (28) when ϕ satisfies assumptions 4-1) $\sim$ 4-5) (Section 4, **high resolution case**).

Moreover, although the above two cases of problem have enough meanings for themselves, they also have a deep connection which is explained in Section 4.

3 Low Resolution Quantization

3.1 Explicit optimal quantization scheme for low resolution case

At first we state the problem formulation dealt with in this section. The next is assumed, which corresponds to that of 3-1) in Section 2.

Assumption 3.1 *$\tilde{\phi}_1$ obeys a uniform distribution in $[-\kappa,\ \kappa]$ (this means y obeys a uniform distribution in $[-\tilde{\theta}_1\kappa,\ \tilde{\theta}_1\kappa] =: [-\kappa',\ \kappa']$).*

As mentioned before, the subject of this paper is mainly for the analysis in order to understand the essential properties of the optimal quantizers. Therefore, although some assumptions do not consist with the original objective of system identification, keep this intention in mind in reading the following of this paper.

Assumption 3.1 automatically guarantees the condition (22) and (26) in Lemma 2.1 (i.e., (31)). Then, the following problem is considered.

Problem 3.1 *Let M_o be the number of the quantized subsections of $[-\kappa,\ \kappa]$. For the system (1) with Assumption 3.1 and a fixed M_o, find a quantizer q that minimizes the variance of (31) such that* $\mathrm{E}_{I_j}\left[\tilde{\phi}_1(i)\cdot e(i)\right] = 0$ *($\forall j$) for the even number M_o or* $\mathrm{E}_{I_{-1}+I_1}\left[\tilde{\phi}_1(i)\cdot e(i)\right] = 0$ $\left(:= \int_{\tilde{\phi}_1(i)\in I_{-1}\cup I_1} \tilde{\phi}_1\cdot e(\tilde{\phi}_1) f(\tilde{\phi}_1) d\tilde{\phi}_1\right)$ *and* $\mathrm{E}_{I_j}\left[\tilde{\phi}_1(i)\cdot e(i)\right] = 0$ *(other j) for the odd number M_o.*

This problem is not only for the case that the identified systems have such special conditions. As we will explain the reasons in Note 4.3 is Section 4, this problem is approximately applicable to the optimal quantization in a local area around the origin of the regressor vector in general.

As described in Section 2, the quantization scheme of $[-\tilde{\theta}_1\kappa,\ \tilde{\theta}_1\kappa] = [-\kappa',\ \kappa']$ 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 $I_{-M}, \ldots, I_{-2}, I_{-1}, I_1, I_2, \ldots, I_M$, where

$$M := \begin{cases} \frac{1}{2}M_o & \text{for even } M_o \\ \frac{1}{2}(M_o+1) & \text{for odd } M_o \end{cases}, \tag{32}$$

and the assigned quantized values

$$\begin{aligned} & q(y),\ y\in S_j \\ &= q(\tilde{\phi}_1),\ \tilde{\phi}_1\in I_j \\ &= \bar{y}_j \end{aligned} \tag{33}$$

for each subsection I_j (see Fig. 1). Therefore, we should find an optimal $I_{-M}, \ldots, I_M$ and $\bar{y}_{-M}, \ldots, \bar{y}_M$ for a fixed M. This is a minimization problem of $\mathrm{V}\left[\tilde{\phi}_1 e(\tilde{\phi}_1)\right]$ of an about $(2M\times 2)$-variables and it seems to be a considerably hard problem in the sense of computation complexity. However we can show that this problem is reduced to be a feasible one by using the following calculations.

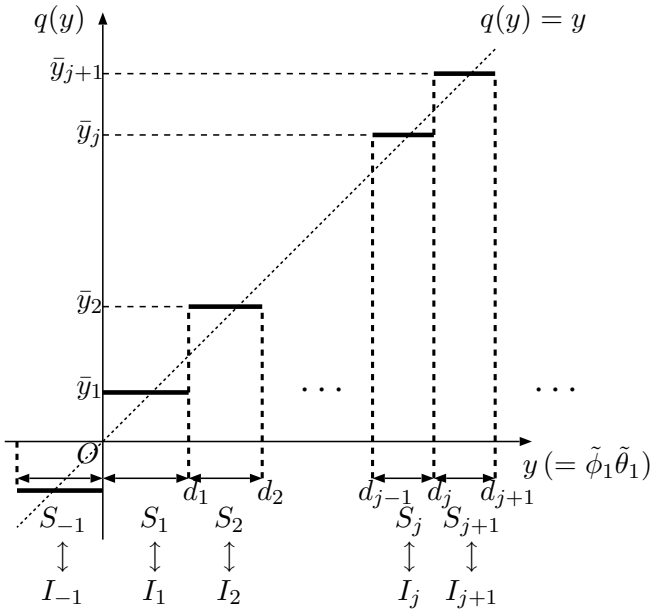

Fig. 1 The quantization scheme of q

Hereafter, we consider the case of even M_o. The case of odd M_o is almost similar to the even case and the differences are explained in Note 3.1.

First, we consider $S_1 = (0,\ d_1]$ (equivalently I_1 on $\tilde{\phi}_1$) and $S_2 = (d_1,\ d_2]$ (equivalently I_2 on $\tilde{\phi}_1$) where their boundaries d_1, d_2 have a relation:

$$d_1 = r_1 d_2,\ r_1\in[0,1], \tag{34}$$

with an appropriate ratio r_1. The quantized values $\bar{y}_1$ and $\bar{y}_2$ for the subsections S_1 on y (or I_1 on $\tilde{\phi}_1$) and S_2 (or I_2) satisfying

$$\mathrm{E}_{I_j}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] = 0,\ j = 1, 2$$

are given as follows. Let $\bar{y}_1 = \frac{d_1}{2} + h_1$, where h_1 is an offset, then,

$$\mathrm{E}_{I_1}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] = \int_{-k_1}^{k_1} \left(\frac{r_1 d_2}{2} + z\right)(z - h_1)\frac{1}{2\kappa'}dz = \frac{1}{2\kappa'}\left(\frac{2}{3}k_1^3 - r_1 d_2 h_1 k_1\right),\ k_1 := \frac{d_1}{2},$$

and therefore,

$$h_1 = \frac{2}{3}\frac{k_1^2}{r_1 d_2} = \frac{1}{6}r_1 d_2. \tag{35}$$

Similarly, let $\bar{y}_2 := \frac{(1+r_1)d_2}{2} + h_2$, where h_2 is the offset, then,

$$\mathrm{E}_{I_2}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] = \int_{-k_2}^{k_2} \left(\frac{d_2 + r_1 d_2}{2} + z\right)(z - h_2)\frac{1}{2\kappa'}dz = \frac{1}{2\kappa'}\left(\frac{2}{3}k_2^3 - (d_2 + r_1 d_2)h_2 k_2\right),$$
$$k_2 := \frac{d_2(1 - r_1)}{2},$$

and therefore,

$$h_2 = \frac{2}{3}k_2^2\frac{1}{d_2(1 + r_1)} = \frac{1}{6}\frac{(1 - r_1)^2}{(1 + r_1)}d_2. \tag{36}$$

Note that in order to make the expectations $\mathrm{E}_{I_1}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right]$ and $\mathrm{E}_{I_2}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right]$ zero, the assigned quantized values $\bar{y}_1$ and $\bar{y}_2$ must be larger than the central values of each subsection S_1 on y (or I_1 on $\tilde{\phi}_1$) and S_2 (or I_2). Hereafter in this section, the quantized values $\bar{y}_i$ are selected as such values.

By using these $\bar{y}_1$ and $\bar{y}_2$, the variances of $\tilde{\phi}_1 e(\tilde{\phi}_1)$ in each subsection also can be calculated as follows. Let $\mathrm{V}_{I_j}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right]$ denote the quantity

$$\mathrm{V}_{I_j}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] := \int_{I_j}\left(\tilde{\phi}_1 \cdot e(\tilde{\phi}_1) - \mathrm{E}_{I_j}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right]\right)^2 f(\tilde{\phi}_1)d\tilde{\phi}_1,$$

then, for the even M_o,

$$\mathrm{V}_{I_1}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] = \int_{-k_1}^{k_1}\left(\frac{r_1 d_2}{2} + z\right)^2 (z - h_1)^2\frac{1}{2\kappa'}dz = \frac{1}{2160}\frac{1}{2\kappa'}d_2^5\left(32r_1^5\right)$$

(note that $\kappa' = \tilde{\theta}_1\kappa$), and similarly

$$\begin{aligned}\mathrm{V}_{I_2}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] &= \int_{-k_2}^{k_2}\left(\frac{d_2(1 + r_1)}{2} + z\right)^2 (z - h_2)^2\frac{1}{2\kappa'}dz \\ &= \frac{1}{2160}\frac{1}{2\kappa'}d_2^5\left\{-18(1 - r_1)^5 + 45(1 + r_1)^2(1 - r_1)^3 + 5(1 - r_1)^7(1 + r_1)^{-2}\right\}.\end{aligned}$$

Therefore, the sum of $\mathrm{V}_{I_1}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right]$ and $\mathrm{V}_{I_2}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right]$ is

$$\mathrm{V}_{I_1}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] + \mathrm{V}_{I_2}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] = \frac{1}{2160}\frac{1}{2\kappa'}d_2^5\psi_1(r_1)$$
$$\psi_1(r_1) := 32r_1^5 - 18(1 - r_1)^5 + 45(1 + r_1)^2(1 - r_1)^3 + 5(1 - r_1)^7(1 + r_1)^{-2}. \tag{37}$$

The minimizer r_1^o of this sum is defined by

$$\begin{aligned} r_1^o &= \arg\min_{r_1 \in [0,1]} \psi_1(r_1) \\ \psi_1^{\min} &:= \psi_1(r_1^o), \end{aligned}$$

and

$$\left(\mathrm{V}_{I_1}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] + \mathrm{V}_{I_2}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right]\right)\Big|_{r_1=r_1^o} = \frac{1}{2160}\frac{1}{2\kappa'}d_2^5\psi_1^{\min}. \tag{38}$$

Note that the optimal r_1^o is independent of the value of d_2, which is the upper boundary of S_2.

Next, we consider another subsection S_3 on y (or I_3 on $\tilde{\phi}_1$) together with S_1 (or I_1), S_{-1} (or I_{-1}) and S_2 (or I_2). Suppose the relation between d_2 and d_3 is:

$$d_2 = r_2 d_3, \tag{39}$$

where r_2 is also an appropriate number in $[0,\ 1]$. Similar to the case of S_1, S_{-1} and S_2, the offset h_3 of $\bar{y}_3$ for the subsection S_3 on y (or I_3 on $\tilde{\phi}_1$) satisfying $\mathrm{E}_{I_3}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] = 0$ and the variance $\mathrm{V}_{I_3}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right]$ can be determined as follows:

$$h_3 = \frac{2}{3}k_3^2\frac{1}{d_3(1+r_2)} = \frac{1}{6}\frac{(1-r_2)^2}{(1+r_2)}d_3,\ k_3 := \frac{d_3(1-r_2)}{2} \tag{40}$$

$$\begin{aligned}
\mathrm{V}_{I_3}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] &= \int_{-k_3}^{k_3}\left(\frac{d_3(1+r_2)}{2} + z\right)^2 (z-h_3)^2\frac{1}{2\kappa'}dz \\
&= \frac{1}{2160}\frac{1}{2\kappa'}d_3^5\left\{-18(1-r_2)^5 + 45(1+r_2)^2(1-r_2)^3 + 5(1-r_2)^7(1+r_2)^{-2}\right\}
\end{aligned}$$

Therefore, the optimal r_2^o that minimizes $\mathrm{V}_{I_1}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] + \mathrm{V}_{I_2}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] + \mathrm{V}_{I_3}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right]$ is given by solving the following minimization problem.

$$\begin{aligned}
r_2^o &:= \arg\min_{r_2}\left(V_{I_1}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] + V_{I_2}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] + V_{I_3}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right]\right) \\
&= \arg\min_{r_2}\frac{1}{2160}\frac{1}{2\kappa'}d_3^5\psi_2(r_2) \\
\psi_2(r_2) &:= \psi_1^{\min}r_2^5 - 18(1-r_2)^5 + 45(1+r_2)^2(1-r_2)^3 + 5(1-r_2)^7(1+r_2)^{-2}.
\end{aligned} \tag{41}$$

Note 3.1 In the case of odd M_o, the quantized values $\bar{y}_1$, $\bar{y}_2$ and $\bar{y}_3$ for the subsection S_{-1}, S_1, S_2 and S_3 on y (correspondingly I_{-1}, I_1, I_2 and I_3 on $\tilde{\phi}_1$) should satisfy

$$\mathrm{E}_{I_1+I_{-1}}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] = 0,\ \mathrm{E}_{I_2}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] = 0,\ \mathrm{E}_{I_3}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] = 0.$$

Therefore, from the symmetry of S_1 and S_{-1}, $\bar{y}_1 = 0$. The quantized values $\bar{y}_2$ and $\bar{y}_3$ are given as similar to the even case and

$$\begin{aligned}
r_1^o &:= \arg\min_{r_1}\left(\frac{1}{2}\mathrm{V}_{I_1+I_{-1}}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] + \mathrm{V}_{I_2}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right]\right) \\
&:= \arg\min_{r_1}\frac{1}{2160}\frac{1}{2\kappa'}d_2^5\psi_1(r_1) \\
\psi_1(r_1) &:= 432r_1^5 - 18(1-r_1)^5 + 45(1+r_1)^2(1-r_1)^3 + 5(1-r_1)^7(1+r_1)^{-2}, \\
r_2^o &:= \arg\min_{r_2}\left(\frac{1}{2}\mathrm{V}_{I_1+I_{-1}}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] + \mathrm{V}_{I_2}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] + \mathrm{V}_{I_3}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right]\right) \\
&= \arg\min_{r_2}\frac{1}{2160}\frac{1}{2\kappa'}d_3^5\psi_2(r_2) \\
\psi_2(r_2) &:= \psi_1^{\min}r_2^5 - 18(1-r_2)^5 + 45(1+r_2)^2(1-r_2)^3 + 5(1-r_2)^7(1+r_2)^{-2}.
\end{aligned}$$

Note that the difference on the formulas between the even case and the odd case is only the coefficient of r_1^5 in $\psi_1(r_1)$. □

By repeating the above process, we obtain the following result.

Theorem 3.1 *The optimal ratios r_j^o for Problem 3.1 are given by solving the following optimization problem iteratively.*

$$r_j^o := \arg \min_{r_j \in [0,\,1]} \psi_j(r_j) \tag{42}$$

$$\psi_j(r_j) := \psi_{j-1}^{\min} r_j^5 - 18(1-r_j)^5 + 45(1+r_j)^2(1-r_j)^3 + 5(1-r_j)^7(1+r_j)^{-2} \tag{43}$$

$$\psi_j^{\min} := \psi_j(r_j^o) \tag{44}$$

$$\psi_0^{\min} := \begin{cases} 32 & \textit{for even } M_o \\ 432 & \textit{for odd } M_o \end{cases} \tag{45}$$

The optimal value of the variance is given by

$$\mathrm{V}_M\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] := \begin{cases} \sum_{j=-M}^{M} \mathrm{V}_{I_j}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] & \textit{for even } M_o \\ \mathrm{V}_{I_1+I_{-1}}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] + \sum_{j=-M,\, j\neq\pm 1}^{M} \mathrm{V}_{I_j}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] & \textit{for odd } M_o \end{cases} \tag{46}$$

$$= \frac{1}{2160}\kappa'^4 \psi_{M-1}^{\min} = \frac{1}{2160}\tilde{\theta}_1^4 \kappa^4 \psi_{M-1}^{\min}. \tag{47}$$

We call this optimal quantization scheme $\mathrm{Q}_{\mathrm{opt}}$ hereafter.

Note 3.2 The original minimization problem of an about $(2M \times 2)$-variables function $\mathrm{V}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right]$ can be reduced to a recursive minimization problem of an only 1-variable rational function. Moreover, from Lemma A.1 in Appendix A, the local minimum of $\psi_j(r_j)$ in $r_j \in (0,\ 1)$ is unique, therefore, finding the minimizer does not require high complexity of calculation.

Every ratio r_j^o can be explicitly given by (42) $\sim$ (45) iteratively, however, understanding the properties of r_j^o is not straightforward from (42) $\sim$ (45) directly. On the asymptotic characteristics of the optimal ratios r_j^o $(j = 1, 2, \ldots)$ and the related quantities, we can derive the following series of Lemma 3.1 $\sim$ 3.4. Their proofs are collected in Appendix A.

Lemma 3.1

$$r_j^o < r_{j+1}^o,\ \forall j > 0 \tag{48}$$

$$r_j^o \to 1,\ j \to \infty \tag{49}$$

Lemma 3.2

$$|S_j| > |S_{j+1}|,\ |I_j| > |I_{j+1}|,\ \forall j > 0, \tag{50}$$

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

Lemma 3.2 shows that the optimal quantization scheme $\mathrm{Q}_{\mathrm{opt}}$ has the property that it is coarse around the origin of y and becomes dense where y goes to the boundaries of $[-\kappa',\ \kappa']$. This property is, in some sense, a dual to the result of the quantization problem for stabilization by [8], that is, the coarsest quantization scheme for stabilization is dense around the origin and becomes coarse at a distance from the origin. These observations suggest that there seems to exist a trade-off between parameter estimation and stabilization in quantization scheme for adaptive type control systems.

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 of the optimality of $\mathrm{Q}_{\mathrm{opt}}$, that is, when a region $[-\gamma,\ \gamma]$ of $\tilde{\phi}_1$ is quantized, the width of I_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 shows this fact. See Appendix A for the proof.

Lemma 3.3

$$\prod_{j=1}^{\infty} \frac{1}{r_j^o} = \infty \tag{51}$$

From Lemma 3.1 to Lemma 3.3, we know the outline of the quantization of the region $[-\kappa', \kappa']$. Next, consider the evaluation of the magnitude of $\mathrm{V}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right]$ with respect to the number of the quantization levels M, and the following lemma shows an asymptotic characteristics of $\psi_M^{\min}$.

Lemma 3.4

$$\psi_M^{\min} \to \Psi_a^b(M),\ M \to \infty \tag{52}$$

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

$$\hat{\psi}(m) - \hat{\psi}(m-1) = a\hat{\psi}^b(m-1) \tag{53}$$

By approximating the difference equation (53) (or (151) in Appendix A) with a differential equation

$$\frac{d\tilde{\psi}(m)}{dm} = (a+\nu)\tilde{\psi}^b(m) \geq a\tilde{\psi}^b(m) + o(\tilde{\psi}^b(m)), \tag{54}$$

where $\nu > 0$ is an appropriate constant number, then, we obtain

$$\tilde{\psi}(m) = \{(-b+1)(a+\nu)m + K\}^{\frac{1}{-b+1}} \tag{55}$$

for an appropriate constant K. From (47) and the convexity of the function (55), the variance $\mathrm{V}_M[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)]$ at sufficiently large M satisfies

$$\begin{aligned}
\mathrm{V}_M\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] &\leq \frac{1}{2160}\kappa^4((-3/2+1)((-5\cdot 3^{-\frac{5}{2}}+\nu)(M-1)+K))^{\frac{1}{-3/2+1}} \\
&= A\kappa^4(M-K')^{-2} \\
A &:= \frac{1}{540}\left(5\cdot 3^{-\frac{5}{2}}-\nu\right)^{-2} \\
K' &:= (5\cdot 3^{-\frac{5}{2}}-\nu)^{-1}K.
\end{aligned} \tag{56}$$

This (56) shows a relation between the optimal variance and the number of quantization levels. In the following section this result is used to evaluate the magnitude of ΔE.

3.2 Numerical simulation

In this subsection, we demonstrate the characteristics of the optimal quantizater by using simple numerical examples.

At first generate 10000 sets of I/O data for the system (1) with $\theta = 1$ and $w = 0$; the 1st order MA model and noise-free case, where $u(i)$ $(= \tilde{\phi}_1(i))$ is an independent random noise of uniform distribution in $[-4, 4]$. Show the histogram of $u = \tilde{\phi}_1$ in Fig. 2.

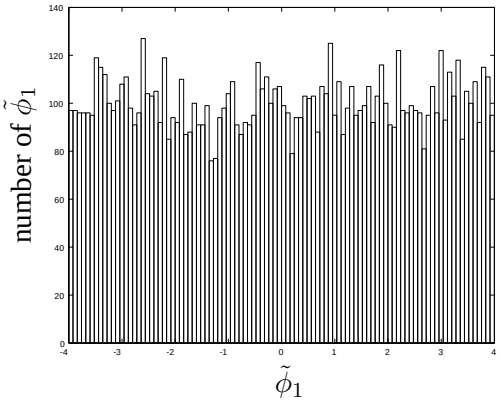

Fig. 2 Histogram of input data $u = \tilde{\phi}_1$

Next quantize the output data y with the optimal quantizers in Theorem 3.1 and uniform quantizers for comparison under the constraint of $M = 5$ ($M_o = 9$), $M = 10$ ($M_o = 19$), and $M = 50$ ($M_o = 99$). Fig. 3, 5, and 7 show the step functions of the optimal quantizers for $M = 5$, $M = 10$, and $M = 50$ respectively and Fig. 4, 6, and 8 show the corresponding step functions of the uniform quantizers. Fig. 3, 5, and 7 show the property of the optimal quantizers, that is, it is coarse around the origin and goes to be dense apart from the origin. Then, calculate the bias term ave. $\tilde{\phi}_1 \cdot e = \frac{1}{10000}\sum_{i=1}^{10000} \tilde{\phi}_1(i) \cdot e(i)$, its variance ave. $\tilde{\phi}_1^2 \cdot e^2 = \frac{1}{10000}\sum_{i=1}^{10000} \tilde{\phi}_1^2(i) \cdot e^2(i)$, and the quantization error term $\frac{1}{10000}(U^{\mathrm{T}}U)^{-1}U^{\mathrm{T}}E$ by using $u = \tilde{\phi}_1$ and the known quantization error e between y and the calculated $\bar{y}$. Table 1, 2, and 3 show the summary of the results. From Table 1, 2, and 3, the optimal quantizers which minimize $\mathrm{E}\left[\tilde{\phi}_1^2 \cdot e^2\right]$ attain lesser ave. $\tilde{\phi}_1^2 \cdot e^2 = \frac{1}{10000}\sum_{i=1}^{10000} \tilde{\phi}_1^2(i) \cdot e^2(i)$ than those of the uniform quantizers and consequently attain lesser $|\Delta E|$.

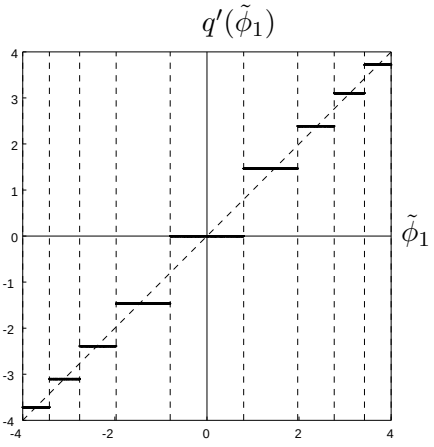

Fig. 3 Optimal quantization scheme $\mathrm{Q_{opt}}$ for $M = 5$

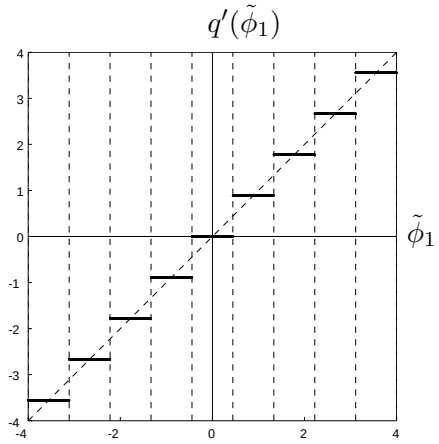

Fig. 4 Uniform quantization scheme for $M = 5$

Table. 1 The bias, variance, and quantization error for $M = 5$

	$\mathrm{Q_{opt}}$	uniform
ave. $\tilde{\phi}_1 \cdot e$	6.61e-004	-4.91e-002
ave. $\tilde{\phi}_1^2 \cdot e^2$	1.79e-001	2.89e-001
$\lvert\Delta E\rvert$	1.18e-004	9.07e-003

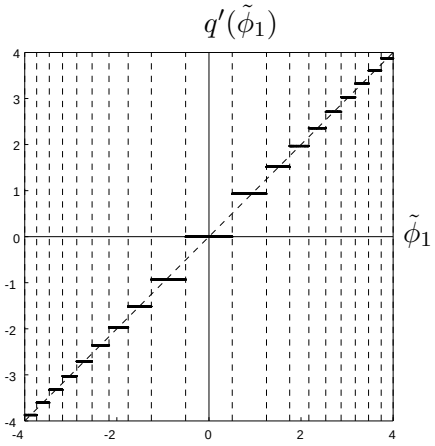

Fig. 5 Optimal quantization scheme $\mathrm{Q_{opt}}$ for $M = 10$

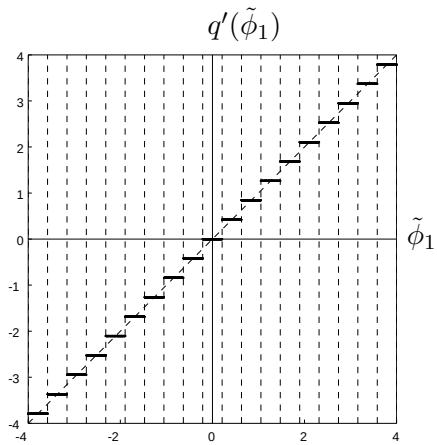

Fig. 6 Uniform quantization scheme for $M = 10$

Table. 2 The bias, variance, and quantization error for $M = 10$

	optimal	uniform
ave. $\tilde{\phi}_1 \cdot e$	1.19e-004	-1.17e-002
ave. $\tilde{\phi}_1^2 \cdot e^2$	4.54e-002	7.03e-002
$\lvert\Delta E\rvert$	2.36e-005	2.16e-003

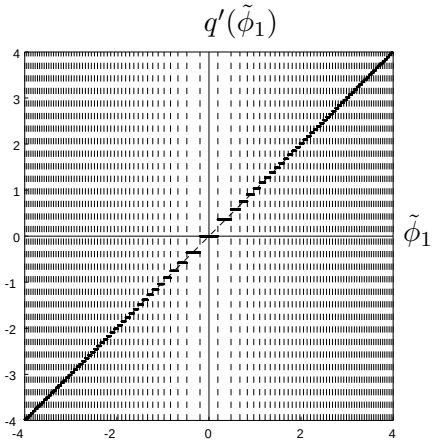

Fig. 7 Optimal quantization scheme $\mathrm{Q_{opt}}$ for $M = 50$

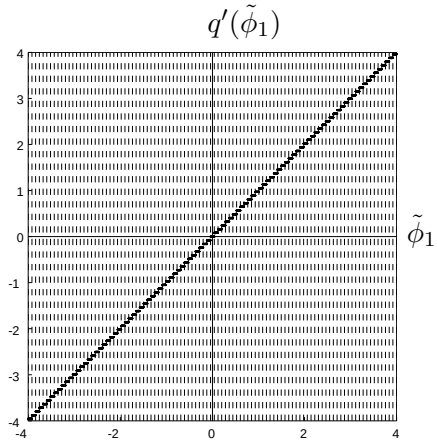

Fig. 8 Uniform quantization scheme for $M = 50$

Table. 3 The bias, variance, and quantization error for $M = 50$

	optimal	uniform
ave. $\tilde{\phi}_1 \cdot e$	-8.24e-005	-8.90e-004
ave. $\tilde{\phi}_1^2 \cdot e^2$	1.88e-003	2.90e-003
$\lvert\Delta E\rvert$	1.68e-005	1.64e-004

4 High Resolution Quantization

In Section 3, we gave an optimal quantizer regardless of its resolution, however, under strong assumptions on the distribution density $f(\phi)$ or $f(\tilde{\phi})$. On the other hand, in this section we give optimal quantizers for general distribution densities $f(\phi)$ where quantizers are assumed to be in high resolution. Here we formally state the assumptions of this section,

Assumption 4.1 *The input u, the distribution density $f(\tilde{\phi}_1)$ and the quantizer satisfy the assumptions 4-1) $\sim$ 4-5).*

At first, the assumption 4-1) gives the reasonability of $\mathrm{V}[U^{\mathrm{T}}E]$ as the minimized function. With the assumptions 4-2) and 4-3), the bias-free condition $\mathrm{E}[U^{\mathrm{T}}E]$ or $\mathrm{E}[\tilde{U}^{\mathrm{T}}E]$ is asymptotically satisfied when the widths of the quantization steps go to 0. Moreover, Lemma 2.2 is derived from the assumption 4-4) and it shows that the variance $\mathrm{V}[U^{\mathrm{T}}E]$ $(= \mathrm{V}[\tilde{U}^{\mathrm{T}}E])$, which is minimized, can be approximated by (28) in high resolution case. Therefore, the highlight of the problem is in the calculation of (28) for general $f(\phi)$ and finding its minimizer.

A key idea to solve the problem is introducing the following quantity on the distribution of quantization subsections, which is a reasonable notion under assumption 4-2) in Section 2.

Definition 4.1 *The quantity $g(\tilde{\phi}_1)$ which satisfies*

$$\tilde{\theta}_1 g(\tilde{\phi}_1) d\tilde{\phi}_1 = \text{number of quantized subsections in } d(\tilde{\theta}_1\tilde{\phi}_1)\,(= \tilde{\theta}_1 d\tilde{\phi}_1) \tag{57}$$

is called distribution density of the number of quantized subsections.

This quantity is the same introduced in [1, 13] and from this definition, $g(\tilde{\phi}_1)^{-1}$ represents the width of the quantization step at $\tilde{\theta}_1\tilde{\phi}_1$.

In Section 3 for the bias-free condition, the quantized value for each subsection is strictly assigned to satisfy that the expectation of the quantization error is zero in each subsection. Although such consideration is indispensable in low resolution case of the quantization, however, the bias-free is asymptotically satisfied in high resolution case and the assignment of the quantized value $\bar{y}_j$ is not critical problem. In particular, at the asymptotic situation of $|I_j| \to 0$, the middle point of each subsection is reasonable to be assigned as the quantized value. Therefore, we fix such quantized values in the following of this section.

Then, we assume the following.

Assumption 4.2 *The density $g(\tilde{\phi}_1)$ satisfies:*

$$\frac{dg(\tilde{\phi}_1)^{-2}}{d\tilde{\phi}_1} \quad < \quad \infty. \tag{58}$$

With this "smoothness" of the density $g(\tilde{\phi}_1)$ and that of $f(\tilde{\phi}_1)$, which is given by the assumption 4-5), we can select the mean value $g_j^{-1} \sim g(\tilde{\phi}_1)^{-1}$ for the subsection I_j and then, we define $f_j \sim f(\tilde{\phi}_1)$ in $\tilde{\phi}_1 \in I_j$ which satisfies the next.

$$p_j := \int_{I_j} f(\tilde{\phi}_1) d\tilde{\phi}_1 =: f_i g_i^{-1}$$

Moreover, with the variance $\sigma(\tilde{\phi}_1)$ of $f(\tilde{\phi})$ at $\tilde{\phi}_1$ defined in (29), the assumption 4-5), Assumption 4.2, and with $\Delta\tilde{\phi} := \max_j \tilde{\phi}_1^{-1}|d_{j+1} - d_j|$, we can derive the followings by direct calculations:

$$\begin{aligned} & \int \left(\sum_{k=1}^{n} \tilde{\phi}_k^2 \right) e^2(\tilde{\phi}_1) f(\tilde{\phi}_1, \ldots, \tilde{\phi}_n) d\tilde{\phi}_1 \cdots d\tilde{\phi}_n \\ = & \int \sigma^2(\tilde{\phi}_1) e^2(\tilde{\phi}_1) f(\tilde{\phi}_1) d\tilde{\phi}_1 \end{aligned}$$

$$
\begin{aligned}
&= \sum_j \int_{I_j} \sigma^2(\tilde{\phi}_1) \cdot e^2(\tilde{\phi}_1) f(\tilde{\phi}_1) d\tilde{\phi}_1 \\
&= \sum_j \int_{\tilde{\phi}_{1j}-\frac{1}{2}g_j^{-1}}^{\tilde{\phi}_{1j}+\frac{1}{2}g_j^{-1}} (\bar{y}_j - x)^2 \cdot \sigma^2(x) f(x) dx + O(\Delta\tilde{\phi}^3) \\
&= \sum_j \int_{\tilde{\phi}_{1j}-\frac{1}{2}g_j^{-1}}^{\tilde{\phi}_{1j}+\frac{1}{2}g_j^{-1}} (\tilde{\theta}_1^2\tilde{\phi}_{1j} - x)^2 \sigma^2(\tilde{\phi}_{1j}) f_j dx + O(\Delta\tilde{\phi}^3) \\
&= \tilde{\theta}_1^2 \sum_j \frac{1}{12} g_j^{-3} \sigma^2(\tilde{\phi}_{1j}) f_j + O(\Delta\tilde{\phi}^3) \\
&= \tilde{\theta}_1^2 \sum_j \int_{\tilde{\phi}_{1j}-\frac{1}{2}g_j^{-1}}^{\tilde{\phi}_{1j}+\frac{1}{2}g_j^{-1}} \frac{1}{12} g_j^{-2} \sigma^2(\tilde{\phi}_{1j}) f_j dx + O(\Delta\tilde{\phi}^3) \\
&= \tilde{\theta}_1^2 \sum_i \int_{\tilde{\phi}_{1j}-\frac{1}{2}g_j^{-1}}^{\tilde{\phi}_{1j}+\frac{1}{2}g_j^{-1}} \frac{1}{12} g(\tilde{\phi}_1)^{-2} \sigma^2(\tilde{\phi}_1) f(\tilde{\phi}_1) d\tilde{\phi}_1 + O(\Delta\tilde{\phi}^3) \\
&= \tilde{\theta}_1^2 \int \frac{1}{12} g(\tilde{\phi}_1)^{-2} \sigma^2(\tilde{\phi}_1) f(\tilde{\phi}_1) d\tilde{\phi}_1 + O(\Delta\tilde{\phi}^3), \qquad (59)
\end{aligned}
$$

where $\tilde{\phi}_{1j}$ is the assigned value I_j satisfying $\tilde{\phi}_{1j} \in I_j$. This says that

$$
\tilde{\theta}_1^2 \int \frac{1}{12} g(\tilde{\phi}_1)^{-2} \sigma^2(\tilde{\phi}_1) f(\tilde{\phi}_1) d\tilde{\phi}_1 \qquad (60)
$$

is an objective function when the assumption 4-2) is satisfied.

In the following we give the optimal quantizers, which minimize (60), under a constraint of the number of quantization steps (Section 4.1) or of the expectation of the code length where the quantized data is optimally encoded (Section 4.2). The former case is referred as "fixed-rate quantization" because it is identical to a "fixed-code length" case, on the other hand, the latter case is referred as "variable-rate quantization" and in fact the code length is not fixed.

4.1 Fixed-rate Quantization

From the above observations, the original optimization problem of (28) (i.e. (60)) can be replaced by the following at $N \to \infty$ and high resolution case:

Problem 4.1

$$
g_{\mathrm{f}}(\tilde{\phi}_1) := \arg\min_g \int \mathcal{F}(g(\tilde{\phi}_1)) d\tilde{\phi}_1 \qquad (61)
$$

$$
\textit{s.t.} \qquad \int_{\tilde{\phi}_1^{\min}}^{\tilde{\phi}_1^{\max}} g(\tilde{\phi}_1) d\tilde{\phi}_1 = M, \qquad (62)
$$

where

$$
\mathcal{F}(g(\tilde{\phi}_1)) := \frac{1}{12} \tilde{\theta}_1^2 g(\tilde{\phi}_1)^{-2} \sigma^2(\tilde{\phi}_1) f(\tilde{\phi}_1). \qquad (63)
$$

We can derive the next result.

Theorem 4.1 *The solution of (61) is:*

$$
\begin{aligned}
g_{\mathrm{f}}(\tilde{\phi}_1) &= K \sigma^{\frac{2}{3}}(\tilde{\phi}_1) f^{\frac{1}{3}}(\tilde{\phi}_1) && (64) \\
K &= D^{-1} M && (65) \\
D &= \int \sigma^{\frac{2}{3}}(\tilde{\phi}_1) f^{\frac{1}{3}}(\tilde{\phi}_1) d\tilde{\phi}_1. && (66)
\end{aligned}
$$

Moreover, the optimized value is given by

$$\int \mathcal{F}(g_{\rm f}(\tilde{\phi}_1)G_{\rm f}(\tilde{\phi}_1))d\tilde{\phi}_1 = \frac{1}{12}\tilde{\theta}_1^2 D^3 M^{-2}. \tag{67}$$

Note 4.1 If the corresponding distribution density of the number of quantized subsections $g(y)$ on y is required, from the equivalence $y = \tilde{\theta}_1\tilde{\phi}_1$ and the corresponding definitions

$$\begin{aligned}
&f_y(y(i), y(i+1), \ldots, y(i+n-1))(\leftrightarrow f_(\tilde{\phi})),\\
&f(y)(\leftrightarrow f(\tilde{\phi}_1)),\\
&\sigma(y)(\leftrightarrow \sigma(\tilde{\phi}_1)),
\end{aligned}$$

we can derive the similar results

$$\begin{aligned}
\mathcal{F}(g(y)) &= \frac{1}{12}g(y)^{-2}\sigma^2(y)f(y)\\
g_{\rm f}(y) &= K\sigma^{\frac{2}{3}}(y)f^{\frac{1}{3}}(y)\\
K &= D^{-1}M\\
D &= \int \sigma^{\frac{2}{3}}(y)f^{\frac{1}{3}}(y)dy\\
\int \mathcal{F}(g_{\rm f}(y),\ G_{\rm f}(y))dy &= \frac{1}{12}D^3M^{-2}.
\end{aligned}$$

Proof of Theorem 4.1 By employing the similar technique in [1, 13], the optimal solution can be given. With the calculus of variations, the following Euler–Lagrange's equation:

$$\frac{d}{d\tilde{\phi}_1}\left(\frac{\partial \mathcal{F}}{\partial g}\right) - \frac{\partial \mathcal{F}}{\partial G} = 0,$$

where

$$G(\tilde{\phi}_1) := \int_{\tilde{\phi}_1^{\min}}^{\tilde{\phi}_1} g(\tilde{\phi}_1)d\tilde{\phi}_1,$$

gives a differential equation:

$$\frac{d}{d\tilde{\phi}_1}\left(-2g(\tilde{\phi}_1)^{-3}\sigma^2(\tilde{\phi}_1)f(\tilde{\phi}_1)\right) = 0, \tag{68}$$

and the solution is:

$$g(\tilde{\phi}_1) = K\sigma^{\frac{2}{3}}(\tilde{\phi}_1)f^{\frac{1}{3}}(\tilde{\phi}_1),\ \ K: \text{constant}.$$

The constant number K is directly calculated by the condition (62), and the value of the objective function is derived as follows.

$$\begin{aligned}
\int \mathcal{F}(g_{\rm f}(\tilde{\phi}_1))d\tilde{\phi}_1 &= \int \frac{1}{12}\tilde{\theta}_1^2(K\sigma^{\frac{2}{3}}(\tilde{\phi}_1)f^{\frac{1}{3}}(\tilde{\phi}_1))^{-2}\sigma^2(\tilde{\phi}_1)f(\tilde{\phi}_1)d\tilde{\phi}_1\\
&= \int \frac{1}{12}\tilde{\theta}_1^2 K^{-2}\sigma^{\frac{2}{3}}(\tilde{\phi}_1)f(\tilde{\phi}_1)^{\frac{1}{3}}d\tilde{\phi}_1 = \frac{1}{12}\tilde{\theta}_1^2 K^{-2}D\\
&= \frac{1}{12}\tilde{\theta}_1^2 D^3 M^{-2}
\end{aligned} \tag{69}$$

□

From this result, the asymptotic optimal quantizations at high resolution case are easily calculated analytically or numerically if the marginal distributions $f(\tilde{\phi}_1)$ are known.

Note 4.2 When $f(\tilde{\phi})$ is a multidimensional normal distribution:

$$\begin{aligned}
f(\tilde{\phi}_1, \tilde{\phi}_2, \ldots, \tilde{\phi}_n) &= \frac{1}{(2\pi)^{\frac{n}{2}}(\det\Gamma)^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\tilde{\phi}^{\rm T}\Gamma^{-1}\tilde{\phi}\right),\\
\Gamma &= \mathrm{diag}(\sigma_o, \sigma_o, \ldots, \sigma_o),
\end{aligned}$$

then

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

Moreover consider a case that the order n of the MA model is enough large, then, in the area at $f(\tilde{\phi})$ has an enough large value (i.e., $\tilde{\phi}_1$ is not large), the variation of $\sigma(\tilde{\phi}_1)$ is relatively small and

$$\sigma(\tilde{\phi}_1) \sim n^{\frac{1}{2}}\sigma_o. \tag{70}$$

Therefore,

$$D \sim n^{\frac{1}{3}}\sigma_o^{\frac{2}{3}} \int f^{\frac{1}{3}}(\tilde{\phi}_1)d\tilde{\phi}_1$$

and

$$\begin{aligned} g_{\mathrm{f}}(\tilde{\phi}_1) &\sim M\left(\int f^{\frac{1}{3}}(\tilde{\phi}_1)d\tilde{\phi}_1\right)^{-1} f^{\frac{1}{3}}(\tilde{\phi}_1), \\ \int \mathcal{F}(g_{\mathrm{f}}(\tilde{\phi}_1))d\tilde{\phi}_1 &\sim \frac{1}{12}\tilde{\theta}_1^2\left(\int f^{\frac{1}{3}}(\tilde{\phi}_1)d\tilde{\phi}_1\right)^3 n\sigma_o^2 M^{-2} \\ &= \frac{1}{12}\tilde{\theta}_1^2 6\sqrt{3}\pi n\sigma_o^4 M^{-2} \sim 0.8658\pi\tilde{\theta}_1^2 n\sigma_o^4 M^{-2}. \end{aligned} \tag{71}$$

Note 4.3 *Here we consider a optimized function:*

$$\mathrm{E}\left[\tilde{\phi}_1^2 \cdot e^2\right] = \int \tilde{\phi}_1^2 e^2(\tilde{\phi}_1) f(\tilde{\phi}_1)d\tilde{\phi}_1.$$

Then, the optimal quantization $g'_{\mathrm{f}}(\tilde{\phi}_1)$ for the above is also given by

$$\begin{aligned} g'_{\mathrm{f}}(\tilde{\phi}_1) &\sim K\tilde{\phi}_1^{\frac{2}{3}} f^{\frac{1}{3}}(\tilde{\phi}_1) \\ K &\sim D^{-1}M \\ D &\sim \int_{\mathcal{R}\backslash[-\epsilon,\,\epsilon]} \tilde{\phi}_1^{\frac{2}{3}} f^{\frac{1}{3}}(\tilde{\phi}_1)d\tilde{\phi}_1, \end{aligned}$$

where $g'_{\mathrm{f}}(\tilde{\phi}_1)$ is defined only for the region $\mathcal{R}\backslash[-\epsilon,\epsilon]$, $\epsilon \ll 1$, because $K\tilde{\phi}_1^{\frac{2}{3}} f^{\frac{1}{3}}(\tilde{\phi}_1)$ is too small in $[-\epsilon,\epsilon]$ to apply the approximation of high resolution case. On the other hand, when $f(\tilde{\phi})$ is normal distribution, uniform distribution or other probable cases, the marginal density $f(\tilde{\phi}_1)$ is approximately uniform around the origin $[-\epsilon,\epsilon]$. Therefore, the optimal quantization $g'_{\mathrm{f}}(\tilde{\phi}_1)$ in the region $[-\epsilon,\epsilon]$ is similar to the solution derived in Section 3. From this reason, the result in Section 3 also indispensable for constructing the optimal quantization in high resolution case.

We illustrate $g_{\mathrm{f}}(\tilde{\phi}_1)$ for the cases that $f(\tilde{\phi}_1)$ is uniform distribution, normal distribution and power law as follows.

In Section 3, we derived the strictly optimal quantization for general resolution case when $f(\tilde{\phi}_1)$ is uniform distribution. Lemma 3.2 shows that the optimal quantization is coarse around the origin of $\tilde{\phi}_1$ and dense near the boundary of $\tilde{\phi}_1$. Such property of the optimal quantization can be also seen in Theorem 4.1 (see Fig. 9). Fig. 9 is an example of a simple case $\sigma(\tilde{\phi}_1) = \tilde{\phi}_1$, and the theorem shows that the growing rate of the resolution against $\tilde{\phi}_1$ is known when $\sigma(\tilde{\phi}_1)$ is given analytically. In this case, the order of the growing rate is $\tilde{\phi}_1^{\frac{2}{3}}$, which is unknown from the results of the previous section.

In the case that $f(\tilde{\phi}_1)$ is normal distribution, the profile of the density $f(\tilde{\phi}_1)$ around the origin is flat, therefore, the optimal quantizer must have the similar profile for the case that $\tilde{\phi}_1$ is uniform distribution around the origin. That is, the resolution grows around it, and we can see such profile of $g_{\mathrm{f}}(\tilde{\phi}_1)$ in Fig. 10. On the other hand, in the area of the tail of $f(\tilde{\phi}_1)$, $g_{\mathrm{f}}(\tilde{\phi}_1)$ goes down, however, against our intuition, the resolution remains high such as $g_{\mathrm{f}}(3) \sim 0.201 \sim 51\%$ of $\max g_{\mathrm{f}}(\tilde{\phi}_1)$ or $g_{\mathrm{f}}(4) \sim 0.0758 \sim 19\%$ of $\max g_{\mathrm{f}}(\tilde{\phi}_1)$, where $f(\tilde{\phi}_1)$ is enough small.

Finally we show the case of $f(\tilde{\phi}_1) \sim \tilde{\phi}_1^{-2}$ at the tail of the distribution as an example of power law. In this case, g_{f} is constant and it is marginal for the existence of the solution (see Fig. 11). This result shows the difficulty of the system identification in an enough accuracy by using finite information on the system when the tail of the distribution density $f(\tilde{\phi}_1)$ is heavier than $O(\tilde{\phi}_1^{-2})$. In other word, it explains the complexity of power law from the view point of parameter estimation of system identification.

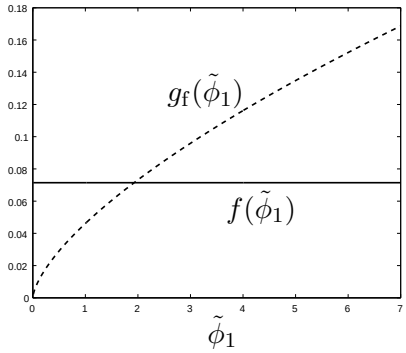

Fig. 9: Uniform distribution density $f(\tilde{\phi}_1)$ of the regressor (solid line) and the distribution density of the number of the optimally quantized subsections $g_{\mathrm{f}}(\tilde{\phi}_1)$ (dashed line) in the case $\sigma(\tilde{\phi}_1) = \tilde{\phi}_1$

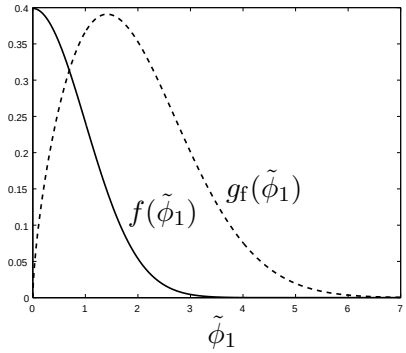

Fig. 10: Normal distribution density $f(\tilde{\phi}_1)$ of the regressor (solid line) and the distribution density of the number of the optimally quantized subsections $g_{\mathrm{f}}(\tilde{\phi}_1)$ (dashed line) in the case $\sigma(\tilde{\phi}_1) = \tilde{\phi}_1$

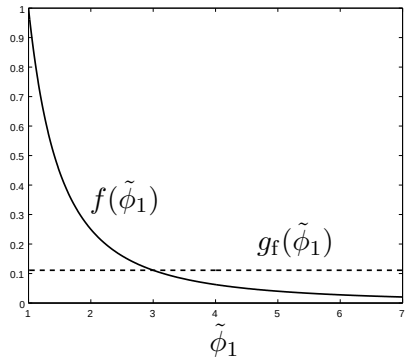

Fig. 11: Power law ($O(\tilde{\phi}_1^{-2})$) case of $f(\tilde{\phi}_1)$ of the regressor (solid line) and the distribution density of the number of the optimally quantized subsections $g_{\mathrm{f}}(\tilde{\phi}_1)$ (dashed line) in the case $\sigma(\tilde{\phi}_1) = \tilde{\phi}_1$

4.2 Variable-rate Quantization

In the previous subsection, we derive an optimal quantizer to minimize the identification error (28) (i.e. (60)) under constraint of the number of quantization steps, i.e., fixed-rate quantization, in the case of high resolution. On the other hand, for the purpose to reduce the information of the observed data from the identified system, it is reasonable to apply variable-rate coding for the quantized signals and measure the mean code length as the quantity of the information. According to this observation, we consider the minimization problem of (28) (i.e. (60)) under constraint of the expectation of the optimal code length, that is, variable-rate quantization, in high resolution case.

Let $C(\cdot)$ be an encoder which is a mapping from source alphabets to code alphabets and $l(\cdot)$ 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 variable-rate optimal code length for a quantized signal has relation with the entropy of the source alphabets from the following well-known source coding theorem.

Proposition 4.1 [16, 4] *Let x be source alphabets, then,*

$$\mathrm{E}\left[l(C(x))\right] \geq H(x), \tag{72}$$

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 (28) (i.e. (60)) under constraint of entropy of the quantized signals.

The basic idea to represent the quantizer in high resolution case is the same of the previous subsection. That is, under the assumption 4-5) and Assumption 4.2, we can get the asymptotic approximation of the entropy of the quantized signal:

$$\begin{aligned} H(f,g) &:= \sum_j -p_j \log p_j \\ &= \sum_j -\int_{I_j} f(\tilde{\phi}_1) d\tilde{\phi}_1 \log f_j g_j^{-1} \\ &\sim \int -f(\tilde{\phi}_1) \log\left(f(\tilde{\phi}_1) g^{-1}(\tilde{\phi}_1)\right) d\tilde{\phi}_1 \\ &= H_{\mathrm{d}}(f) + \int -f(\tilde{\phi}_1) \log\left(g^{-1}(\tilde{\phi}_1)\right) d\tilde{\phi}_1, \end{aligned} \tag{73}$$

where $H_{\mathrm{d}}(f) := \int -f(\tilde{\phi}_1)\log f(\tilde{\phi}_1)d\tilde{\phi}_1$. By using this asymptotic approximation of the entropy (73), we consider the following problem.

$$g_{\mathrm{v}}(\tilde{\phi}_1) := \arg\min_g \int \mathcal{F}(g(\tilde{\phi}_1))d\tilde{\phi}_1 \tag{74}$$

$$\text{s.t. } H(f,g) = \log M \tag{75}$$

Note that M is an expected number of quantization steps in the sense of (75).

We can derive the next result.

Theorem 4.2 *The solution of (74) is:*

$$\begin{aligned} g_{\mathrm{v}}(\tilde{\phi}_1) &= KM\sigma(\tilde{\phi}_1) && (76)\\ K &= \exp L && (77)\\ L &:= -H(f) - \int f\log\sigma(\tilde{\phi}_1)d\tilde{\phi}_1 \\ &= \int f(\tilde{\phi}_1)\log\frac{f(\tilde{\phi}_1)}{\sigma(\tilde{\phi}_1)}d\tilde{\phi}_1 && (78)\end{aligned}$$

Moreover, the optimized value is given by

$$\int \mathcal{F}(g_{\mathrm{v}}(\tilde{\phi}_1))d\tilde{\phi}_1 = \frac{1}{12}\tilde{\theta}_1^2 K^{-2}M^{-2}. \tag{79}$$

Proof We employ the similar technique in [10, 2]. Let λ be a Lagrange multiplier and consider the minimization problem of the following quantity.

$$\begin{aligned} &\int \mathcal{F}(g(\tilde{\phi}_1))d\tilde{\phi}_1 + \lambda H(f,g) \\ &= \int \frac{1}{12}\tilde{\theta}_1^2\left(\frac{1}{g(\tilde{\phi}_1)}\right)^2\sigma^2(\tilde{\phi}_1)f(\tilde{\phi}_1) - \lambda f(\tilde{\phi}_1)\log\left(g^{-1}(\tilde{\phi}_1)\right)d\tilde{\phi}_1 + \lambda H(f) \\ &= \int \frac{1}{12}\tilde{\theta}_1^2 f(\tilde{\phi}_1)\left(g^{-2}(\tilde{\phi}_1)\sigma^2(\tilde{\phi}_1) + \lambda\log g(\tilde{\phi}_1)\right)d\tilde{\phi}_1 + \lambda H(f) \end{aligned} \tag{80}$$

By applying Eular–Lagrange's differential equation, we get

$$\begin{aligned} \frac{\partial}{\partial g}\left(g^{-2}\sigma^2(\tilde{\phi}_1) + \lambda\log g\right) &= -2g^{-3}\sigma^2(\tilde{\phi}_1) + \lambda g^{-1} \\ &= \text{constant}. \end{aligned} \tag{81}$$

Fix the constant to be zero, then,

$$g = \left(\frac{2}{\lambda}\right)^{\frac{1}{2}}\sigma(\tilde{\phi}_1), \tag{82}$$

and by substituting it for $H(f,\, g)$, we get

$$\begin{aligned} H(f,\, g) &= \int -f\log g^{-1}f d\tilde{\phi}_1 \\ &= \log\left(\frac{2}{\lambda}\right)^{\frac{1}{2}} + \int -f\log\frac{f}{\sigma(\tilde{\phi}_1)}d\tilde{\phi}_1 \\ &= \log M. \end{aligned} \tag{83}$$

Therefore,

$$\left(\frac{2}{\lambda}\right)^{\frac{1}{2}} = \exp\left(\int f\log\frac{f}{\sigma(\tilde{\phi}_1)}d\tilde{\phi}_1 + \log M\right), \tag{84}$$

and (76) is derived. By substituting $g_{\rm v}$ for the objective integral, the following is derived.

$$\begin{aligned}\int \frac{1}{12}\tilde{\theta}_1^2 g_{\rm v}^{-2}(\tilde{\phi}_1)\sigma^2(\tilde{\phi}_1)f(\tilde{\phi}_1)d\tilde{\phi}_1 &= \frac{1}{12}\tilde{\theta}_1^2\frac{\lambda}{2} \\ &= \frac{1}{12}\tilde{\theta}_1^2 K^{-2}M^{-2} \end{aligned} \tag{85}$$

□

Note 4.4 Interesting fact is that the optimal $g_{\rm v}$ is a simple linear function of $\sigma(\tilde{\phi}_1)$. The constant coefficient is also linear to the number of the quantization steps M. On the other hand, the convergence rate of the minimized variance of the quantization error term is M^{-2} and this fact is common with the fixed-rate quantization.

Note 4.5 When $f_{\tilde{\phi}}$ is a multidimensional normal distribution and n is large as discussed in Note 4.2, by using (70),

$$\begin{aligned} L &\sim -H(f) - \log(\sigma_o n^{\frac{1}{2}}), \\ K &\sim \exp(-H(f)) \cdot (\sigma_o n^{\frac{1}{2}})^{-1}, \end{aligned}$$

and

$$\begin{aligned} g_{\rm v}(\tilde{\phi}_1) &= KM\sigma(\tilde{\phi}_1) \\ &\sim M \cdot \exp(-H(f)) \cdot (\sigma_o n^{\frac{1}{2}})^{-1} \cdot \sigma_o n^{\frac{1}{2}} \\ &= M \cdot \exp(-H(f)) \end{aligned}$$

$$\begin{aligned}\int \mathcal{F}(g_{\rm v}(\tilde{\phi}_1))d\tilde{\phi}_1 &\sim \frac{1}{12}\tilde{\theta}_1^2 \exp(2H(f))n\sigma_o^2 M^{-2} \\ &= \frac{1}{12}\tilde{\theta}_1^2 2e\pi n\sigma_o^4 M^{-2} \sim 0.4533\pi\tilde{\theta}_1^2 n\sigma_o^4 M^{-2}. \end{aligned} \tag{86}$$

The comparison of (71) and (86) tells us that the case of the variable-rate optimal coding attains about a half magnitude of the variance of the quantization error compared with $g_{\rm f}$ for fixed-rate quantization.

5 Pre-quantizer Noise Case

In Section 2, we showed two forms of exogenius noise w as (1) and (2). In the case of (1), the relationship between noise error term ΔW and the quantization error term ΔE is simple since they are independent and we can simply evaluate their magnitude separately. On the other hand, in the case of the pre-quantizer noise (2), although which is more realistic case, the effects of noise and quantization on the identification error are complexly correlated each other and their evaluation is not straightforward. This shows the necessity of the analysis on the effect of the correlation between noise and quantization on the parameter error for the case of (2).

For the system (2), we define $\hat{y}$ and the error e_q between $\hat{y}$ and y as follows.

$$\hat{y} := q(y),\ e_q := \hat{y} - y \tag{87}$$

The above $\hat{y}$ and e_q can be regarded as an imaginary quantized signal and the corresponding quantization error. The error between y_o defined in (1) and $\hat{y}$:

$$e_w := y_o - \hat{y} \tag{88}$$

can be regarded as an imaginary noise. Therefore, the observed signal y_o can be represented by the imaginary quantized error e_q and the imaginary noise e_w as

$$y_o = y + e_q + e_w. \tag{89}$$

In order to evaluate the essential characteristics of the effect of noise on parameter error, here we deal with the quantity:

$$\mathrm{V}[\tilde{\phi}_1(e_q + e_w)] = \mathrm{E}[\tilde{\phi}_1^2(e_q^2 + e_w^2 + 2e_q e_w)]. \tag{90}$$

Note that the imaginary quantization error e_q is definitely given on the event in $\tilde{\phi}_1$, on the other hand, the imaginary noise error e_w is probabilistically realized and its distribution density depends on $\tilde{\phi}_1$. Therefore, we get

$$\begin{aligned} \mathrm{V}\left[\tilde{\phi}_1(e_q + e_w)\right] &= \mathrm{E}\left[\tilde{\phi}_1^2(e_q^2 + e_w^2 + 2e_q e_w)\right] \\ &= \mathrm{E}\left[\tilde{\phi}_1^2\left(e_q^2(\tilde{\phi}_1) + \mathrm{E}_{\tilde{\phi}_1}\left[e_w^2\right] + 2e_q(\tilde{\phi}_1)\mathrm{E}_{\tilde{\phi}_1}\left[e_w\right]\right)\right], \end{aligned} \tag{91}$$

where

$$\mathrm{E}_{\tilde{\phi}_1}[x] := \int x dP(x|\tilde{\phi}_1) \tag{92}$$

and $P(x|\tilde{\phi}_1)$ is a conditional probability of x given $\tilde{\phi}_1$. We here assume that w is a random noise obeying uniform distribution of a section $[-\epsilon,\ \epsilon]$. Let $\bar{y}_j$ be the middle point of S_j for simplicity and we calculate hereafter two terms $\mathrm{E}_{\tilde{\phi}_1}[e_w^2]$ and $\mathrm{E}_{\tilde{\phi}_1}[e_w]$ which depend on e_w in the right hand side of (91).

Assume $\tilde{\phi}_1(i)$ is in a subsection I_j of width $\tilde{\theta}_1^{-1}\delta$ (i.e. $y(i) \in S_j = (d_{j-1}, d_j]$ of width δ) and satisfies $\tilde{\theta}_1\tilde{\phi}_1 = \frac{d_{j-1}+d_j}{2} + h$. Moreover, assume that ϵ and δ have a relation $\epsilon = (\frac{1}{2} + s)\delta$ where s is an integer for simplicity of the following analysis and ϵ is enough small such that $|d_j - d_{j-1}|$, $|d_{j+1} - d_j|$, $|d_{j+2} - d_{j+1}|$, $\ldots$, can be considered to be a constant δ in the region $[\bar{y}_j - \epsilon,\ \bar{y}_j + \epsilon]$. Then, in the case $h > 0$, we can derive

$$\begin{aligned} &\mathrm{E}_{\tilde{\phi}_1}\left[e_w^2\right] \\ &= P(\tilde{\theta}_1\tilde{\phi}_1 + w \in S_{j-s}) \cdot e_w^2|_{\tilde{\theta}_1\tilde{\phi}_1 + w \in S_{j-s}} + P(\tilde{\theta}_1\tilde{\phi}_1 + w \in S_{j-(s-1)}) \cdot e_w^2|_{\tilde{\theta}_1\tilde{\phi}_1 + w \in S_{j-(s-1)}} + \cdots \\ &\quad + P(\tilde{\theta}_1\tilde{\phi}_1 + w \in S_{j+(s-1)}) \cdot e_w^2|_{\tilde{\theta}_1\tilde{\phi}_1 + w \in S_{j+(s-1)}} + P(\tilde{\theta}_1\tilde{\phi}_1 + w \in S_{j+s}) \cdot e_w^2|_{\tilde{\theta}_1\tilde{\phi}_1 + w \in S_{j+s}} \\ &\quad + P(\tilde{\theta}_1\tilde{\phi}_1 + w \in S_{j+(s+1)}) \cdot e_w^2|_{\tilde{\theta}_1\tilde{\phi}_1 + w \in S_{j+(s+1)}} \\ &= \int_{-(\frac{1}{2}+s)\delta+h}^{-(\frac{1}{2}+s-1)\delta} ((-s)\delta)^2 \frac{1}{2\epsilon} d(\tilde{\theta}_1\tilde{\phi}_1 + w) + \int_{-(\frac{1}{2}+s-1)\delta}^{-(\frac{1}{2}+s-2)\delta} ((-(s-1))\delta)^2 \frac{1}{2\epsilon} d(\tilde{\theta}_1\tilde{\phi}_1 + w) \\ &\quad + \cdots + \int_{(\frac{1}{2}+s-2)\delta}^{(\frac{1}{2}+s-1)\delta} ((s-1)\delta)^2 \frac{1}{2\epsilon} d(\tilde{\theta}_1\tilde{\phi}_1 + w) + \int_{(\frac{1}{2}+s-1)\delta}^{(\frac{1}{2}+s)\delta} (s\delta)^2 \frac{1}{2\epsilon} d(\tilde{\theta}_1\tilde{\phi}_1 + w) \\ &\quad + \int_{(\frac{1}{2}+s)\delta}^{(\frac{1}{2}+s+1)\delta+h} ((s+1)\delta)^2 \frac{1}{2\epsilon} d(\tilde{\theta}_1\tilde{\phi}_1 + w) \\ &= \left\{(1^2 + 2^2 + \cdots + s^2) \cdot 2\delta + ((s+1)^2 - s^2) \cdot h\right\} \frac{\delta^2}{2\epsilon} \\ &= a(s,\delta) + b(\delta, h) \end{aligned} \tag{93}$$

where

$$\begin{aligned} a(s,\delta) &:= \left\{\left(\frac{1}{s}\right)^2 + \left(\frac{2}{s}\right)^2 + \cdots + \left(\frac{s}{s}\right)^2\right\} \frac{1}{s + \frac{1}{2}} \cdot (s\delta)^2 \\ b(\delta, h) &:= \delta \cdot h. \end{aligned}$$

On the contrary, when $h < 0$, we can derive

$$\mathrm{E}_{\tilde{\phi}_1}\left[e_w^2\right] = a(s,\delta) - b(\delta, h). \tag{94}$$

For the other terms which depends on e_w, we can also derive the following for $h > 0$.

$\mathrm{E}_{\tilde{\phi}_1}\left[e_w\right]$

$$
\begin{aligned}
&= P(\tilde{\theta}_1\tilde{\phi}_1 + w \in S_{j-s}) \cdot e_w|_{\tilde{\theta}_1\tilde{\phi}_1+w\in S_{j-s}} + P(\tilde{\theta}_1\tilde{\phi}_1 + w \in S_{j-(s-1)}) \cdot e_w|_{\tilde{\theta}_1\tilde{\phi}_1+w\in S_{j-(s-1)}} + \cdots \\
&\quad + P(\tilde{\theta}_1\tilde{\phi}_1 + w \in S_{j+(s-1)}) \cdot e_w|_{\tilde{\theta}_1\tilde{\phi}_1+w\in S_{j+(s-1)}} + P(\tilde{\theta}_1\tilde{\phi}_1 + w \in S_{j+s}) \cdot e_w|_{\tilde{\theta}_1\tilde{\phi}_1+w\in S_{j+s}} \\
&\quad + P(\tilde{\theta}_1\tilde{\phi}_1 + w \in S_{j+(s+1)}) \cdot e_w|_{\tilde{\theta}_1\tilde{\phi}_1+w\in S_{j+(s+1)}} \\
&= \int_{-(\frac{1}{2}+s)\delta+h}^{-(\frac{1}{2}+s-1)\delta} (-s)\delta\frac{1}{2\epsilon}d(\tilde{\theta}_1\tilde{\phi}_1 + w) + \int_{-(\frac{1}{2}+s-1)\delta}^{-(\frac{1}{2}+s-2)\delta} (-(s-1))\delta\frac{1}{2\epsilon}d(\tilde{\theta}_1\tilde{\phi}_1 + w) \\
&\quad + \cdots + \int_{(\frac{1}{2}+s-2)\delta}^{(\frac{1}{2}+s-1)\delta} (s-1)\delta\frac{1}{2\epsilon}d(\tilde{\theta}_1\tilde{\phi}_1 + w) + \int_{(\frac{1}{2}+s-1)\delta}^{(\frac{1}{2}+s)\delta} s\delta\frac{1}{2\epsilon}d(\tilde{\theta}_1\tilde{\phi}_1 + w) \\
&\quad + \int_{(\frac{1}{2}+s)\delta}^{(\frac{1}{2}+s+1)\delta+h} (s+1)\delta\frac{1}{2\epsilon}d(\tilde{\theta}_1\tilde{\phi}_1 + w) \\
&= \frac{(2s+1)\delta}{2\epsilon}h = h
\end{aligned} \tag{95}
$$

From the above, we can get the following which is used for the calculation of the right hand side of (91).

$$
e_q^2 + \mathrm{E}_{\tilde{\phi}_1}\left[e_w^2\right] + 2e_q \cdot \mathrm{E}_{\tilde{\phi}_1}\left[e_w\right] = (-h)^2 + a(s,\delta) + |b(\delta,h)| + 2(-h)\cdot h = a(s,\delta) + |\delta\cdot h| - h^2 \tag{96}
$$

When the quantization step width is enough small, $\tilde{\phi}_1$ is almost constant in a quantization subsection, therefore, we get an approximation of the following partial integral of (91):

$$
\begin{aligned}
&\int_{-\frac{\delta'}{2}}^{\frac{\delta'}{2}} \tilde{\phi}_1^2\left(a(s,\delta) + |\delta\cdot\tilde{\theta}_1 h'| - (\tilde{\theta}_1 h')^2\right) f(\tilde{\phi}_1)\tilde{\theta}_1 dh' \\
&= \tilde{\phi}_1^2 f(\tilde{\phi}_1)\tilde{\theta}_1\left(\left[a(s,\delta)\cdot h' + \frac{1}{2}\delta\cdot\tilde{\theta}_1 h'^2 - \frac{1}{3}\tilde{\theta}_1^2 h'^3\right]_0^{\frac{\delta'}{2}} + \left[a(s,\delta)\cdot h' - \frac{1}{2}\delta\cdot\tilde{\theta}_1 h'^2 - \frac{1}{3}\tilde{\theta}_1^2 h'^3\right]_{-\frac{\delta'}{2}}^{0}\right) \\
&= \tilde{\phi}_1^2 f(\tilde{\phi}_1)\tilde{\theta}_1\left(a(s,\delta)\delta' + \frac{1}{6}\tilde{\theta}_1^2\delta'^3\right),
\end{aligned} \tag{97}
$$

where $\delta' := \tilde{\theta}_1^{-1}\delta$ and $h' := \tilde{\theta}_1^{-1}h$.

By using the above result, (91) can be approximated as follows.

$$
\begin{aligned}
\mathrm{E}\left[\tilde{\phi}_1^2\left(e_q^2(\tilde{\phi}_1) + \mathrm{E}_{\tilde{\phi}_1}\left[e_w^2\right] + 2e_q(\tilde{\phi}_1)\mathrm{E}_{\tilde{\phi}_1}\left[e_w\right]\right)\right] &\sim \sum \tilde{\phi}_1^2 f(\tilde{\phi}_1)\tilde{\theta}_1(a(s,\delta) + \frac{1}{6}\tilde{\theta}_1^2\delta'^2)\delta' \\
&\sim \int \tilde{\phi}_1^2 f(\tilde{\phi}_1)(a(s,\tilde{\theta}_1 g^{-1}(\tilde{\phi}_1)) + \frac{1}{6}\tilde{\theta}_1^2 g^{-2}(\tilde{\phi}_1))d\tilde{\phi}_1
\end{aligned} \tag{98}
$$

On the other hand, in the noise-free case we get:

$$
\begin{aligned}
\mathrm{E}\left[\tilde{\phi}_1^2\left(e_q^2(\tilde{\phi}_1)\right)\right] + \mathrm{E}\left[\tilde{\phi}_1^2 w^2\right] &\simeq \sum_j \tilde{\phi}_1^2 f(\tilde{\phi}_1)\int_{-\frac{\delta}{2}}^{\frac{\delta}{2}} h^2 dh + \int \tilde{\phi}_1^2 f(\tilde{\phi}_1)\frac{1}{3}\epsilon^3 d\tilde{\phi}_1 \qquad (99) \\
&= \sum_j \tilde{\phi}_1^2 f(\tilde{\phi}_1)(\frac{1}{12}\delta^2)\delta + \int \tilde{\phi}_1^2 f(\tilde{\phi}_1)\frac{1}{3}\epsilon^3 d\tilde{\phi}_1 \\
&\simeq \int \tilde{\phi}_1^2 f(\tilde{\phi}_1)(\frac{1}{12}\tilde{\theta}_1^2 g^{-2}(\tilde{\phi}_1))d\tilde{\phi}_1 + \int \tilde{\phi}_1^2 f(\tilde{\phi}_1)\frac{1}{3}\epsilon^3 d\tilde{\phi}_1. \qquad (100)
\end{aligned}
$$

The results (98) and (100) show the effect of the pre-quantizer noise.

The term $\int \tilde{\phi}_1^2 f(\tilde{\phi}_1)a(s,\tilde{\theta}_1 g^{-1}(\tilde{\phi}_1))d\tilde{\phi}_1$ in (98) is an quantized version of the noise error $\int \tilde{\phi}_1^2 f(\tilde{\phi}_1)\frac{1}{3}\epsilon^3 d\tilde{\phi}_1$ in (100) and we can confirm that the former converges to the latter when $\delta \to 0$. The remainders $\int \tilde{\phi}_1^2 f(\tilde{\phi}_1)(\frac{1}{6}\tilde{\theta}_1^2 g^{-2}(\tilde{\phi}_1))d\tilde{\phi}$ and $\int \tilde{\phi}_1^2 f(\tilde{\phi}_1)(\frac{1}{12}\tilde{\theta}_1^2 g^{-2}(\tilde{\phi}_1))d\tilde{\phi}_1$ in (98) and (100) can be regarded as the equivalent quantization error and the interesting fact that the former is twice of the latter. This suggests that the pre-quantizer noise equivalently increases the magnitude of the imaginary quantization error twice compared with the post-quantizer noise case.

6 Resolution of Quantizer and I/O Data Length

By using the results in the previous sections, we evaluate the magnitudes of the error term ΔE and ΔW based on the approach in [20] and then, compare the effects of the resolution of quantizers and the I/O data length. First, we evaluate the magnitude of $(U^{\mathrm{T}}U)^{-1}$.

Lemma 6.1 [20] *Suppose that $u(i)$ are i.i.d. random variables with $\mathrm{E}[u(i)] = 0$, $\mathrm{V}[u(i)] = \sigma_u^2$, $\mathrm{V}[u^2(i)] = \eta$. Then, for any reliability index β_1, where $1 - \beta_1 > 0$, and $\sigma_u^2 N - n\sqrt{\frac{N}{\beta_1}}\left(\sqrt{\eta} + (n-1)\sigma_u^2\right) > 0$, the following inequality is satisfied.*

$$\mathrm{Prob}\left(\|(U^{\mathrm{T}}U)^{-1}\|_1 \geq \epsilon_1\right) \leq \beta_1 \tag{101}$$

$$\epsilon_1 := \frac{1}{\sigma_u^2 N - n\sqrt{\frac{N}{\beta_1}}\left(\sqrt{\eta} + (n-1)\sigma_u^2\right)} \tag{102}$$

When $u(i)$ has a uniform distribution: $u_i \in [-\kappa, \kappa]$, that is, $\sigma_u^2 = \frac{1}{3}\kappa^2$, $\eta = \frac{4}{45}\kappa^4$, then,

$$\epsilon_1 = \frac{1}{\kappa^2\left(\frac{1}{3}N - n\left(\sqrt{\frac{4}{45}} + \frac{1}{3}(n-1)\right)\sqrt{\frac{N}{\beta_1}}\right)}.$$

By employing Lemma 6.1, we can evaluate $|\Delta\tilde{E}_1|$ in the following theorem.

Theorem 6.1 *For the system (1) with the optimal quantizer $q(y)$ defined by (3) $\sim$ (5), (42) $\sim$ (45), assume Assumption 3.1. Then, for reliability indices β_1, β_2, a length of data N and the number of quantization levels $2M$ in $[-\tilde{\theta}_1\kappa,\ \tilde{\theta}_1\kappa]$, where $1 - \beta_1 - \beta_2 > 0$, $M \gg K'$, where K' is defined in (56), and $\sigma_{\tilde{\phi}_1}^2 N - n\sqrt{\frac{N}{\beta_1}}\left(\sqrt{\eta} + (n-1)\sigma_{\tilde{\phi}_1}^2\right) > 0$, where $\sigma_{\tilde{\phi}_1} = \frac{1}{3}\kappa^2$, the following inequality holds.*

$$\mathrm{Prob}\left(|\Delta\tilde{E}_1| \leq \epsilon_1\epsilon_2\right) \geq 1 - \beta_1 - \beta_2 \tag{103}$$

$$\epsilon_1 := \frac{1}{\sigma_{\tilde{\phi}_1}^2 N - n\sqrt{\frac{N}{\beta_1}}\left(\sqrt{\eta} + (n-1)\sigma_{\tilde{\phi}_1}^2\right)},\ \epsilon_2 := \frac{A^{\frac{1}{2}}\kappa^2}{M - K'}\sqrt{\frac{nN}{\beta_2}} \tag{104}$$

From this theorem, we know that the convergence rate of the error term $|\Delta\tilde{E}_1|$ has an order of M^{-1} at sufficiently large M and of $N^{-1/2}$. Approximately, the total amount of information on the quantized output transmitted from identified systems to the observers is about $N\log_2 2M$ using a binary coding. Therefore, under a constraint of such a total amount of information, a large M is preferable to large N. Of course, this fact is valid only for the error term ΔE_1 and the situation is different for the noise error term ΔW. We introduce the result for ΔW in the following proposition.

Proposition 6.1 [20] *Suppose that $u(i)$ and $w(i)$ are i.i.d. random variables with $\mathrm{E}[u(i)] = 0$, $\mathrm{V}[u(i)] = \sigma_u^2$, and $\mathrm{V}[w(i)] = \sigma_w^2$, respectively. Then, for reliability indices β_1, β_2, and a length of data N, where $1 - \beta_1 - \beta_2 > 0$, and $\sigma_u^2 N - n\sqrt{\frac{N}{\beta_1}}\left(\sqrt{\eta} + (n-1)\sigma_u^2\right) > 0$, the following inequality holds.*

$$\mathrm{Prob}\left(\|\Delta W\|_\infty \leq \epsilon_1\epsilon_2\right) \geq 1 - \beta_1 - \beta_2 \tag{105}$$

$$\epsilon_1 := \frac{1}{\sigma_u^2 N - n\sqrt{\frac{N}{\beta_1}}\left(\sqrt{\eta} + (n-1)\sigma_u^2\right)},\ \epsilon_2 := \sigma_u\sigma_w\sqrt{\frac{nN}{\beta_2}} \tag{106}$$

By combining Theorem 6.1 and Proposition 6.1, we conclude there exists a trade-off between ΔE and ΔW for reducing the total identification error under the constraint of the amount of information transmitted from the identified systems to the observers.

7 Conclusion

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

1) When the regressor vector obeys a kind of uniform distribution, the optimal quantization problem for system identification is reduced to a recursive minimization of 1-variable rational function (Section 3).

2) This quantizer is coarse around the origin of the output and goes to be dense apart from the origin (Section 3).

3) General cases of the distribution of regressor vector can be dealt under a condition of high resolution quantizer by introducing a notion of the density of quantization subsections (Section 4).

4) The above optimization problem is reduced to a minimization of a functional and the solution can be given by solving Eular–Lagrange's differential equation (Section 4).

5) The pre-quantizer noise equivalently increases the magnitude of the quantization error twice compared with the post-quantizer noise.

6) Under a limitation of the total quantity of information of the quantized I/O data, there exists a trade-off between the magnitudes of the quantization error and noise error.

In this paper, we restrict the model to SISO MA model. For more realistic situation, we should extend the results to a) ARMA model, or MIMO system, b) quantized input signal, c) on-line system identification, adaptive control, and these are left to the future work.

References

[1] W. R. Bennett. Spectra of quantized signals. *The Bell System Technical Journal*, 27:446–472, 1948.

[2] T. Berger. Optimum quantizers and permutation codes. *IEEE Transactions on Information Theory*, IT-18-6:759–765, 1972.

[3] R. W. Brockett and D. Liberzon. Quantized feedback stabilization of linear systems. *IEEE Trans. Automat. Control*, AC-45-7:1279–1289, 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. Curry. *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*, 13:365–372, 1989.

[7] D. F. Delchamps. Stabilizing a linear system with quantized state feedback. *IEEE Trans. Automat. Control*, AC-35-8:916–924, 1990.

[8] N. Elia and S. K. Mitter. Stabilization of linear systems with limited information. *IEEE Trans. on Automatic Control*, AC-46-9:1384–1400, 2001.

[9] M. Gevers and G. Li. *Parametrization in control, estimation and filtering problems: Accuracy aspects*. Communications and control engineering series. Springer-Verlag, Berlin, 1993.

[10] H. Gish and J. N. Pierce. Asymptotically efficient quantizing. *IEEE Transactions on Information Theory*, IT-14-5:676–683, 1968.

[11] R. M. Gray and D. L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, IT-44-6:1–63, 1998.

[12] L. Ljung. *System Identification - Theory for the User*. Information and System Sciences Series. Prentice-Hall, Englewood Cliffs, N.J., 1987.

[13] S. P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, IT-28-2:129–137, 1982.

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

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

[16] C. E. Shannon. A mathematical theory of communication. *Bell Sys. Tech. Journal*, 27:379–423,623–656, 1948.

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

[18] 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).

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

[20] K. Tsumura and Y. Oishi. Optimal length of data for identification of time varying system. In *Proceedings of the 38th C.D.C.*, pages 3224–3229, 1999.

[21] W. S. Wong and R. W. Brockett. Systems with finite communication bandwidth constraints – part I: State estimation problems. *IEEE Trans. Automat. Control*, AC-42-9:1294–1299, 1997.

[22] W. S. Wong and R. W. Brockett. Systems with finite communication bandwidth constraints – II: Stabilization with limited information feedback. *IEEE Trans. Automat. Control*, AC-44-5:1049–1053, 1999.

A Appendix

Proposition A.1 (Chebyshev's inequality (see [12])) *Let x be an independent random variable and* Prob $(x^2) = \sigma_x^2$. *Then for any $c > 0$,*

$$\text{Prob}\left(|x - \text{E}[x]| \geq c\sigma_x\right) \leq \frac{1}{c^2}. \tag{107}$$

Proof of Lemma 2.1

The left hand side of (26) is extended as:

$$\begin{aligned}\text{E}\left[\left(\sum_{i=1}^{N}\tilde{\phi}_k(i)e(\tilde{\phi}_1(i))\right)^2\right] &= \text{E}\left[\sum_{i=1}^{N}\tilde{\phi}_k^2(i)e^2(\tilde{\phi}_1(i))\right] + \text{E}\left[\sum_{i=1}^{N}\tilde{\phi}_k(i)e(\tilde{\phi}_1(i))\tilde{\phi}_k(i+1)e(\tilde{\phi}_1(i+1))\right] + \cdots \\ &= N\text{E}\left[\tilde{\phi}_k^2e^2(\tilde{\phi}_1)\right] + 2(N-1)\text{E}\left[\tilde{\phi}_ke(\tilde{\phi}_1)\tilde{\phi}_{k+1}e(\tilde{\phi}_2)\right] + \cdots. \end{aligned} \tag{108}$$

In (108), terms of the form $\text{E}\left[\tilde{\phi}_he(\tilde{\phi}_i)\tilde{\phi}_je(\tilde{\phi}_k)\right]$ appear and in general, when (25) holds, $\text{E}\left[\tilde{\phi}_he(\tilde{\phi}_i)\tilde{\phi}_je(\tilde{\phi}_k)\right]$ can be calculated as follows.

In the case of $h \neq i \neq j \neq k$,

$$\begin{aligned}
\mathrm{E}\left[\tilde{\phi}_h e(\tilde{\phi}_i)\tilde{\phi}_j e(\tilde{\phi}_k)\right] &= \int \tilde{\phi}_h e(\tilde{\phi}_i)\tilde{\phi}_j e(\tilde{\phi}_k) f(\tilde{\phi}_h, \tilde{\phi}_i, \tilde{\phi}_j, \tilde{\phi}_k) d\tilde{\phi}_h d\tilde{\phi}_i d\tilde{\phi}_j d\tilde{\phi}_k \\
&= \int e(\tilde{\phi}_i)\tilde{\phi}_j e(\tilde{\phi}_k) \left(\int \tilde{\phi}_h f(\tilde{\phi}_h, \tilde{\phi}_i, \tilde{\phi}_j, \tilde{\phi}_k) d\tilde{\phi}_h \right) d\tilde{\phi}_i d\tilde{\phi}_j d\tilde{\phi}_k \\
&= \int e(\tilde{\phi}_i)\tilde{\phi}_j e(\tilde{\phi}_k) \times 0 \times d\tilde{\phi}_i d\tilde{\phi}_j d\tilde{\phi}_k \\
&= 0, \qquad (109)
\end{aligned}$$

and also in the case of $h = i \neq j \neq k$,

$$\begin{aligned}
\mathrm{E}\left[\tilde{\phi}_h e(\tilde{\phi}_i)\tilde{\phi}_j e(\tilde{\phi}_k)\right] &= \int \tilde{\phi}_h e(\tilde{\phi}_h)\tilde{\phi}_j e(\tilde{\phi}_k) f(\tilde{\phi}_h, \tilde{\phi}_j, \tilde{\phi}_k) d\tilde{\phi}_h d\tilde{\phi}_j d\tilde{\phi}_k \\
&= \int \tilde{\phi}_h e(\tilde{\phi}_h) e(\tilde{\phi}_k) \left(\int \tilde{\phi}_j f(\tilde{\phi}_h, \tilde{\phi}_i, \tilde{\phi}_j, \tilde{\phi}_k) d\tilde{\phi}_j \right) d\tilde{\phi}_h d\tilde{\phi}_k \\
&= \int \tilde{\phi}_h e(\tilde{\phi}_h) e(\tilde{\phi}_k) \times 0 \times d\tilde{\phi}_h d\tilde{\phi}_k \\
&= 0. \qquad (110)
\end{aligned}$$

On the other hand, $h = j \neq i \neq k$ or $i = k \neq j \neq k$ is not the case of (108). Finally in the case of $h = j$, $i = k$,

$$\mathrm{E}\left[\tilde{\phi}_h e(\tilde{\phi}_i)\tilde{\phi}_j e(\tilde{\phi}_k)\right] = \mathrm{E}\left[\tilde{\phi}_h^2 e^2(\tilde{\phi}_i)\right]. \qquad (111)$$

The other cases are essentially equal to one of the above cases (for example, the case $h = k \neq i \neq j$ is equal to the case $h = i \neq j \neq k$).

From the above calculations, we get the following:

$$\mathrm{E}\left[\left(\sum_{i=1}^{N} \tilde{\phi}_k(i) e(\tilde{\phi}_1(i))\right)^2\right] = N\mathrm{E}\left[\tilde{\phi}_k^2 e^2(\tilde{\phi}_1)\right]. \qquad (112)$$

□

Proof of Lemma 2.2

The outline of the proof is similar to that of Lemma 2.1 and we evaluate the value of $\mathrm{E}\left[\tilde{\phi}_h e(\tilde{\phi}_i)\tilde{\phi}_j e(\tilde{\phi}_k)\right]$ for each possible case.

Let I^i, I^j, I^h, or I^k be a quantized subsection of the axis of $\tilde{\phi}_i$, $\tilde{\phi}_j$, $\tilde{\phi}_h$, or $\tilde{\phi}_k$ respectively and define a subset in the space of ϕ :

$$\mathcal{I} := \left\{ \tilde{\phi}_i, \tilde{\phi}_j, \tilde{\phi}_h, \tilde{\phi}_k | \tilde{\phi}_i \in I^i, \tilde{\phi}_j \in I^j, \tilde{\phi}_h \in I^h, \tilde{\phi}_k \in I^k \right\}.$$

Moreover let $\overline{\tilde{\phi}_i}$, $\overline{\tilde{\phi}_j}$, $\overline{\tilde{\phi}_h}$, and $\overline{\tilde{\phi}_k}$ be the quantized values which are middle points of I^i, I^j, I^h, and I^k respectively. The partial integral of $\mathrm{E}\left[\tilde{\phi}_h e(\tilde{\phi}_i)\tilde{\phi}_j e(\tilde{\phi}_k)\right]$ restricted to this subset is

$$\int_{\mathcal{I}} \tilde{\phi}_h e(\tilde{\phi}_i)\tilde{\phi}_j e(\tilde{\phi}_k) f(\tilde{\phi}_h, \tilde{\phi}_i, \tilde{\phi}_j, \tilde{\phi}_k) d\tilde{\phi}_h d\tilde{\phi}_i d\tilde{\phi}_j d\tilde{\phi}_k. \qquad (113)$$

Let $2\Delta\tilde{\phi}$ be the width of the largest side of the possible rectangulars parallelepiped in $\tilde{\phi}$, then, in the case of $h \neq i \neq j \neq k$,

$$\begin{aligned}
&\int_{\mathcal{I}} \tilde{\phi}_h e(\tilde{\phi}_i)\tilde{\phi}_j e(\tilde{\phi}_k) f(\tilde{\phi}_h, \tilde{\phi}_i, \tilde{\phi}_j, \tilde{\phi}_k) d\tilde{\phi}_h d\tilde{\phi}_i d\tilde{\phi}_j d\tilde{\phi}_k \\
&= \int_{\tilde{\phi}_h \in I^h, \tilde{\phi}_j \in \Phi_j} \tilde{\phi}_h \tilde{\phi}_j \left(\int_{\tilde{\phi}_i \in I^i, \tilde{\phi}_k \in I^k} e(\tilde{\phi}_i) e(\tilde{\phi}_k) f(\tilde{\phi}_h, \tilde{\phi}_i, \tilde{\phi}_j, \tilde{\phi}_k) d\tilde{\phi}_i d\tilde{\phi}_k \right) d\tilde{\phi}_h d\tilde{\phi}_j
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\tilde{\phi}_h \in I^h, \tilde{\phi}_j \in I^j} \tilde{\phi}_h \tilde{\phi}_j \left(\int_{\tilde{\phi}_i \in I^i} (\overline{\tilde{\phi}_i} - \tilde{\phi}_i)(H_i + K_i(\overline{\tilde{\phi}_i} - \tilde{\phi}_i) + O((\overline{\tilde{\phi}_i} - \tilde{\phi}_i)^2)) d\tilde{\phi}_i \right. \\
&\quad \left. \times \int_{\tilde{\phi}_k \in I^k} (\overline{\tilde{\phi}_k} - \tilde{\phi}_k)(H_k + K_k(\overline{\tilde{\phi}_k} - \tilde{\phi}_k) + O((\overline{\tilde{\phi}_k} - \tilde{\phi}_k)^2)) d\tilde{\phi}_k \right) \\
&\quad \times (H_h + K_h(\overline{\tilde{\phi}_h} - \tilde{\phi}_h) + O((\overline{\tilde{\phi}_h} - \tilde{\phi}_h)^2))(H_j + K_j(\overline{\tilde{\phi}_j} - \tilde{\phi}_j) + O((\overline{\tilde{\phi}_j} - \tilde{\phi}_j)^2)) d\tilde{\phi}_h d\tilde{\phi}_j \\
&= \overline{\tilde{\phi}_h \tilde{\phi}_j} H_h H_j K_i K_k \frac{2^4}{3^2} \Delta\tilde{\phi}^8 + O(\Delta\tilde{\phi}^9), \qquad (114)
\end{aligned}
$$

and similarly, in the case of $h = i \neq j = k$,

$$
\int_{\mathcal{I}} \tilde{\phi}_h e(\tilde{\phi}_i) \tilde{\phi}_j e(\tilde{\phi}_k) f(\tilde{\phi}_h, \tilde{\phi}_i, \tilde{\phi}_j, \tilde{\phi}_k) d\tilde{\phi}_h d\tilde{\phi}_i d\tilde{\phi}_j d\tilde{\phi}_k = \overline{\tilde{\phi}_h \tilde{\phi}_i \tilde{\phi}_j \tilde{\phi}_k} H_i H_k K_h K_j \frac{2^4}{3^2} \Delta\tilde{\phi}^8 + O(\Delta\tilde{\phi}^9). \quad (115)
$$

On the other hand, in the case of $h = i = j = k$,

$$
\begin{aligned}
&\int_{\mathcal{I}} \tilde{\phi}_h e(\tilde{\phi}_i) \tilde{\phi}_j e(\tilde{\phi}_k) f(\tilde{\phi}_h, \tilde{\phi}_i, \tilde{\phi}_j, \tilde{\phi}_k) d\tilde{\phi}_h d\tilde{\phi}_i d\tilde{\phi}_j d\tilde{\phi}_k \\
&= \int_{\tilde{\phi}_i \in I^i, \tilde{\phi}_j \in I^j, \tilde{\phi}_k \in I^k} \left(\int_{\tilde{\phi}_h \in I^h} \tilde{\phi}_h^2 e^2(\tilde{\phi}_h)(H_h + K_h(\overline{\tilde{\phi}_h} - \tilde{\phi}_h) + O((\overline{\tilde{\phi}_h} - \tilde{\phi}_h)^2)) d\tilde{\phi}_h \right) \\
&\quad \times (H_i + K_i(\overline{\tilde{\phi}_i} - \tilde{\phi}_i) + O((\overline{\tilde{\phi}_i} - \tilde{\phi}_i)^2))(H_j + K_j(\overline{\tilde{\phi}_j} - \tilde{\phi}_j) + O((\overline{\tilde{\phi}_j} - \tilde{\phi}_j)^2)) \\
&\quad \times (H_k + K_k(\overline{\tilde{\phi}_k} - \tilde{\phi}_k) + O((\overline{\tilde{\phi}_k} - \tilde{\phi}_k)^2)) d\tilde{\phi}_i d\tilde{\phi}_j d\tilde{\phi}_k \\
&= \overline{\tilde{\phi}_h}^2 \overline{\tilde{\phi}_i \tilde{\phi}_j \tilde{\phi}_k} H_h H_i H_j H_k \frac{2^4}{3} \Delta\tilde{\phi}^6 + O(\Delta\tilde{\phi}^7) \qquad (116)
\end{aligned}
$$

and similarly, in the case of $h = j \neq i = k$,

$$
\int_{\mathcal{I}} \tilde{\phi}_h e(\tilde{\phi}_i) \tilde{\phi}_j e(\tilde{\phi}_k) f(\tilde{\phi}_h, \tilde{\phi}_i, \tilde{\phi}_j, \tilde{\phi}_k) d\tilde{\phi}_h d\tilde{\phi}_i d\tilde{\phi}_j d\tilde{\phi}_k = \overline{\tilde{\phi}_h}^2 \overline{\tilde{\phi}_j \tilde{\phi}_k} H_h H_i H_j H_k \frac{2^4}{3} \Delta\tilde{\phi}^6 + O(\Delta\tilde{\phi}^7). \quad (117)
$$

The above show that, when $\Delta\tilde{\phi} \to 0$, the rate of convergence of (114) and (115) to 0 is faster than that of (116) and (117), therefore, we get the following:

$$
\mathrm{E}\left[\left(\sum_{i=1}^{N} \tilde{\phi}_k(i) e(\tilde{\phi}_1(i)) \right)^2 \right] \underset{\Delta y_{\max} \to 0}{\to} N\mathrm{E}\left[\tilde{\phi}_k^2 e^2(\tilde{\phi}_1) \right]. \qquad (118)
$$

□

Lemma A.1

$$
\psi(r) := kr^5 - 18(1-r)^5 + 45(1+r)^2(1-r)^3 + 5(1-r)^7(1+r)^{-2} \qquad (119)
$$

has only one local minimum in $r \in (0,\ 1)$ *when* $0 < k$.

Proof The derivative of $\psi(r)$ is calculated by

$$
\begin{aligned}
\frac{d\psi(r)}{dr} &= (1+r)^{-3}\nu(r) \\
\nu(r) &:= 5k(1+r)^3 r^4 + 90(1+r)^3(1-r)^4 + 90(1+r)^4(1-r)^3 - 35(1+r)(1-r)^6 - 10(1-r)^7 \\
&= (5k-160)r^7 + (15k-480)r^6 + (15k-240)r^5 + (5k+1040)r^4 + 1200r^3 - 1200r^2 - 160r, \qquad (120)
\end{aligned}
$$

therefore, the condition that $\frac{d\psi(r)}{dr}$ has only one zero in $r \in (0,\ 1)$ is equal to that of $\nu(r)$. Note that from (120), we can calculate

$$\frac{d\nu(r)}{dr} = 7(5k-160)r^6 + 6(15k-480)r^5 + 5(15k-240)r^4 + 4(5k+1040)r^3 + 3600r^2 - 2400r - 160 \tag{121}$$

$$\frac{d^3\nu(r)}{dr^3} = 210(5k-160)r^4 + 120(15k-480)r^3 + 60(15k-240)r^2 + 24(5k+1040)r + 7200, \tag{122}$$

and also

$$\nu(0) = 0,\ \ \nu(1) = 5k > 0, \tag{123}$$

$$\frac{d\nu(0)}{dr} = -160 < 0,\ \ \frac{d\nu(1)}{dr} = 220k > 0. \tag{124}$$

When $k = k'$, $0 < k' \ll 1$, it is known that (122) is concave and

$$\frac{d^3\nu(0)}{dr^3} = 7200 > 0,\ \ \frac{d^3\nu(1)}{dr^3} = -73440 + \epsilon(k') < 0. \tag{125}$$

This shows that the sign of $\frac{d^3\nu(r)}{dr^3}$ changes once from positive to negative, that is, the curvature of $\frac{d\nu(r)}{dr}$ changes once from positive to negative, when r increase from 0 to 1 with enough small $k = k'$. From this fact and (124), when $k = k'$, $\frac{d\nu(r)}{dr}$ has only one zero (denote r_z) in $r \in (0,\ 1)$ and the sign of $\frac{d\nu(r)}{dr}$ changes from negative to positive when r increases. Moreover, $\frac{d\nu(r)}{dr}$ is convex from $r = 0$ to the local mininum (denote $r_{\min}$) and increases from $r_{\min}$ to r_z. If $k \ll 1$, $\frac{d\nu(r)}{dr}$ at $k'(\ll 1)$ is added a convex and increasing function:

$$(k-k')(35r^6 + 90r^5 + 75r^4 + 20r^3). \tag{126}$$

Therefore, when $k > 0$, $\frac{d\nu(r)}{dr}$ is convex between 0 and $r_{\min}$ and increases from $r_{\min}$ to r_z. This implies $\frac{d\nu(r)}{dr}$ has only one zero between 0 and r_z. Of course, $\frac{d\nu(r)}{dr}$ has no zero between r_z and $r = 1$ when $k > 0$. In conclusion, $\frac{d\nu(r)}{dr}$ has only one zero at $r \in (0,\ 1)$ and also the sign changes once from negative to positive for all $k > 0$. With this fact and (123), ν has only one zero at $r \in (0,\ 1)$ and its sign changes from negative to positive for all $k > 0$ and we finally conclude the statement of the lemma. □

Proof of Lemma 3.1

From Lemma A.1, it is known that $\psi_1(r_1)$ has only one local minimum in $r_1 \in (0, 1)$. Moreover, from

$$\psi_j(0) = 32,\ \forall j,\ \psi_j(1) = \psi_{j-1}^{\min},\ \psi_0^{\min} = 32 \ \text{ or } \ 432$$

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

$$\psi_1^{\min} < 32. \tag{127}$$

Next, $\psi_2(r_2)$ satisfies

$$\psi_2(0) = 32,\ \psi_2(1) = \psi_1^{\min} < 32,$$

and also $\psi_2(r_2)$ has only one local minimum in $r_2 \in (0, 1)$. This means

$$\psi_1^{\min} > \psi_2^{\min}.$$

Moreover, the term r_1^5 and r_2^5 is a strictly increasing function in $(0, 1]$. Therefore, with $\psi_0^{\min} > \psi_1^{\min}$,

$$r_1^o < r_2^o < 1. \tag{128}$$

By repeating the same process, we finally obtain

$$r_1^o < r_2^o < r_3^o < \cdots < 1.$$

Next show $\lim_{j\to\infty} r_j^o = 1$. Let $\lim_{j\to\infty} r_j^o = r_\infty$. Then, r_∞ satisfies

$$\begin{aligned} r_\infty &:= \arg\min_{r\in[0,1]} \psi_\infty(r) \\ \psi_\infty(r) &:= \psi_\infty^{\min} r^5 - 18(1-r)^5 + 45(1+r)^2(1-r)^3 + 5(1-r)^7(1+r)^{-2} \\ \psi_\infty^{\min} &:= \psi_\infty(r_\infty). \end{aligned} \tag{129}$$

Note that if $\psi_\infty^{\min} > 0$, $\psi_\infty(r)$ has also only one local minimum in $r \in (0,\ 1)$. On the other hand, when $\psi_\infty^{\min} = 0$, it is also known that $\psi_\infty(r)$ is a decreasing function in $r \in [0,\ 1]$ from the proof of Lemma A.1 and $\min_r \psi_\infty(r) = \psi_\infty(1)$. From (129), $\psi_\infty(1) = \psi_\infty^{\min}$, and the minimum is attained at $r = 1$. This means $r_\infty = 1$ (and $\psi_\infty^{\min} = 0$). □

Proof of Lemma 3.2

Consider the subsections I_j (S_j) and I_{j+1} (S_{j+1}), the general case for (34) $\sim$ (41), and from

$$\int_{-k_j}^{k_j} \left(\frac{d_j + d_{j+1}}{2} + z\right)(z - h_j)\, dz = \frac{2}{3}k_j^3 - (d_j + d_{j+1})h_j k_j, \tag{130}$$

the offsets h_j and h_{j+1} such that $\mathrm{E}_{I_j}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] = 0$ and $\mathrm{E}_{I_{j+1}}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] = 0$ are given as

$$h_j = \frac{2}{3}\frac{1}{d_j + d_{j+1}}k_i^2,\ k_j := \frac{d_{j+1} - d_j}{2},\ h_{j+1} = \frac{2}{3}\frac{1}{d_{j+1} + d_{j+2}}k_{j+1}^2,\ k_{j+1} := \frac{d_{j+2} - d_{j+1}}{2} \tag{131}$$

On the other hand, the variance is calculated as

$$\begin{aligned} \mathrm{V}_{I_j}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right] &= \int_{-k_j}^{k_j} \left(\frac{d_j + d_{j+1}}{2} + z\right)^2 (z - h_j)^2\, dz \\ &= A\,(d_{j+1} - d_j)^5 + B\,(d_j + d_{j+1})^2\,(d_{j+1} - d_j)^3\,, \end{aligned} \tag{132}$$

where

$$A := \frac{1}{5\cdot 2^4} - \frac{1}{3^2\cdot 2^3} < 0,\ B := \frac{1}{3\cdot 2^4} > 0. \tag{133}$$

Therefore,

$$\begin{aligned} \mathrm{V}_{I_j} + \mathrm{V}_{I_{j+1}} &= A(d_{j+1} - d_j)^5 + B(d_{j+1} + d_j)^2(d_{j+1} - d_j)^3 \\ &\quad + A(d_{j+2} - d_{j+1})^5 + B(d_{j+2} + d_{j+1})^2(d_{j+2} - d_{j+1})^3 =: Z(d_{j+1}). \end{aligned} \tag{134}$$

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 $(d_{j+2} + d_{j+1})^2$, it is known that $Z(d_{j+1})$ has its minimum at $d_o > \frac{d_j + d_{j+2}}{2}$. This means $|I_j| > |I_{j+1}|$, that is, $|S_j| > |S_{j+1}|$. The same discussion is applicable for arbitrary sections I_j and I_{j+1}, and we can conclude the statement is true. □

Proof of Lemma 3.3

We show a contradiction of an assumption of $\prod_{j=1}^{\infty} \frac{1}{r_j^o} = \gamma < \infty$. At first, define another quantization scheme Q' based on $\mathrm{Q_{opt}}$. In this proof, we refer only to the positive section of the region $[-\tilde{\theta}_1\kappa,\ \tilde{\theta}_1\kappa]$ from the symmetry of the quantization. The partition scheme of Q' is the same that of $\mathrm{Q_{opt}}$ except for the regions on $\tilde{\phi}_1$ corresponding to I_1 and $\bigcup_{j=m+1}^{\infty} I_j$ where m is an appropriate number. Let $2k_m$ denote the width of I_m. The scheme Q' divides the regions on $\tilde{\phi}_1$ corresponding to I_1 and $\bigcup_{j=m+1}^{\infty} I_j$ of $\mathrm{Q_{opt}}$ uniformly into small subsections of a width $2k_m$ and the remainders. Here let I'_j and M' denote the subsections of Q' and their maximum index respectively. Similar to $\mathrm{V}_{I_1}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right]$, define the total variance $\mathrm{V}'_{I_1,k_m}\left[\tilde{\phi}_1 \cdot e(\tilde{\phi}_1)\right]$ in the region I_1 with the quantization scheme Q'. Then,

$$\mathrm{V}_{I_1} - \mathrm{V}'_{I_1,k_m} \tag{135}$$

monotonically increases as $k_m \to 0$. On the other hand,

$$\sum_{j=m+1}^{M'} \mathrm{V}'_{I'_j} - \sum_{j=m+1}^{\infty} \mathrm{V}_{I_j} =: \alpha > 0, \tag{136}$$

and when $m \to \infty$, α monotonically decreases to 0. From the above two observations, there exists a number m that satisfies

$$\mathrm{V}_{I_1} - \mathrm{V}'_{I_1,k_m} > \alpha + \beta \tag{137}$$

for some $\beta > 0$. Then, from (136) and (137), we derive

$$\sum_{j=1}^{M'} \mathrm{V}'_{I'_j} + \beta < \sum_{j=1}^{\infty} \mathrm{V}_{I_j}. \tag{138}$$

For such m, define

$$J := \left(\frac{|I_1|}{2k_m}\right) + \left(\frac{\sum_{j=m+1}^{\infty} |I_j|}{2k_m}\right), \tag{139}$$

where $|\cdot|$ denotes the width of the subsections. Then, from the assumption $\prod_{j=1}^{\infty} \frac{1}{r_j^o} < \infty$, there exists M such that

$$M > m - 1 + J, \ \sum_{j=1}^{M} \mathrm{V}_{I_j} + \beta > \sum_{j=1}^{\infty} \mathrm{V}_{I_j}. \tag{140}$$

From (138),

$$\sum_{j=1}^{M} \mathrm{V}_{I_j} + \beta > \sum_{j=1}^{m+J} \mathrm{V}'_{I'_j} + \beta, \tag{141}$$

and therefore,

$$\sum_{j=1}^{M} \mathrm{V}_{I_j} > \sum_{j=1}^{m+J} \mathrm{V}'_{I'_j} \tag{142}$$

Now define

$$J' := \left(\frac{|I_1|}{2k_m}\right) + \left(\frac{\sum_{j=m+1}^{M} |I_j|}{2k_m}\right), \tag{143}$$

then, $J \geq J'$ and

$$\sum_{j=1}^{M} \mathrm{V}_{I_j} > \sum_{j=1}^{m+J'} \mathrm{V}'_{I'_j}. \tag{144}$$

Note that

$$\bigcup_{j=1}^{M} I_j = \bigcup_{j=1}^{m+J'} I'_j \tag{145}$$

and $M > m + J'$. This means that the number of the quantization levels of Q' in the subsection $\bigcup_{j=1}^{M} I_j$ is less than that of $\mathrm{Q}_{\mathrm{opt}}$, and the variance of the former is also less than that of the latter. This contradicts the optimality of $\mathrm{Q}_{\mathrm{opt}}$. □

Proof of Lemma 3.4

From Lemma 3.1 and its proof, it is known that when $j \to \infty$, r_j^o and $\psi_j^{\min}$ converge to 1 and 0, respectively. Therefore, by employing the Taylor series expansion, $\psi_j(r)$ can be represented by

$$\psi_j(r) = \psi_{j-1}^{\min}(1 - 5(1-r) + 10(1-r)^2 - 10(1-r)^3) + 45 \cdot 2^2(1-r)^3 + O((1-r)^4) \tag{146}$$

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

$$\psi_j(\epsilon) = \psi_{j-1}^{\min}(1 - 5\epsilon + 10\epsilon^2 - 10\epsilon^3) + 180\epsilon^3 + O(\epsilon^4) \tag{147}$$

at $\epsilon \to 0$. Denote the local minimum of $\psi_j(\epsilon)$ as ϵ_j, then ϵ_j should satisfies

$$\psi_{j-1}^{\min}(-5 + 20\epsilon_j - 30\epsilon_j^2) + 540\epsilon_j^2 + O(\epsilon_j^3) = 0. \tag{148}$$

From (148), it is easy to verify

$$\epsilon_j = \left(\frac{1}{108}\psi_{j-1}^{\min}\right)^{1/2} + o\left(\left(\psi_{j-1}^{\min}\right)^{1/2}\right). \tag{149}$$

On the other hand, from (147), $\psi_j^{\min}$ is represented by

$$\psi_j^{\min} = \psi_{j-1}^{\min}(1 - 5\epsilon_j + 10\epsilon_j^2 - 10\epsilon_j^3) + 180\epsilon_j^3 + O(\epsilon_j^4), \tag{150}$$

and with (149), we get

$$\begin{aligned} \psi_j^{\min} - \psi_{j-1}^{\min} &= -5\left(\frac{1}{108}\right)^{1/2} {\psi_{j-1}^{\min}}^{3/2} + 180\left(\frac{1}{108}\right)^{3/2} {\psi_{j-1}^{\min}}^{3/2} + O({\psi_{j-1}^{\min}}^2) \\ &= -5 \cdot 3^{-\frac{5}{2}} {\psi_{j-1}^{\min}}^{\frac{3}{2}} + O({\psi_{j-1}^{\min}}^2). \end{aligned} \tag{151}$$

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

Proof of Lemma 6.1 [20]

The diagonal elements of $U^{\mathrm{T}}U$ are in the form of

$$u^2(-k+1) + u^2(-k+2) + \cdots + u^2(-k+N).$$

From the assumption that every signal u_i is independent, then,

$$\begin{aligned} \mathrm{E}\left[(U^{\mathrm{T}}U)_{ii}\right] &= \mathrm{E}\left[u^2(-k+1) + u^2(-k+2) + \cdots + u^2(-k+N)\right] \\ &= \sum_{j=1}^{N} \mathrm{E}\left[u^2(-k+j)\right] \\ &= N\sigma_u^2. \end{aligned} \tag{152}$$

The variance can be calculated as

$$\mathrm{V}\left[(U^{\mathrm{T}}U)_{ii}\right] = \sum_{j=1}^{N} \mathrm{V}[u(-k+j)^2] = N\eta. \tag{153}$$

On the other hand, the non-diagonal elements $(U^{\mathrm{T}}U)_{ij}$ $(i \neq j)$ are in the form of

$$u(-k+1)u(-l+1) + u(-k+2)u(-l+2) + \cdots + u(-k+N)u(-l+N),\ k \neq l.$$

Then, their expectations are given by

$$\begin{aligned} \mathrm{E}\left[(U^{\mathrm{T}}U)_{ij}\right] &= \mathrm{E}\left[u(-k+1)u(-l+1) + u(-k+2)u(-l+2) + \cdots + u(-k+N)u(-l+N)\right] \\ &= \sum_{m=1}^{N} \mathrm{E}\left[u(-k+m)u(-l+m)\right] \\ &= 0. \end{aligned} \tag{154}$$

The variance is given by noting that $\mathrm{E}\left[(u(k+m)u(l+m)) \times (u(k+n)u(l+n))\right] = 0$, even if $u(l+m) = u(k+n)$ or $u(k+m) = u(l+n)$.

$$\begin{aligned}
\mathrm{V}\left[(U^{\mathrm{T}}U)_{ij}\right] &= \mathrm{E}\left[(u(-k+1)u(-l+1) + u(-k+2)u(-l+2) + \cdots + u(-k+N)u(-l+N))^2\right] \\
&= \sum_{m=1}^{N} \mathrm{E}\left[u(-k+m)^2 u(-l+m)^2\right],\ k \neq l \\
&= N\sigma_u^4
\end{aligned} \tag{155}$$

Here we decompose $U^{\mathrm{T}}U$ as

$$U^{\mathrm{T}}U = (U^{\mathrm{T}}U - N\sigma_u^2 I) + N\sigma_u^2 I,$$

and by employing the norm inequality we obtain

$$\|U^{\mathrm{T}}U\|_1 \geq \|N\sigma_u^2 I\|_1 - \|U^{\mathrm{T}}U - N\sigma_u^2 I\|_1. \tag{156}$$

The value of the first term of the right hand side in (156) is $N\sigma_u^2$, and in the second term, by employing Chebyshev's inequality with (152) and (154), we obtain

$$\mathrm{Prob}\left(|(U^{\mathrm{T}}U - N\sigma_u^2 I)_{ij}| \geq \sqrt{\frac{\mathrm{V}[(U^{\mathrm{T}}U)_{ij}]}{r}}\right) \leq r,$$

and

$$\begin{aligned}
&\mathrm{Prob}\left(\sum_{j=1}^{n} |(U^{\mathrm{T}}U - N\sigma_u^2 I)_{ij}| \geq \sqrt{\frac{\mathrm{V}[(U^{\mathrm{T}}U)_{ii}]}{r}} + (n-1)\sqrt{\frac{\mathrm{V}[(U^{\mathrm{T}}U)_{ij}]}{r}}\right) \\
&= \mathrm{Prob}\left(\sum_{j=1}^{n} |(U^{\mathrm{T}}U - N\sigma_u^2 I)_{ij}| \geq \sqrt{\frac{N}{r}}\left(\sqrt{\eta} + (n-1)\sigma_u^2\right)\right) \leq nr.
\end{aligned}$$

Therefore,

$$\mathrm{Prob}\left(\|U^{\mathrm{T}}U - N\sigma_u^2 I\|_1 = \max_i \sum_{j=1} |(U^{\mathrm{T}}U - N\sigma_u^2 I)_{ij}| \geq \sqrt{\frac{N}{r}}\left(\sqrt{\eta} + (n-1)\sigma_u^2\right)\right) \leq n^2 r.$$

Noting that

$$\begin{aligned}
\|(U^{\mathrm{T}}U)^{-1}\| &= \frac{1}{\inf_x \frac{\|U^{\mathrm{T}}Ux\|}{\|x\|}} = \frac{1}{\inf_x \frac{\|\sigma_u^2 NI + (U^{\mathrm{T}}U - \sigma_u^2 NI)x\|}{\|x\|}} \\
&\leq \frac{1}{\sigma_u^2 N - \sup_x \frac{\|(U^{\mathrm{T}}U - \sigma_u^2 NI)x\|}{\|x\|}},
\end{aligned}$$

this means

$$\mathrm{Prob}\left(\|(U^{\mathrm{T}}U)^{-1}\|_1 \geq \frac{1}{N\sigma_u^2 - \sqrt{\frac{N}{r}}\left(\sqrt{\eta} + (n-1)\sigma_u^2\right)}\right) \leq rn^2.$$

By denoting $\beta_1 := rn^2$ for simplicity, we obtain the statement. □

Proof of Theorem 6.1

First evaluate the magnitude of $\tilde{U}^{\mathrm{T}}E$. Its 1st element $(\tilde{U}^{\mathrm{T}}E)_1$ is of form

$$\tilde{\phi}_1(1)e(1) + \tilde{\phi}_1(2)e(2) + \cdots + \tilde{\phi}_1(N)e(N).$$

From the independence of $\tilde{\phi}_1(i)$ and (56), the expectation and the variance of $(\tilde{U}^{\mathrm{T}}E)_1$ are given as:

$$\mathrm{E}\left[(\tilde{U}^{\mathrm{T}}E)_1\right] = 0,\ \mathrm{V}\left[(\tilde{U}^{\mathrm{T}}E)_1\right] \leq NA\kappa^4(M - K')^{-2}$$

Then by Chebyshev's inequality, we obtain

$$\text{Prob}\left(|\tilde{U}^{\mathrm{T}}E|_1 \geq \sqrt{\frac{A\kappa^4 N}{\beta_2(M-1)^2}}\right) \leq \beta_2,$$

for a reliability index β_2. Combine $(\tilde{U}^{\mathrm{T}}\tilde{U})^{-1}$ and $\tilde{U}^{\mathrm{T}}E$ using a norm inequality:

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

and this gives

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

Therefore we prove the statements. □